Reliability of the TMS-evoked potential in dorsolateral prefrontal cortex

Juha Gogulski[1,2,3], Christopher C. Cline[1,2], Jessica M. Ross[4,1,2], Sara Parmigiani[1,2], Corey J. Keller[1,2,4,δ]

[1]Department of Psychiatry & Behavioral Sciences, Stanford School of Medicine, Stanford, CA, 94305, USA

[2]Wu Tsai Neuroscience Institute, Stanford University, Stanford, CA, USA

[3]Department of Clinical Neurophysiology, HUS Diagnostic Center, Clinical Neurosciences, Helsinki University Hospital and University of Helsinki, Helsinki, FI-00029 HUS, Finland

[4]Veterans Affairs Palo Alto Healthcare System, and the Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Palo Alto, CA, 94394, USA

[δ]Correspondence:
Corey Keller, MD, PhD
Stanford University
Department of Psychiatry and Behavioral Sciences
401 Quarry Road
Stanford, CA 94305-5797
Email: ckeller1@stanford.edu
Phone: +1 8025786292

Running title: Reliability of the prefrontal TEP

Keywords: Transcranial magnetic stimulation (TMS), electroencephalography (EEG), dorsolateral prefrontal cortex (dlPFC), TMS-evoked potentials (TEPs), reliability, cortical excitability

Abstract

Background: We currently lack a robust and reliable method to probe cortical excitability noninvasively from the human dorsolateral prefrontal cortex (dlPFC), a region heavily implicated in psychiatric disorders. We recently found that the strength of *early* and *local* dlPFC single pulse transcranial magnetic stimulation (TMS)-evoked potentials (EL-TEPs) varied widely depending on the anatomical subregion probed, with more medial regions eliciting stronger responses than anterolateral sites. Despite these differences in *amplitude* of response, the *reliability* at each target is not known.

Objective: To evaluate the reliability of EL-TEPs across the dlPFC.

Methods: In 15 healthy subjects, we quantified within-session reliability of dlPFC EL-TEPs after single pulse TMS to six dlPFC subregions. We evaluated the concordance correlation coefficient (CCC) across targets and analytical parameters including time window, quantification method, region of interest, sensor- vs. source-space, and number of trials.

Results: At least one target in the anterior and posterior dlPFC produced reliable EL-TEPs (CCC>0.7). The medial target was most reliable (CCC = 0.78) and the most anterior target was least reliable (CCC = 0.24). ROI size and type (sensor vs. source space) did not affect reliability. Longer (20-60 ms, CCC = 0.62) and later (30-60 ms, CCC = 0.61) time windows resulted in higher reliability compared to earlier and shorter (20-40 ms, CCC 0.43; 20-50 ms, CCC = 0.55) time windows. Peak-to-peak quantification resulted in higher reliability than the mean of the absolute amplitude. Reliable EL-TEPs (CCC up to 0.86) were observed using only 25 TMS trials for a medial dlPFC target.

Conclusions: Medial TMS location, wider time window (20-60ms), and peak-to-peak quantification improved reliability. Highly reliable EL-TEPs can be extracted from dlPFC after only a small number of trials.

Highlights

- Medial dlPFC target improved EL-TEP reliability compared to anterior targets.
- After optimizing analytical parameters, at least one anterior and one posterior target was reliable (CCC>0.7).
- Longer (20-60 ms) and later (30-60 ms) time windows were more reliable than earlier and shorter (20-40 ms or 20-50 ms) latencies.
- Peak-to-peak quantification resulted in higher reliability compared to the mean of the absolute amplitude.
- As low as 25 trials can yield reliable EL-TEPs from the dlPFC.

1. Introduction

Although transcranial magnetic stimulation (TMS) to the dorsolateral prefrontal cortex (dlPFC) is an effective treatment for depression (1), the neural response to TMS is not fully understood. Developing a marker of cortical excitability measurable in the clinic that can 1) stratify patients into different treatment protocols and 2) adapt treatment in real-time using closed-loop methods, would represent significant advances. One such promising metric is derived from TMS coupled with electroencephalography (TMS-EEG), which produces a causal measure of cortical excitability (2,3). Here, the '*EL-TEP'* or the *local* and short-latency (*early,* 20-60 ms) response to single TMS pulses applied to the dlPFC (4) represents an exciting interpretable neurophysiological metric with neural underpinnings (5). EL-TEPs may reflect aspects of prefrontal GABAergic inhibition and NMDA receptor function (6), but these studies need to be further validated and tested in the dlPFC. Furthermore, pilot studies suggest dlPFC EL-TEPs may track with depression pathophysiology (7) and clinical response to TMS (8,9). In summary, EL-TEPs represent an exciting potential brain biomarker for TMS. As multiple dlPFC subregions represent potential novel treatment targets (10–13), understanding how the EL-TEP changes across dlPFC subregions is critical prior to use as a clinical biomarker.

Recent work in our lab investigated the effect of applying single TMS pulses at different dlPFC subregions on EL-TEP amplitude (4). We hypothesized that because the anterolateral portion of the dlPFC activates scalp muscle that induces large early artifacts in EEG (0-20 ms after a single pulse of TMS (14)) which are difficult to remove, the 'true' underlying neural response (e.g. EL-TEPs) may be unintentionally masked or suppressed by offline removal techniques and therefore these targets might produce smaller recorded EL-TEPs. Relatedly, we hypothesized that larger EL-TEPs would be observed from stimulation of the posterior and medial dlPFC due to weaker muscle activation (and therefore smaller masking artifacts). Our findings confirmed these hypotheses, with posterior and medial targets eliciting larger EL-TEPs compared to anterolateral targets (4). However, stronger EL-TEP amplitudes at specific subregions importantly do not necessarily translate to higher reliability. To date, the reliability of EL-TEPs across dlPFC subregions has not been characterized but is critically important to study for prefrontal physiology, pathophysiology, and neuromodulation (15). Furthermore, the effect of analytical parameters including analysis time window and region of interest, how the evoked potential is quantified (e.g. peak-to-peak or mean of the absolute amplitude), and the minimal number of trials needed to obtain a reliable prefrontal EL-TEP, have to date not been characterized. While previous studies have evaluated the reliability of prefrontal TEPs (16,17), this question is worth re-investigating due to an improved understanding of the neural and non-neural artifacts that are present in the TEP. Later (>100 ms) prefrontal TEPs are consistently reliable (16–19), but due to the now well-known contribution of off-target sensory activation (20,21), much of this reliability may be of non-specific sensory brain responses to both the auditory click and tactile sensation to TMS. Reliability measures from earlier components of the TEP (e.g. EL-TEP), less likely to be sensory in nature, have been mixed (16,17). Kerwin *et al*. (17) reported moderate reliability (CCC < 0.6) of early TEPs (the N40 peak) within the same day across different blocks. Conversely, Lioumis *et al*. (16) observed strong correlations for early dlPFC TEP peaks when comparing assessments across different days. For these reasons, and because neither study thoroughly investigated the effects of dlPFC subregion and analytic parameters on reliability, we chose to revisit these critical questions.

In this study, we sought to characterize reliability of EL-TEPs across clinically relevant dlPFC stimulation targets and analytic parameters including ROI size, type (sensor and source space), time window, and amplitude quantification method. Additionally, we evaluated the minimum number of trials needed to obtain highly reliable prefrontal EL-TEPs (CCC>0.7). We focused on within-day reliability to provide reliability thresholds for studies that quantify the acute single-day neural effects of TMS. Here, we hypothesized that 1) because EL-TEP reliability may scale with amplitude of response, a TMS target in posterior and medial aspect of the dlPFC would yield the most reliable prefrontal EL-TEPs (4); 2) how the EL-TEP is quantified (*analytic parameters*) would greatly impact reliability; 3) at least 100 trials would be required for a highly reliable prefrontal EL-TEP (17). Our results demonstrated that 1) the most medial dlPFC target was as predicted the most reliable, while the most anterior target was the least reliable; 2) ROI size did not significantly impact reliability; 3) analytic parameters affected reliability, with later (e.g. starting 30 ms after TMS) and longer (e.g. 40 ms) time windows, peak-to-peak quantification method, and sensor-space yielding more reliable EL-TEPs; and 4) TMS to the most medial dlPFC subregion delivered reliable EL-TEPs with as few as 25 single pulses. These findings demonstrate that the EL-TEP can be reliably measured, and that the choice of stimulation target and analytical parameters can greatly impact EL-TEP reliability.

2. Methods

2.1. Participants

22 healthy participants (23-64 years old, mean=39.45, SD=13.00, 7F/15M) were recruited into the study and provided written informed consent under a protocol approved by the Stanford University Institutional Review Board. Details on the demographics have been reported previously (4). Briefly, the participants were aged 18-65, fluent in English, fully vaccinated against COVID-19, and without moderate or severe depression. Exclusion criteria included psychiatric/neurological disorders including moderate or severe depression: QIDS ≥11 (22),

substance abuse during the month before screening, cardiac events, current pregnancy, contraindications for TMS or MRI (23,24), and lifetime history of psychotropic medication use. Out of the initial participants, 6 were excluded (3 before the experimental session, 1 for TMS intolerability assessed by delivering test single pulses before the experiment, 1 for lower back pain, and 2 for technical difficulties), leaving 15 subjects for analysis after completing MRI and TMS-EEG sessions.

2.2. Transcranial Magnetic Stimulation

Single-pulse TMS was delivered using a MagVenture Cool-B65 A/P figure-of-eight coil with a MagPro X100 stimulator (MagVenture, Denmark). A thin (0.5 mm) foam pad was attached to the TMS coil to minimize electrode movement, scalp sensation, and bone-conducted auditory response (25). The TMS coil was held and positioned automatically by a robotic arm (TMS Cobot, Axilum Robotics, France). MRI-based neuronavigation (Localite TMS Navigator, Localite, Germany) was used to align the coil to pre-planned individualized coil orientations for each stimulation target (see section 2.2.1). To estimate resting motor threshold (rMT), biphasic single pulses of TMS were delivered to the hand region of the left motor cortex with the coil held tangentially to the scalp and at 45° from midline (26,27). The optimal motor hotspot was defined as the coil position from which TMS produced the largest and most consistent visible twitch in relaxed right first dorsal interosseous muscle (FDI; (27)). rMT was determined to be the minimum intensity that elicited a visible twitch in relaxed FDI in ≥ 5 out of 10 pulses (28,29). Stimulation intensity was adjusted relative to motor threshold based on electric field strength at the cortical surface, as reported previously (4).

2.2.1 dlPFC Target Selection

For each subject, single pulses of TMS were delivered to six different targets within the left dlPFC (Fig 1B). Targeting details have been reported previously (4). Briefly, the number of targets and spatial locations were chosen to balance the following: sampling of the anterior / posterior and medial / lateral planes in the middle frontal gyrus and sampling of the many utilized clinical dlPFC targets for depression (12). *t1* is close to the commonly used 5 cm average clinical target (10). *t2* corresponds to the EEG F3 site (11). *t3* is located midway between two targets: the BA9 definition (30) and a treatment site with resting fMRI activity negatively correlated to anterior subgenual cingulate cortex (sgACC) (10). *t4* corresponds to the treatment site with activity that is most negatively correlated to the sgACC (10). *t5* roughly corresponds to the TMS treatment which corresponded to better treatment success in the study by Herbsman *et al*. 2009 (13). *t6* roughly corresponds to the 5.5 cm average target (31). The six targets were registered to each individual brain using a non-linear inverse transformation from MNI space (32).

2.2.2. Experimental Design

The order of stimulation of the targets was pseudorandomized within the session. 150 single pulse TMS trials were collected for each experiment block. Two blocks of single pulse data were collected on the same day, separated by an average of 97 ∓ 37 minutes. TMS pulses were delivered every 1.5 s, jittered +/- 25% (33). Work with intracranial stimulation has shown that an interstimulus interval of 1 s may not produce acute effects in cortico-cortical evoked potentials (34). To maintain levels of attention, subjects were shown neutral images on a screen and were instructed to keep eyes open during stimulation. Subjects were monitored throughout the experiment.

2.4. Electroencephalography

EEG was recorded using a 64-channel TMS-compatible amplifier (ActiCHamp Plus, Brain Products GmbH, Germany) with a 25 kHz sampling rate. Slim, freely rotatable, active electrodes (actiCAP slim, Brain Products GmbH, Germany) were used in a standard montage labeled according to the extended 10-20 international system. EEG data were online referenced to the 'Cz' electrode and recorded using BrainVision Recorder software (Brain Products GmbH, Germany). Impedances were monitored and percentage of channels with impedances <10 kΩ was 97.54 ∓ 2.31%. TMS "click" frequency matched noise masking and earmuffs, as described in detail previously (35), were applied to reduce off-target sensory effects.

2.4.1. Preprocessing of TMS-EEG data

TMS-EEG preprocessing was performed with the fully automated AARATEP pipeline (version 2, (36)). Epochs were extracted from 800 ms before to 1000 ms after each TMS pulse. Data between 2 ms before to 12 ms after each pulse were replaced with values interpolated by autoregressive extrapolation and blending (36), downsampled to 1 kHz, and baseline-corrected based on mean values between 500 to 10 ms before the pulse. Epochs were then high-pass filtered above 1 Hz with a modified filtering approach to reduce the spread of pulse-related artifact into baseline time periods. Bad channels were rejected via quantified noise thresholds and replaced with spatially interpolated values (see (37) for all details on channel deletion and interpolation), with 1.8 +/- 2.0% (mean +/- standard deviation) channels rejected on average. Eye blink artifacts were attenuated by a dedicated round of independent component analysis (ICA) and eye-specific component labeling and rejection using ICLabel (38), a modification from the original AARATEP pipeline introduced in version 2. An average of 1.7 +/- 0.7 components were rejected at this stage. Various non-neuronal noise sources were attenuated with SOUND (39). Decay artifacts were reduced via a specialized decay fitting and removal procedure (36). Line noise was attenuated with a bandstop filter between 58-62 Hz. Additional artifacts were attenuated with a second stage of ICA and ICLabel labeling and rejection, with rejection criteria targeted at removing any clearly non-neural signals (see (37) for all data deletion criteria). 60.6% +/- 10.3% of components were rejected, accounting for 30.3% +/- 20.5% of variance in the signal at this stage. Data were again interpolated between -2 and 12 ms with autoregressive extrapolation and blending, low-pass filtered below 100 Hz, and average re-referenced. For complete details of the pipeline implementation, see (36) and source code at github.com/chriscline/AARATEPPipeline. For each target and analytical parameter combination (see section 2.5.), trial-averaged EL-TEPs were calculated with a 10% trimmed mean to prevent a small number of outlier trials from confounding results.

A subset of recorded conditions were processed two additional times with identical pipeline parameters as above to quantify variability between preprocessing runs due to non-deterministic parts of the pipeline, including random variation in ICA decomposition (40).

2.4.2. Quantification of reliability

The ability of a biomarker to distinguish different individuals is best assessed by an intraclass correlation coefficient (ICC), which represents the ratio of the inter-participant variance to the total variance (41,42). To quantify within-session reliability of prefrontal EL-TEPs, we applied the concordance correlation coefficient (CCC), a form of ICC optimally modified to assess test-retest reliability (17,43), and defined the following correlation strength categories: low [<0.5], moderate [0.5, 0.7], and high [0.7, 1.0] correlation. For each target and analytical parameter combination, the magnitude of subjects' EL-TEP of the first block was compared to that of the second block (Fig 1A) to obtain a group-level CCC value. Mean and 95% confidence intervals of the CCC were calculated across target and analytical parameter combinations. To compare CCCs across conditions, we quantified the means and ranges (Fig 2-3). Reliability calculations of different trial counts is described in section 2.4.3.

In a supplementary analysis, we studied how reproducible each TEP is across the two experimental blocks. We calculated cosine similarity (44) across blocks for each stimulation target (Fig S1). Here, we did not restrict the analysis to any predetermined ROI but instead used all EEG channels as input. Cosine similarity was calculated within and between subjects.

2.4.3. Subset analysis of TMS trial counts

In order to explore what the minimum number of trials were needed to obtain a reliable EL-TEP (CCC>0.7), 150-trial stimulation blocks were divided into 25, 50, 75, 100 and 125 trial subsets, including trials 1-25, 1-50, 1-75, 1-100, and 1-125. Each subset was preprocessed separately to simulate an experiment with a reduced number of trials, prior to EL-TEP amplitude quantification. EL-TEP was quantified using a fixed set of parameters based on our other CCC results which showed that a medium size ROI with six prefrontal channels, 20-60 ms time window, and peak-to-peak amplitude quantification is one of the most reliable parameter combinations (Fig 3). To obtain CCC, EL-TEP amplitude was compared either to 1) the amplitude of the same block when using a full set of 150 trials (*reliability within blocks*), 2) to the amplitude of another experimental block of the same target using a full set of 150 trials (*reliability between blocks*), or 3) to the amplitude of another block using the same amount of trials (*test-retest reliability*). *Reliability within blocks* was calculated to estimate the minimum number of trials needed to obtain a consistent EL-TEP response compared to the EL-TEP from a large number of trials, with minimal variation in coil positioning, brain state, or other experiment factors that may change over longer time scales. *Reliability between blocks* was calculated to evaluate the consistency of EL-TEPs over a longer time scale, varying just the number of trials in one block but using the maximum amount of available data (150 trials) in the other block as a proxy for a "ground truth" best TEP estimate. *Test-retest reliability* was performed to mimic experimental conditions in which one has less than 150 trials in each of both blocks, and provide an estimate of a minimum number of trials to obtain reliable EL-TEPs within a session.

2.4.4 Quantification of within-block noise and SNR

As part of a supplemental analysis, we examined how within-block variance related to across-block reliability. We employed the bootstrapped standardized measurement error (bSME) (45) to quantify uncertainty in EL-TEP peak-to-peak and mean(abs) amplitude measures with 10000 bootstrap repeats per estimate, accounting for trial-to-trial variation in responses and number of trials aggregated. With bSME calculated on logarithmically-scaled dBμV values, we obtained bSME measures in units of dBμV, and calculated per-block SNR as the mean response amplitude (in dBμV) minus the estimated bSME value for the same block. Supplemental Fig S15 shows the results of analyses examining possible relationships between within-block variance and across-block reliability. While correlations between within-block response amplitude and within-block variance were identified (Fig S15 A, B, E, F), there were no strong correlations identified between within-block SNR and reliability across blocks (Fig S15 C, D, G, H).

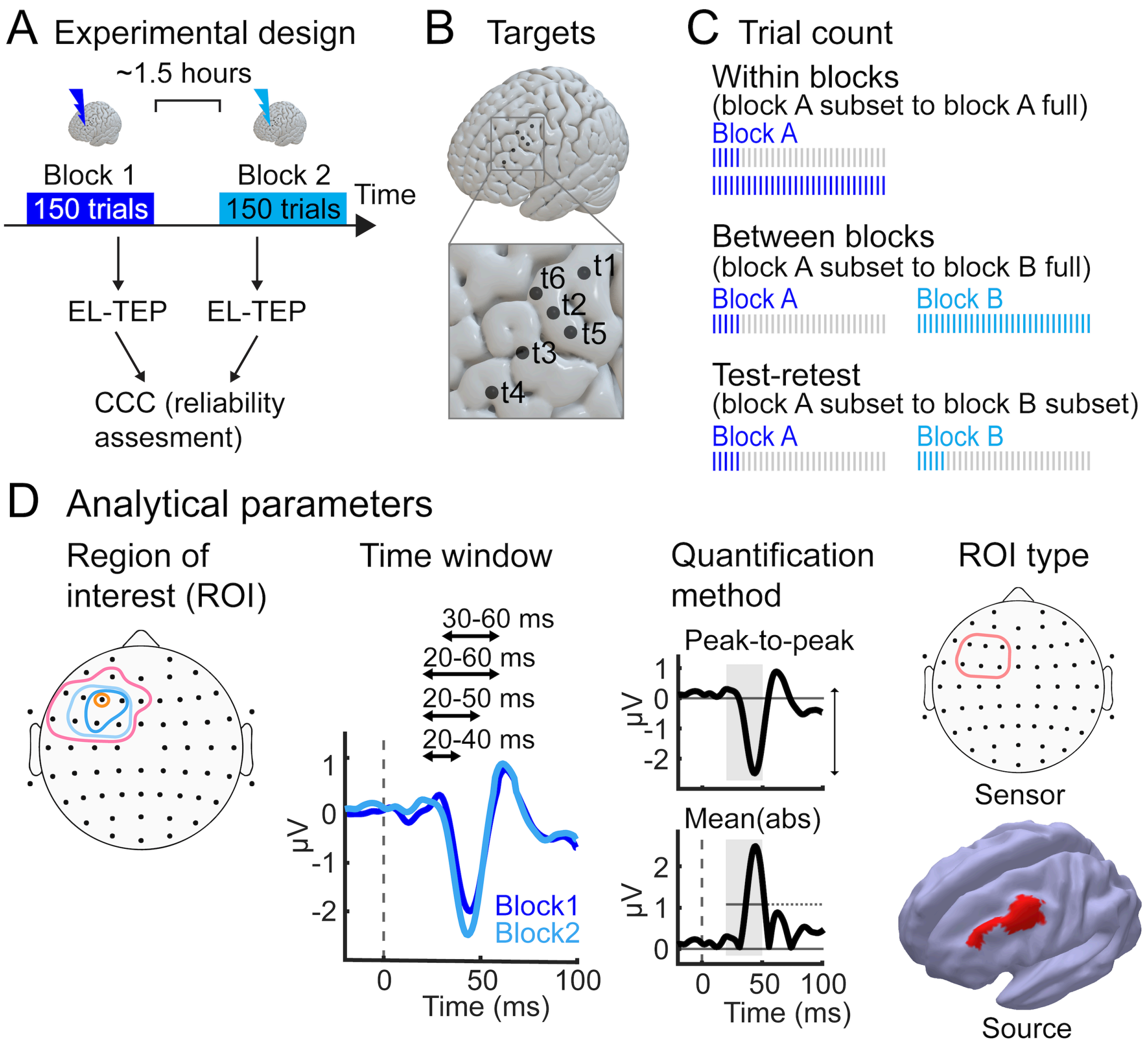


Figure 1 - Study schematic. A) Experiment consisted of two single-pulse TMS-EEG blocks, separated by approximately 1.5 hours. For each subject, EL-TEPs were quantified for both experimental blocks. B) TMS was delivered to six different targets within the dlPFC. C) Retrospectively, each stimulation block was divided into subsets consisting of a smaller number of trials than the full sample (150 trials). Each subset was preprocessed separately. D) In addition to experimental manipulations (stimulation target and number of trials), we investigated how reliability of the dlPFC EL-TEP varied as a function of analytic parameters. Within the sensor space, four different ROI sizes and time windows, as well as two different quantification methods, were evaluated. Reliability of sensor space EL-TEPs was also compared to source space EL-TEPs.

2.5. Analytical parameters

Our goal was to characterize the reliability of prefrontal EL-TEPs across dlPFC stimulation targets and analytical parameters, defined as follows: quantification method, time window, ROIs size, and ROI type (Fig 1).

2.5.1. Quantification method

We quantified EL-TEP amplitude by computing both a peak-to-peak amplitude metric and mean of the absolute amplitudes (mean(abs)), averaged across samples within each time window and ROI (see 2.5.2.). Because the number and timing of EL-TEP peaks from stimulation of the dlPFC are variable across targets and subjects (4,16), we quantified the peak-to-peak amplitude across different time windows using subtraction of the signal minimum from the signal maximum within a time window of interest (46), rather than the difference in amplitude of specific predefined TEP peaks. Both peak-to-peak and mean(abs) metrics were quantified in units of dBμV as linear-scale (μV) responses exhibited a roughly log-normal distribution (Fig S2).

2.5.2. Time window

We focused on the early and local local response of the TEP (EL-TEP), which captures previously reported peaks (2,47,48) including the P30, N40, and P50, because 1) there is evidence for the neural basis of the early TEP components (34,49); 2) this time window is commonly studied in the literature (2,47,48); and 3) off-target sensory confounds are less prominent in these earlier components of the TEP (20,35,50–53). However, the optimal time window for quantifying prefrontal EL-TEPs remains unknown. Therefore, we assessed the test-retest reliability of EL-TEPs by calculating it within four distinct time windows: 20-40 ms, 20-50 ms, 30-60 ms, and 20-60 ms.

2.5.3. ROI size (sensor space)

As there is no agreed upon optimal ROI size, we evaluated the effect of four predetermined prefrontal sensor-space ROIs on reliability (Fig 1D). These ROIs included a single electrode (F3), a small ROI (F3, F1, FC3), a medium ROI (F5, F3, F1, FC5, FC3, FC1), and a large ROI (F5, F3, F1, FC5, FC3, FC1, FT7, F7, AF3, Fz). We selected these ROIs to sample from and be centered over our dlPFC stimulation targets. As a followup analysis, to ensure these ROIs were not biasing results, we repeated the analyses with shifted ROIs (Fig S3 and Fig S4). See Supplementary Material for more details.

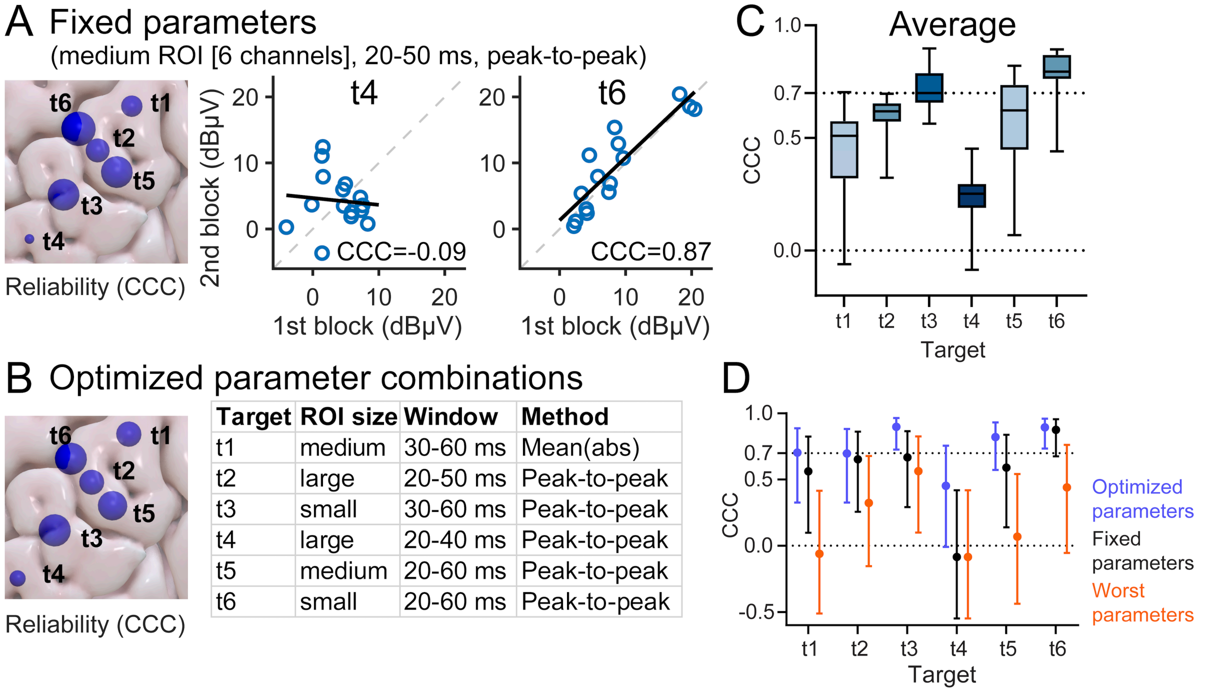
2.5.4. Source space estimation

Subject-specific differences in gyral anatomy can cause underlying common cortical sources to project to the scalp in different topographies across subjects (54). To account for this and other related consequences of EEG volume conduction, we performed EEG source estimation and asked whether the source-space EL-TEPs would be more reliable compared to the sensor space TEPs. Using digitized electrode locations and individual head models constructed from subjects' anatomical MRI data (55), subject-specific forward models of signal propagation from dipoles distributed over and oriented perpendicular to the cortical surface to electrodes on the scalp were constructed (56,57). Inverse kernels mapping measured scalp EEG activity to underlying cortical sources were estimated using weighted minimum-norm estimation (wMNE) as implemented in Brainstorm (58). A cortical region of interest (ROI) was constructed from a combination of HCP MMP1 atlas (59) parcels (46 L, 8Av L, and 8C L; Fig 1D lower right), and mapped from a common group template surface (ICBM152) to subject-specific cortical surfaces via surface-based morphometry.

3. Results

3.1. Reliability across stimulation targets

We first investigated the effect of EL-TEP reliability across different targets within the dlPFC (Fig 1B). We characterized the effect of region of interest (ROI) size and type, time window, and quantification method – together referred to as *analytic parameters* – on reliability. Across all analytic parameters in sensor space (ROI size, time window, and quantification method), targets t3 (mean CCC_{t3}=0.72, range 0.56 to 0.90) and t6 (mean CCC_{t6}=0.78, range 0.44 to 0.89) were highly reliable (CCC>0.7) on average, whereas all other targets had average CCCs under 0.7 (Fig 2C; CCC_{t1}=0.44, CCC_{t2}=0.59, CCC_{t4}=0.24, CCC_{t5}=0.56).



Target	ROI size	Window	Method
t1	medium	30-60 ms	Mean(abs)
t2	large	20-50 ms	Peak-to-peak
t3	small	30-60 ms	Peak-to-peak
t4	large	20-40 ms	Peak-to-peak
t5	medium	20-60 ms	Peak-to-peak
t6	small	20-60 ms	Peak-to-peak

Figure 2 - EL-TEP reliability across dlPFC targets for fixed and optimized analytic parameters. A) EL-TEP reliability across dlPFC targets using a fixed parameter combination previously used in an earlier study (medium ROI size, 20-50 ms time window, peak-to-peak amplitude; (4)). Scatter plots show EL-TEP amplitudes of the first and second experimental

block for targets t4 and t6. B) Analytical parameter combinations that produced the most reliable EL-TEP amplitude for each target. On the left, marker size represents the CCC of the most reliable parameter combination. The table shows the most reliable parameters for each target. C) Average reliability across analytical parameters. D) Best and worst parameter combination compared to the fixed parameter combination for each target. In A and B, the blobs visualized on the brain surface represent CCCs but are not drawn to scale. In C, boxes represent the distribution of CCCs across analytical parameters in sensor space. In D, dots denote mean CCC and whiskers denote 95% CIs.

When using an *a priori* chosen fixed combination of parameters used previously to quantify the EL-TEP (medium sized sensor space ROI, 20-50 ms, peak-to-peak) (4), only target t6 was highly reliable (Fig 2A and D; CCC_{t6}=0.87, CI [0.68, 0.96]). However, when considering the set of all tested analytic parameters in sensor space, all targets except t4 (CCC_{t4}=0.62, CI [0.18, 0.85]) had a mean CCC of greater than 0.7 for at least one parameter combination (Fig 2B and D; CCC_{t1}=0.70, CI [0.33, 0.89]; CCC_{t2}=0.74, CI [0.42, 0.90]; CCC_{t3}=0.98, CI [0.96, 0.99]; CCC_{t5}=0.89, CI [0.70, 0.96]; CCC_{t6}=0.95, CI [0.86, 0.98]). Of those, targets t3 and t6 had a lower 95% CI bound of greater than 0.7 indicating that at these two targets, with optimized parameters, almost all data should be expected to be highly reliable. Conversely, using a poor combination of parameters for a target could be expected to result in unreliable EL-TEPs; using the worst possible combination of analytic parameters (sensor space combination that yields the smallest CCC for each target) resulted in none of the targets being highly reliable (Fig 2D; all targets' CCC < 0.6).

We also explored the spatial consistency of maximal EL-TEP amplitude across targets. When using a parameter combination of 20-60 ms time window, medium ROI and peak-to-peak quantification, 9/15 subjects exhibited the largest EL-TEP amplitude in the same target for both the first and second stimulation blocks. Moreover, 14/15 subjects exhibited the largest EL-TEP amplitude in either the same or adjacent stimulation target.

We next quantified cosine similarity across the two experimental blocks when all EEG channels were used (Fig S1). For similarity within subjects, no significant difference was found between the stimulation targets.

3.2. Reliability across analytical parameters

We next investigated the impact of various analytical parameters on the test-retest reliability of prefrontal EL-TEPs. Specifically, we examined the influence of ROI size and type, time window, and amplitude quantification method on the reliability. Table 1 shows the sensor space conditions where the lower bound of the 95% confidence interval for CCC exceeded 0.7, indicating a high level of reliability for most of the data. Of note, these reliable conditions used either the t6 or t3 targets and peak-to-peak quantification method. Median CCCs (across targets) for all analytical parameters in sensor space are shown in Fig 3B. CCCs for all targets and sensor-space parameters are shown in Fig S5 and Fig S6.

Target	ROI size	Time window	Quantification method	Mean CCC	95% CI (lower)	95% CI (higher)
t6	small	20-60 ms	Peak-to-peak	0.89	0.73	0.96
t6	medium	20-60 ms	Peak-to-peak	0.89	0.73	0.96
t6	medium	30-60 ms	Peak-to-peak	0.89	0.72	0.96
t3	small	30-60 ms	Peak-to-peak	0.90	0.73	0.96

Table 1 - Experimental conditions that yield highly reliable dlPFC EL-TEPs. The table shows sensor-space conditions in which the 95% confidence interval lower bound of the mean CCC was >0.7, indicating high reliability for most of the data.

3.2.1. ROI size

We next examined the effect of sensor-space ROI size on prefrontal EL-TEP reliability (Fig 3C-F). We hypothesized that larger ROIs would be more reliable due to less susceptibility to the influence of artifact outliers in individual channels. The results showed that when averaging across all stimulation targets and time windows, EL-TEP reliability does not vary widely across ROIs tested (Fig 3D; *one electrode* mean CCC: 0.53, range -0.06 to 0.87; *small ROI* mean CCC: 0.56, range 0.040 to 0.90; *medium ROI* mean CCC: 0.55, range -0.085 to 0.89; *large ROI* mean CCC: 0.56, range 0.0023 to 0.87). When utilizing the most reliable time window (20-60 ms; Fig 3H) and amplitude quantification method (peak-to-peak; Fig 3L), both target t3 and target t6 demonstrated high reliability across various analytical parameters. However, in terms of the lower bound of the 95% confidence interval, only target t6 exhibited values exceeding 0.7, in small and medium ROIs (Fig 3F; $\text{CCC}_{\text{small}}$=0.89, CI [0.73, 0.96]; $\text{CCC}_{\text{medium}}$=0.89, CI [0.73, 0.96]).

We next examined the influence of shifting ROIs anteriorly or posteriorly (Fig S3 and Fig S4). While other targets had stable CCCs across the shifted ROIs, for the most anterior and posterior targets (t1 and t4) the shift in ROI positioning had a large effect on the CCCs (Fig S4).

3.2.2. Time window

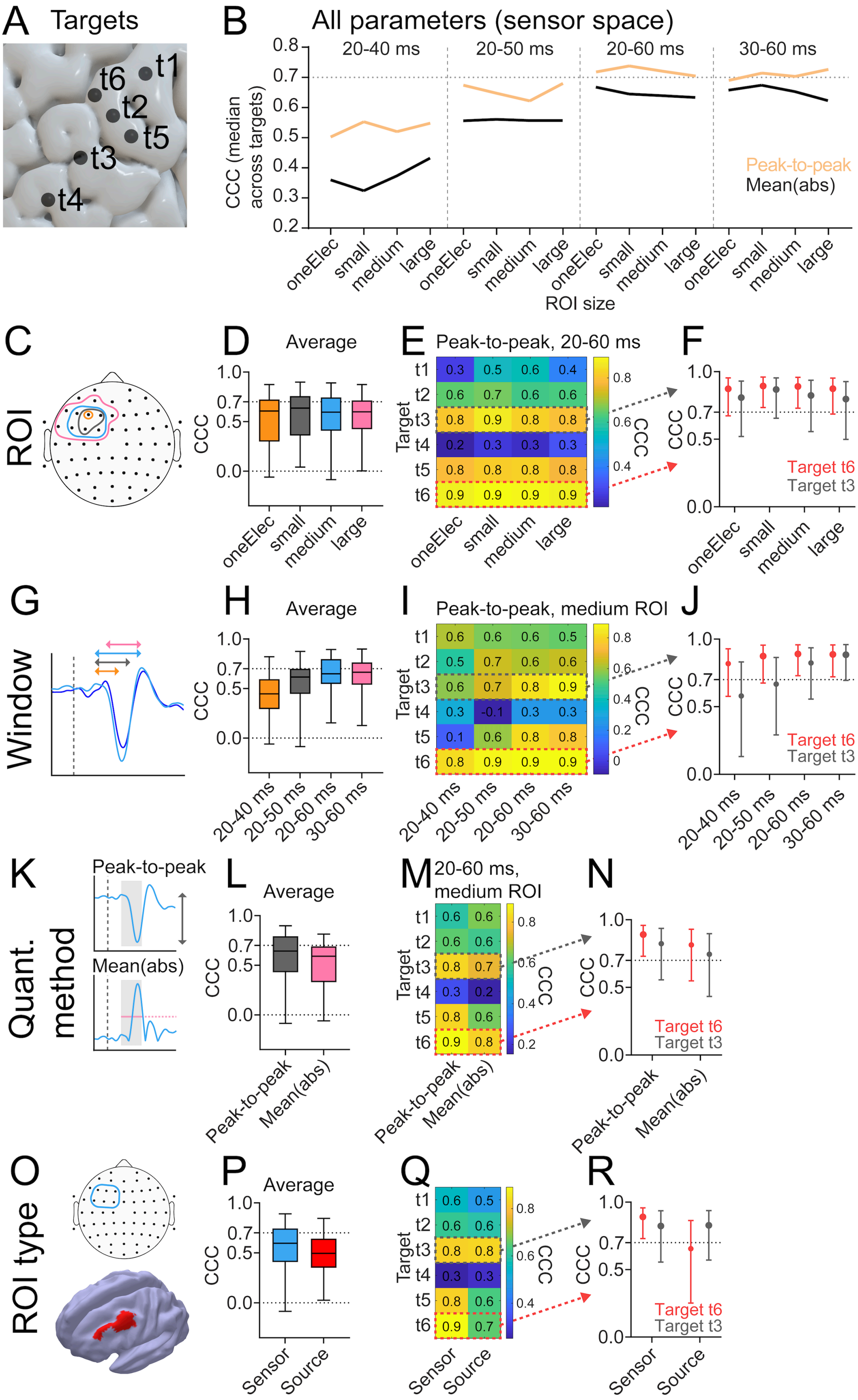
Next, we studied the effect of the time window on prefrontal EL-TEP reliability in sensor space (Fig 3G-J and Fig S7-10). Because of numerous peaks previously reported within the EL-TEP time window (51,60,61), we hypothesized that a longer window (20-60 ms) that captured the P30/N40/P50 peaks would have greater reliability than shorter time windows (30-60 ms, 20-40 ms, 20-50 ms). When averaging across stimulation targets and ROI sizes, the longer time window (20-60 ms) did in fact have greater reliability (Fig 3H; mean $CCC_{20\text{-}60ms}$=0.62, range 0.15 to 0.89) compared to the shorter 20-40 ms and 20-50 ms time windows (Fig 3H; mean $CCC_{20\text{-}40ms}$=0.43, range -0.06 to 0.82; mean $CCC_{20\text{-}50ms}$=0.55, range -0.085 to 0.87). However, one shorter time window that included later latencies, 30-60 ms, also had high reliability (Fig 3H; mean $CCC_{30\text{-}60ms}$=0.61, range 0.13 to 0.90), suggesting that high EL-TEP reliability might be driven more by the later latencies of the EL-TEP rather than the length of the time window analyzed. When employing a medium ROI and peak-to-peak amplitude quantification, target t6 exhibited the greatest reliability, with CCC values exceeding 0.8 across all included time windows (Fig 3J; $CCC_{t6,20\text{-}40ms}$=0.82, CI [0.58, 0.93]; $CCC_{t6,20\text{-}50ms}$=0.87, CI [0.68, 0.96]; $CCC_{t6,20\text{-}60ms}$=0.89, CI [0.73, 0.96]; $CCC_{t6,30\text{-}60ms}$=0.89, CI [0.72, 0.96];). Target t3 demonstrated CCCs greater than 0.8 in the longer 20-60 ms and later 30-60 ms time windows (Fig 3J; $CCC_{t3,20\text{-}40ms}$=0.58, CI [0.13, 0.83]; $CCC_{t3,20\text{-}50ms}$=0.67, CI [0.29, 0.86]; $CCC_{t3,20\text{-}60ms}$=0.82, CI [0.56, 0.94]; $CCC_{t3,30\text{-}60ms}$=0.88, CI [0.69, 0.96];)). Peak-to-peak amplitudes for all time windows in medium size ROI are shown in Supplementary Material (Fig S7-10 for dBμV and Fig S11-14 for μV units). In summary, longer (20-60 ms) and later (30-60 ms) time windows increased EL-TEP reliability.

3.2.3. Amplitude quantification method

We also studied the effect of EL-TEP quantification method on prefrontal EL-TEP reliability (in sensor space, Fig 3K-N). On average, peak-to-peak quantification was more reliable than mean(abs) (Fig 3L; $CCC_{peak\text{-}to\text{-}peak}$=0.59, range -0.085 to 0.90; $CCC_{mean(abs)}$: 0.52, range -0.060 to 0.81). When using a medium size ROI and 20-60 ms time window, both targets t6 and t3 had mean CCC>0.7 for both quantification methods (Fig 3N; $CCC_{t6,peak\text{-}to\text{-}peak}$=0.89, CI [0.73, 0.96]; $CCC_{t6,mean(abs)}$=0.81, CI [0.55, 0.93]; $CCC_{t3,peak\text{-}to\text{-}peak}$=0.82, CI [0.56, 0.94]; $CCC_{t3,mean(abs)}$=0.74, CI [0.43, 0.90]). In summary, peak-to-peak quantification increased reliability compared to the mean(abs).

3.2.4. ROI type

We next compared the reliability of prefrontal EL-TEPs in sensor-space to those in source-space (ROI type; Fig 3O-R). We quantified the peak-to-peak EL-TEP amplitudes in sensor and source space and compared the reliability of dlPFC source-based ROI (see Methods) to a medium-size sensor space ROI. We hypothesized that because inverse estimation may better focus on the local stimulation response, source space analysis would yield more reliable EL-TEPs. In contrast, on average, EL-TEPs were more reliable in sensor space compared to source space (Fig 3P; mean CCC_{sensor}=0.55, range -0.085 to 0.89; mean CCC_{source}=0.49, range 0.026 to 0.85). When utilizing peak-to-peak quantification and a time window of 20-60 ms, both targets t6 and t3 exhibited mean CCC values exceeding 0.8 in sensor space (Fig 3R; $CCC_{t6,sensor}$=0.89, CI [0.73, 0.96]; $CCC_{t3,sensor}$=0.82, CI [0.56, 0.94]). However, in source space t6 displayed lower reliability compared to t3 (Fig 3R; $CCC_{t6,source}$=0.65, CI [0.25, 0.86]; $CCC_{t3,source}$=0.83, CI [0.57, 0.94]). In summary, EL-TEPs in sensor space (medium size ROI) were more reliable than our source-space EL-TEPs metric.



E Peak-to-peak, 20-60 ms

Target	oneElec	small	medium	large
t1	0.3	0.5	0.6	0.4
t2	0.6	0.7	0.6	0.6
t3	0.8	0.9	0.8	0.8
t4	0.2	0.3	0.3	0.3
t5	0.8	0.8	0.8	0.8
t6	0.9	0.9	0.9	0.9

I Peak-to-peak, medium ROI

Target	20-40 ms	20-50 ms	20-60 ms	30-60 ms
t1	0.6	0.6	0.6	0.5
t2	0.5	0.7	0.6	0.6
t3	0.6	0.7	0.8	0.9
t4	0.3	-0.1	0.3	0.3
t5	0.1	0.6	0.8	0.8
t6	0.8	0.9	0.9	0.9

M 20-60 ms, medium ROI

Target	Peak-to-peak	Mean(abs)
t1	0.6	0.6
t2	0.6	0.6
t3	0.8	0.7
t4	0.3	0.2
t5	0.8	0.6
t6	0.9	0.8

Q

Target	Sensor	Source
t1	0.6	0.5
t2	0.6	0.6
t3	0.8	0.8
t4	0.3	0.3
t5	0.8	0.6
t6	0.9	0.7

Figure 3 - EL-TEP reliability across all analytical parameters. A) Stimulation targets in the dlPFC. B) Median CCCs across targets for all analytical parameters in sensor space. C-F) Effect of ROI size on EL-TEP reliability (in E and F peak-to-peak amplitude within 20-60 ms time window). G-J) Effect of time window on EL-TEP reliability (in I and J peak-to-peak amplitude within medium-size ROI). K-N) Effect of amplitude quantification method on EL-TEP reliability (in M and N medium-size ROI, 20-60 ms time window). B-N are in sensor space. O-R) Effect of ROI type (sensor versus source space) on EL-TEP reliability. In P-R the sensor space data represents the medium size ROI with peak-to-peak quantification and 20-60 ms time window. The box plots in D, H and L represent the distribution of CCCs across targets and other analytical parameters in sensor space. Last column (F, J, N and R) represents the two most reliable targets (t3 and t6) across the parameters (dots: mean, error bars: 95% CI). The dots denote mean CCC and whiskers denote 95% CIs.

3.3. Influence of trial counts on EL-TEP reliability

Next, we explored how many single pulses of TMS (*trials*) are required to yield a reliable EL-TEP (Fig 4). We first calculated the CCCs to assess the reliability *within* blocks by comparing each *n*-trial subset to the full set of trials from the same block. This allowed us to compare trial subsets to a full “gold standard set” of trials, answering the question *how close to the full EL-TEP block can one get from a subset of trials*? We found that for all targets except t4, 50 trials were enough to produce similar EL-TEPs to the full block (Fig 4C-D; CCCs > 0.7). For targets t3 and t6, only 25 trials were needed to yield high reliability compared to the full block ($CCC_{t3,within,25}$=0.74; $CCC_{t6,within,25}$=0.86).

We then calculated the reliability *between* blocks to evaluate the consistency of EL-TEPs over a longer time scale (~1.5 hours). To do so, we studied the effect of the number of trials in the first block against the maximum amount of available data (150 trials) in the other block as a proxy for a “ground truth” best TEP estimate (Fig 4 C-D). For target t6, only 25 trials were enough to yield high reliability between blocks ($CCC_{t6,between,25}$=0.78). For targets t1, t2 and t3, subsets of 50 or more trials produced moderate to high reliability ($CCC_{t1,between,50}$=0.63; $CCC_{t2,between,50}$=0.65; $CCC_{t3,between,50}$=0.71). Target t4 yielded low reliability between blocks regardless of the trial count (CCC < 0.3).

Next, to mimic experimental conditions in which one may record fewer than 150 trials per block, we evaluated the *test-retest* reliability of experimental blocks by comparing the amplitude of one block to the amplitude of another block using the same number of trials (Fig 4C-D). For target t6, only 25 trials were enough to produce high test-retest reliability ($CCC_{t6,test-retest,25}$=0.73). Target t3 produced high and t5 produced moderate test-retest reliability with more than 125 trials ($CCC_{t3,test-retest,125}$=0.72; $CCC_{t5,test-retest,125}$=0.52). Target t4 produced low test-retest reliability regardless of the trial count (CCC < 0.3).

In summary, when stimulating at targets t3 or t6, as few as 25 trials were enough to yield EL-TEPs that were reliable within an experimental block. Moreover, for target t6 as few as 25 trials were also sufficient for high test-retest and between-block reliability. For other dlPFC targets, reliable EL-TEPs between blocks may require more trials.

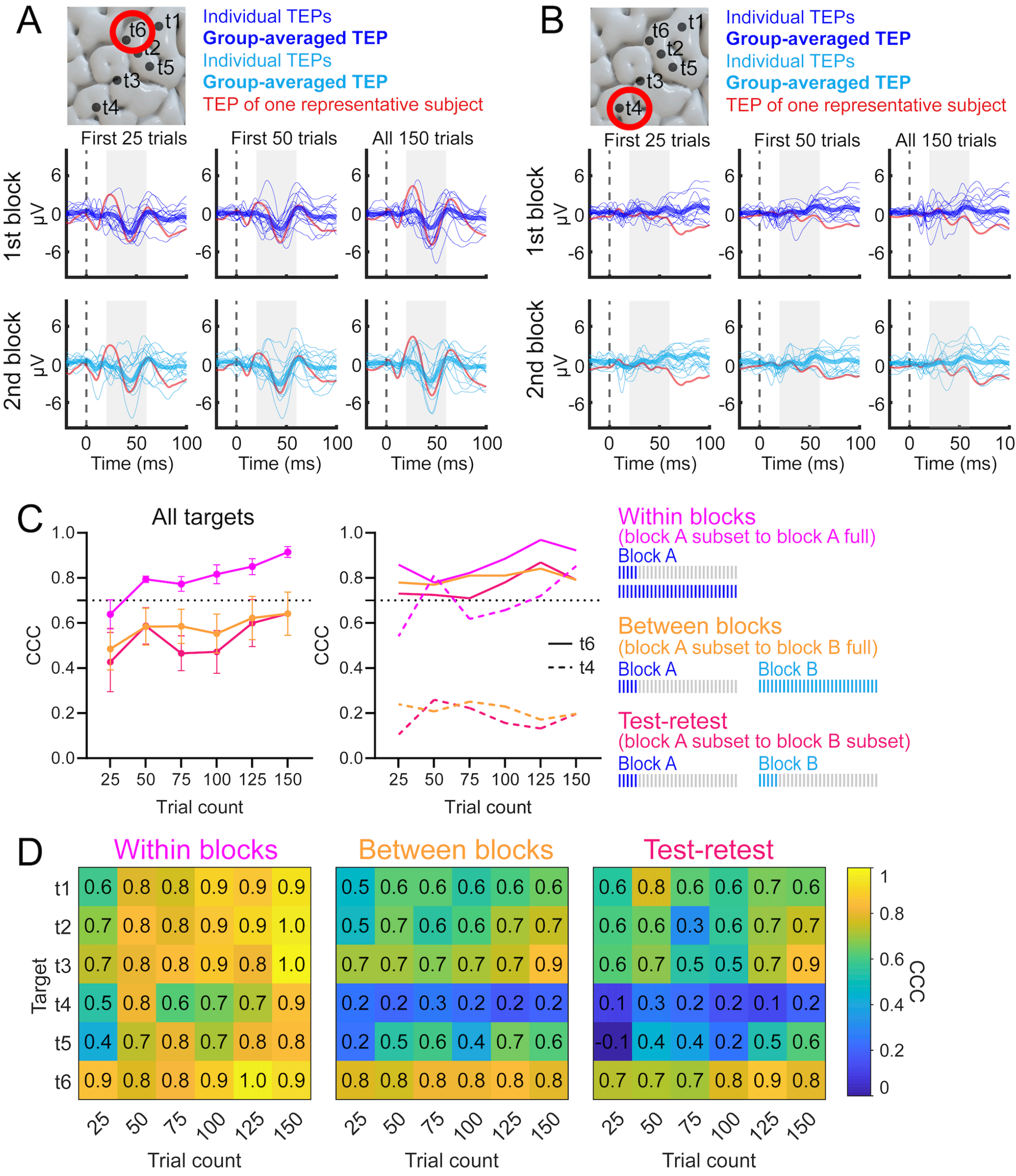


Figure 4 - Effects of trial count on EL-TEP reliability. A-B) Each thin trace represents the TEP of one subject, averaged over electrodes of a medium size ROI. Thick traces represent

the average TEP across subjects. Crimson color highlights TEPs from one representative subject. A) Example TEP traces for a reliable target (t6). B) Example TEPs for a non-reliable target (t4). TEPs for other targets are shown in Fig S17. C) Number of trials needed to produce reliable EL-TEPs. Left subplot shows the average CCC across dlPFC targets and the right subplot shows targets t6 and t4. Blue lines represent reliability within blocks, i.e. each subset of trials is compared to the full sample of 150 trials of the same block. Orange lines represent reliability between blocks, i.e. each subset of trials is compared to the full sample of 150 trials of a different block. Pink lines represent the test-retest reliability, i.e. each subset compared to the same subset of a different block. In the left subplot the dots denote mean CCC across targets ±SEM. D) Within-block, between-block and test-retest reliability across trial count and target.

4. Discussion

Summary of results. In the current study, we investigated how reliability of the prefrontal EL-TEP varied as a function of both stimulation target and analytic parameters. We applied single pulse TMS to six clinically-relevant targets within the dlPFC. We hypothesized that due to larger muscle artifacts observed after stimulation of more anterior targets (4), EL-TEPs would be more reliable after stimulation of posterior and medial dlPFC targets. Regarding the effect of analytic parameters (region of interest, time window, quantification method, trial count) on EL-TEP reliability, we hypothesized that later time windows within the early TEP, larger ROIs, source-space TEPs rather than sensor-space TEPs, and more trials would all increase reliability. Our main findings are summarized as follows: 1) The medial dlPFC TMS target was most reliable, while the most anterior dlPFC target was least reliable (Fig 2C); 2) targets in both the anterior and posterior dlPFC were reliable after optimizing analysis parameters (Fig 2D); 3) analyzing EL-TEPs using peak-to-peak quantification, longer and later time windows, and sensor-space EL-TEPs strengthened EL-TEP reliability (Fig 3); 4) ROI size did not appear to have a strong influence on EL-TEP reliability; 5) reliable EL-TEPs can be achieved with only 25 trials when stimulating the medial dlPFC. Together, these findings highlight that, a) EL-TEPs can be reliable, particularly if stimulating a medial prefrontal target; b) there is an urgent need to fine-tune analytic parameters to optimize reliability; c) ROI size used for analysis may not be as important for achieving a reliable EL-TEP biomarker; and d) the number of trials needed for obtaining a reliable EL-TEP may not be as high as previously assumed.

Effect of stimulation target on reliability. How reliable are TMS-EEG measurements of excitability (EL-TEP) when stimulating the dlPFC? Our current findings underscore the significance of target selection, as it emerged as a crucial factor to obtaining highly reliable

EL-TEPs. Specifically, we observed that with a medial target (t6), reliable EL-TEPs were obtained regardless of the chosen ROI size and time window (CCC from 0.72 to 0.89 for sensor-space peak-to-peak quantification). The most anterior site (t4) exhibited unreliable EL-TEPs irrespective of ROI size and time window (CCC from -0.085 to 0.45 for sensor-space peak-to-peak quantification), but another anterior (albeit less so) site (t3) exhibited reliable EL-TEPs for some parameter combinations (CCC from 0.56 to 0.90 for sensor-space peak-to-peak quantification). To our knowledge this is the first study evaluating differences in reliability of TEP across dlPFC subregions. There are a few earlier prefrontal TEP reliability studies, however. Kerwin *et al.* (17) reported low reliability of CCC < 0.6 for early TEPs (N40 peak) across blocks in a given day. Lioumis *et al.* (47) reported high Pearson correlations of between 0.88 to 0.96 for early dlPFC TEP peaks across separate days. These inconsistencies across the two papers may be due to several factors, including differences in noise masking, which is now common in the field (53), and differences in pre- and post-processing. In the current study, we used the newest sensory masking technique which involves auditory masking, foam separator between TMS coil and the scalp and over-the-ear protection (62). Future TEP studies should continue to apply the newest experimental and analytic techniques to reduce off-target sensory effects and continue to optimize EL-TEP reliability by exploring the effect of EEG amplifier, TMS coil type, pulse waveform, stimulation intensity, variations in TMS-EEG preprocessing, demographics (15). Finally, it is important to systematically investigate the reliability of EL-TEPs in other brain regions outside of dlPFC. In summary, we observed higher within-session EL-TEP reliability after TMS to the medial dlPFC, and that other dlPFC targets can produce reliable EL-TEPs after appropriate analytic considerations.

Relationship between EL-TEP *amplitude, SNR,* and *reliability* across stimulation targets. We recently reported that posteromedial targets exhibited larger *amplitude* EL-TEPs compared to anteromedial sites (4). We would like to point out that *amplitude* does not necessarily co-fluctuate with *reliability*. Case in point, while our previous work demonstrated anterior dlPFC targets to elicit smaller amplitude EL-TEP responses (4), in this study we report some anterior sites with low reliability (t4) and others with high reliability (t3). Thus, both amplitude and reliability should be assessed when considering applying it to study the neural effect of an intervention (15,63–66).

We also hypothesized that within-block signal-to-noise would positively correlate with across-block reliability. However, our supplemental analysis did not reveal strong trends supporting this hypothesis (Fig S15), suggesting that factors contributing to within-block variance (e.g. trial-to-trial noise in EEG signals) may not be the main factors contributing to across-block degradation in reliability. Our trial subset analyses (Fig 4) indicating high within-block reliability in EL-TEP measures with just 25-50 trials further support this conclusion. Instead, longer time scale factors such as slow changes in brain state, drifts in stimulation coil positioning, or changes in electrode impedances may be more dominant factors and warrant further investigation in future experiments focused on improving EL-TEP reliability.

Contribution of analytic parameters across stimulation targets. We observed that the size of the grouping of prefrontal electrodes used for analysis (ROI) did not significantly impact reliability of the EL-TEP (Fig 3C-F). We found that, on the contrary, the time window used in the analysis did impact reliability: earlier / narrower time windows (20-40 ms, 20-50 ms) produced EL-TEPs with lower reliability for all targets except t1 (most posterior target), whereas later (30-60 ms) and longer windows (20-60 ms) produced EL-TEPs with higher reliability (Fig 3G-J). The choice of amplitude quantification method also influenced reliability: peak-to-peak was more reliable than the mean of the absolute value (Fig 3K-N), although both had moderate reliability. Our metric for absolute amplitude, the mean of the absolute values within a time window, is equivalent to the common area-under-the-curve (AUC) metrics of TEP amplitude (7,35,67), except with a normalization for window length. Our method for peak-to-peak calculation, similar to that used in (68), differs from some other approaches that rely on temporal peak detection (e.g. (69,70)). Therefore, it should not be assumed that our reported reliability of peak-to-peak amplitude would be equal to different peak-to-peak metrics used in other EEG studies. When specific EL-TEP peaks are more well-defined, other methods for peak-to-peak calculation should be examined for reliability. Finally, we found that source localized EL-TEPs were less reliable than sensor-space EL-TEPs (Fig 3O-R). It is worth noting that for targets such as t3 (anterior target) where EL-TEP amplitude is low, the choice of quantification parameters may substantially impact reliability. Notably, the optimal parameter combination for achieving higher reliability varied for each target, as illustrated in Figure 2B. If researchers opt for a fixed parameter set across multiple stimulation targets, based on our results we recommend utilizing peak-to-peak amplitude quantification with a time window of either 20-60 ms or 30-60 ms in sensor space to maximize reliability (Fig 3B). However, from this work we find that to maximize target-specific reliability for EL-TEPs, each target may require different analytic parameters. When evaluating new stimulation targets, a distinct early data collection stage focused on reliability testing will help inform choice of optimal stimulation target to maximize EL-TEP reliability. This separate, early stage reliability evaluation could also be used to optimize and fix analytic parameters prior to primary data collection, allowing the opportunity to follow best practices in brain imaging research including pre-registering details of study design and outcome metrics.

Influence of trial number on reliability. Here, we investigated the number of single pulses of TMS (trials) required to obtain a reliable prefrontal EL-TEP. To address this, we subdivided the data into different sets of *n*-trials and compared reliability across resulting EL-TEPs. An important note about our methods is that we preprocessed each set of *n*-trials separately in order to mimic the preprocessing that would be done in an experiment that only collected *n* number of trials. We found that for all except the most anterior dlPFC target (t4), averaging only 50 trials to compute the EL-TEP yielded a similar EL-TEP response compared to the EL-TEP from 150 trials (reliability *within* blocks; Fig 4D). Moreover, for three of the targets (t1, t2 and t3), a moderate to high consistency of EL-TEPs over a longer time scale (reliability *between* blocks) was obtained with only 50 trials (Fig 4D). When comparing less than 150 trials across both blocks (*test-retest* reliability), two targets produced high to moderate reliability with 125 trials. Notably, for the most medial target (t6), as few as 25 trials were sufficient to obtain high reliability both within and between blocks, as well as high test-retest reliability (Fig 4A, C and D). For the least reliable target (t4), reliability between blocks and test-retest reliability were low regardless of the trial count (Fig 4B, C and D).

These results are somewhat inconsistent with Kerwin *et al*. (17) which reported that the early TEP (in this case, the ‘N40’) was only moderately reliable (CCC ~ 0.5) across experimental blocks, regardless of how many trials were used. However, several important differences between this study and (17) could account for this inconsistency: different EEG recording systems, different preprocessing pipelines, and slightly different stimulation targets. On the other hand, our results are compatible with the supplementary results described by Casarotto *et al*. (64). They estimated the minimum amplitude of the EL-TEPs that can be recovered from baseline EEG after averaging a certain number of trials, and the minimum number of trials required for signal reliability. They found that, if data has a good signal-to-noise ratio, as low as 20 trials are enough. Our results are also consistent with Lioumis *et al.* (47) in showing high reliability of the EL-TEP, though in that study they only examined 100 trials per condition and did not investigate what the minimum number of trials should be to obtain a reliable EL-TEP. Moreover, they used a different way of quantifying EL-TEPs compared to the current study. We recommend that future studies aiming to achieve reliable EL-TEPs from stimulation of the dlPFC should collect at least 50 trials per target. If requiring 25 or fewer trials, for example for real-time monitoring and closed-loop adaptive stimulation, our data suggest that the t6 target (in MNI coordinates x=-32, y=26, z=52) would be the optimal target. Regardless, reliability testing should be performed when using other equipment and experimental and analytic procedures.

Limitations. While this study presents several important findings, numerous points limit generalizability and warrant careful consideration. First, the reliability assessment was limited to within-session measurements, as data between different days were not collected in this study. Within-session reliability is useful to determine how stable EL-TEPs are prior to an intervention, and as such, how confident one can be that a pre/post-intervention change *on a single day* is attributed to the intervention. For longitudinal studies, this type of test-retest reliability investigation must be performed across days. Second, it is important to note that a highly reliable *signal* does not necessarily imply high *biological validity* (15). Coupling these noninvasive studies with invasive investigations (5,34,49) to determine the neural response contributing to the EL-TEP will help reveal whether the EL-TEP is *biologically valid*. Third, details of the TMS-EEG preprocessing pipeline can affect reliability, and future improvements in preprocessing methods may increase reliability in a manner that reduces sensitivity to a specific stimulation target or analytic parameters (Fig S16). Fourth, our results may not generalize to other TEP quantification metrics, such as methods based on peak detection (16) or other time windows. Fifth, this study primarily focused on a predetermined set of analytic parameters. Future work should explore reliability across a broader range of parameter combinations, encompassing more refined (and potentially personalized) time windows, an expanded (and potentially personalized) set of sensor-space and source-space ROIs, and features in the spectral domain, which could further enhance reliability. Finally, it is important to acknowledge the influence of the experimental setup and hardware employed in this study, including the TMS coil, EEG system, and the use of a robotic arm for accurate targeting. Consequently, even if the dlPFC targets are the same as those reported here, for each hardware/software setup it may be beneficial for researchers to perform a reliability analysis to fine-tune experimental and analytic parameters and maximize TEP reliability (15).

Conclusions. We characterized the test-retest reliability of the EL-TEP after stimulation of different dlPFC subregions. We found that EL-TEP reliability was sensitive to stimulation target in the dlPFC, with medial and anterior targets producing the most and least reliable EL-TEPs, respectively. The EL-TEP was sensitive to the quantification method, where peak-to-peak measurement in the 20-60 ms time window measured in sensor space typically provided the most reliable EL-TEPs. Finally, reliable EL-TEPs were produced after only 25 single TMS pulses to the medial dlPFC. These results have important implications for the EL-TEP as a measure of cortical excitability and potential clinical biomarker.

Acknowledgements. We extend our deep gratitude to all of our research participants. We would also like to acknowledge the generous contributions of the members of the Personalized Neurotherapeutics Laboratory for helpful feedback on the manuscript and throughout the course of the study. We acknowledge Recep Ozdemir for his valuable assistance in implementing the original cosine similarity code adapted for this study. This research was supported by the National Institute of Mental Health under award number R01MH126639, R01MH129018, and a Burroughs Wellcome Fund Career Award for Medical Scientists (CJK). JG was supported by personal grants from Orion Research Foundation, the Finnish Medical Foundation and Emil Aaltonen Foundation. JMR was supported by the Department of Veterans Affairs Office of Academic Affiliations Advanced Fellowship Program in Mental Illness Research and Treatment, the Medical Research Service of the Veterans Affairs Palo Alto Health Care System and the Department of Veterans Affairs Sierra-Pacific Data Science Fellowship.

Declaration of Interest. Corey Keller holds equity in Alto Neuroscience, Inc. All other authors have nothing to disclose.

References

1. Lefaucheur J-P, Aleman A, Baeken C, Benninger DH, Brunelin J, Di Lazzaro V, *et al.* (2020): Evidence-based guidelines on the therapeutic use of repetitive transcranial magnetic stimulation (rTMS): An update (2014–2018). *Clin Neurophysiol* 131: 474–528.
2. Tremblay S, Rogasch NC, Premoli I, Blumberger DM, Casarotto S, Chen R, *et al.* (2019): Clinical utility and prospective of TMS-EEG. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol* 130: 802–844.
3. Casarotto S, Lauro LJR, Bellina V, Casali AG, Rosanova M, Pigorini A, *et al.* (2010): EEG Responses to TMS Are Sensitive to Changes in the Perturbation Parameters and Repeatable over Time. *PLOS ONE* 5: e10281.
4. Gogulski J, Cline CC, Ross JM, Truong J, Sarkar M, Parmigiani S, Keller CJ (2023, January 20): Mapping cortical excitability in the human dorsolateral prefrontal cortex. bioRxiv, p 2023.01.20.524867.
5. Wang JB, Bruss JE, Oya H, Uitermarkt BD, Trapp NT, Gander PE, *et al.* (2022, January 21): Effects of transcranial magnetic stimulation on the human brain recorded with intracranial electrocorticography: First-in-human study. bioRxiv, p 2022.01.18.476811.
6. Belardinelli P, König F, Liang C, Premoli I, Desideri D, Müller-Dahlhaus F, *et al.* (2021): TMS-EEG signatures of glutamatergic neurotransmission in human cortex [no. 1]. *Sci Rep* 11: 8159.
7. Voineskos D, Blumberger DM, Zomorrodi R, Rogasch NC, Farzan F, Foussias G, *et al.* (2019): Altered Transcranial Magnetic Stimulation–Electroencephalographic Markers of Inhibition and Excitation in the Dorsolateral Prefrontal Cortex in Major Depressive Disorder. *Biol Psychiatry* 85: 477–486.
8. Eshel N, Keller CJ, Wu W, Jiang J, Mills-Finnerty C, Huemer J, *et al.* (2020): Global connectivity and local excitability changes underlie antidepressant effects of

repetitive transcranial magnetic stimulation. *Neuropsychopharmacology* 45: 1018–1025.

9. Voineskos D, Blumberger DM, Rogasch NC, Zomorrodi R, Farzan F, Foussias G, *et al.* (2021): Neurophysiological effects of repetitive transcranial magnetic stimulation (rTMS) in treatment resistant depression. *Clin Neurophysiol* 132: 2306–2316.
10. Fox MD, Buckner RL, White MP, Greicius MD, Pascual-Leone A (2012): Efficacy of transcranial magnetic stimulation targets for depression is related to intrinsic functional connectivity with the subgenual cingulate. *Biol Psychiatry* 72: 595–603.
11. Herwig U, Lampe Y, Juengling FD, Wunderlich A, Walter H, Spitzer M, Schönfeldt-Lecuona C (2003): Add-on rTMS for treatment of depression: a pilot study using stereotaxic coil-navigation according to PET data. *J Psychiatr Res* 37: 267–275.
12. Cash RFH, Weigand A, Zalesky A, Siddiqi SH, Downar J, Fitzgerald PB, Fox MD (2021): Using Brain Imaging to Improve Spatial Targeting of Transcranial Magnetic Stimulation for Depression. *Biol Psychiatry* 90: 689–700.
13. Herbsman T, Avery D, Ramsey D, Holtzheimer P, Wadjik C, Hardaway F, *et al.* (2009): More lateral and anterior prefrontal coil location is associated with better repetitive transcranial magnetic stimulation antidepressant response. *Biol Psychiatry* 66: 509–515.
14. Mäki H, Ilmoniemi RJ (2011): Projecting out muscle artifacts from TMS-evoked EEG. *NeuroImage* 54: 2706–2710.
15. Parmigiani S, Ross JM, Cline CC, Minasi CB, Gogulski J, Keller CJ (2022): Reliability and Validity of Transcranial Magnetic Stimulation–Electroencephalography Biomarkers. *Biol Psychiatry Cogn Neurosci Neuroimaging* 0. https://doi.org/10.1016/j.bpsc.2022.12.005
16. Lioumis P, Kičić D, Savolainen P, Mäkelä JP, Kähkönen S (2009): Reproducibility of TMS—Evoked EEG responses. *Hum Brain Mapp* 30: 1387–1396.
17. Kerwin LJ, Keller CJ, Wu W, Narayan M, Etkin A (2018): Test-retest reliability of transcranial magnetic stimulation EEG evoked potentials. *Brain Stimulat* 11: 536–

544.

18. Bertazzoli G, Esposito R, Mutanen TP, Ferrari C, Ilmoniemi RJ, Miniussi C, Bortoletto M (2021): The impact of artifact removal approaches on TMS–EEG signal. *NeuroImage* 239: 118272.
19. Biabani M, Fornito A, Mutanen T, Morrow J, Rogasch N (2019): Sensory contamination in TMS-EEG recordings: Can we isolate TMS-evoked neural activity? *Brain Stimul Basic Transl Clin Res Neuromodulation* 12: 473.
20. Conde V, Tomasevic L, Akopian I, Stanek K, Saturnino GB, Thielscher A, *et al.* (2019): The non-transcranial TMS-evoked potential is an inherent source of ambiguity in TMS-EEG studies. *NeuroImage* 185: 300–312.
21. ter Braack EM, de Vos CC, van Putten MJAM (2015): Masking the Auditory Evoked Potential in TMS–EEG: A Comparison of Various Methods. *Brain Topogr* 28: 520–528.
22. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, *et al.* (2003): The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* 54: 573–583.
23. Rossi S, Hallett M, Rossini PM, Pascual-Leone A (2011): Screening questionnaire before TMS: an update. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol* 122: 1686.
24. Rossi S, Hallett M, Rossini PM, Pascual-Leone A (2009): Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin Neurophysiol* 120: 2008–2039.
25. ter Braack EM, de Vos CC, van Putten MJAM (2015): Masking the Auditory Evoked Potential in TMS-EEG: A Comparison of Various Methods. *Brain Topogr* 28: 520–528.
26. Rossi S, Antal A, Bestmann S, Bikson M, Brewer C, Brockmöller J, *et al.* (2021): Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines. *Clin*

Neurophysiol 132: 269–306.

27. Rossini PM, Burke D, Chen R, Cohen LG, Daskalakis Z, Di Iorio R, *et al.* (2015): Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee. *Clin Neurophysiol* 126: 1071–1107.
28. Stokes MG, Chambers CD, Gould IC, English T, McNaught E, McDonald O, Mattingley JB (2007): Distance-adjusted motor threshold for transcranial magnetic stimulation. *Clin Neurophysiol* 118: 1617–1625.
29. Pridmore S, Fernandes Filho JA, Nahas Z, Liberatos C, George MS (1998): Motor threshold in transcranial magnetic stimulation: a comparison of a neurophysiological method and a visualization of movement method. *J ECT* 14: 25–27.
30. Rajkowska G, Goldman-Rakic PS (1995): Cytoarchitectonic definition of prefrontal areas in the normal human cortex: II. Variability in locations of areas 9 and 46 and relationship to the Talairach Coordinate System. *Cereb Cortex N Y N 1991* 5: 323–337.
31. Weigand A, Horn A, Caballero R, Cooke D, Stern AP, Taylor SF, *et al.* (2018): Prospective Validation That Subgenual Connectivity Predicts Antidepressant Efficacy of Transcranial Magnetic Stimulation Sites. *Biol Psychiatry* 84: 28–37.
32. Dunås T, Wåhlin A, Ambarki K, Zarrinkoob L, Malm J, Eklund A (2017): A Stereotactic Probabilistic Atlas for the Major Cerebral Arteries. *Neuroinformatics* 15: 101–110.
33. Leodori G, Rocchi L, Mancuso M, De Bartolo MI, Baione V, Costanzo M, *et al.* (2022): The effect of stimulation frequency on transcranial evoked potentials. *Transl Neurosci* 13: 211–217.
34. Keller CJ, Huang Y, Herrero JL, Fini ME, Du V, Lado FA, *et al.* (2018): Induction and Quantification of Excitability Changes in Human Cortical Networks. *J Neurosci Off J Soc Neurosci* 38: 5384–5398.
35. Ross JM, Sarkar M, Keller CJ (2022): Experimental suppression of transcranial magnetic

stimulation-electroencephalography sensory potentials. *Hum Brain Mapp*. https://doi.org/10.1002/hbm.25990

36. Cline CC, Lucas MV, Sun Y, Menezes M, Etkin A (2021): Advanced Artifact Removal for Automated TMS-EEG Data Processing. *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)* 1039–1042.
37. Cline CC, Lucas MV, Sun Y, Menezes M, Etkin A (2021): Advanced Artifact Removal for Automated TMS-EEG Data Processing. *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)* 1039–1042.
38. Pion-Tonachini L, Kreutz-Delgado K, Makeig S (2019): ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* 198: 181–197.
39. Mutanen TP, Metsomaa J, Liljander S, Ilmoniemi RJ (2018): Automatic and robust noise suppression in EEG and MEG: The SOUND algorithm. *NeuroImage* 166: 135–151.
40. Himberg J, Hyvärinen A, Esposito F (2004): Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 22: 1214–1222.
41. Schambra H, Ogden RT, Martínez-Hernández I, Lin X, Chang YB, Rahman A, *et al.* (2015): The reliability of repeated TMS measures in older adults and in patients with subacute and chronic stroke. *Front Cell Neurosci* 9. Retrieved June 1, 2023, from https://www.frontiersin.org/articles/10.3389/fncel.2015.00335
42. Shrout PE, Fleiss JL (19790101): Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 86: 420.
43. Lin LI-K (1989): A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45: 255–268.
44. Ozdemir RA, Tadayon E, Boucher P, Sun H, Momi D, Ganglberger W, *et al.* (2021): Cortical responses to noninvasive perturbations enable individual brain fingerprinting. *Brain Stimulat* 14: 391–403.
45. Luck SJ, Stewart AX, Simmons AM, Rhemtulla M (2021): Standardized measurement

error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology* 58. https://doi.org/10.1111/psyp.13793

46. Tervo AE, Nieminen JO, Lioumis P, Metsomaa J, Souza VH, Sinisalo H, *et al.* (2021): *Closed-Loop Optimization of Transcranial Magnetic Stimulation with Electroencephalography Feedback*. Bioengineering. https://doi.org/10.1101/2021.08.31.458148
47. Lioumis P, Kicić D, Savolainen P, Mäkelä JP, Kähkönen S (2009): Reproducibility of TMS-Evoked EEG responses. *Hum Brain Mapp* 30: 1387–1396.
48. Kähkönen S, Komssi S, Wilenius J, Ilmoniemi RJ (2005): Prefrontal transcranial magnetic stimulation produces intensity-dependent EEG responses in humans. *NeuroImage* 24: 955–960.
49. Huang Y, Hajnal B, Entz L, Fabó D, Herrero JL, Mehta AD, Keller CJ (2019): Intracortical Dynamics Underlying Repetitive Stimulation Predicts Changes in Network Connectivity. *J Neurosci Off J Soc Neurosci* 39: 6122–6135.
50. Biabani M, Fornito A, Mutanen TP, Morrow J, Rogasch NC (2019): Characterizing and minimizing the contribution of sensory inputs to TMS-evoked potentials. *Brain Stimulat* 12: 1537–1552.
51. Ross JM, Ozdemir RA, Lian SJ, Fried PJ, Schmitt EM, Inouye SK, *et al.* (2022): A structured ICA-based process for removing auditory evoked potentials. *Sci Rep* 12: 1391.
52. Freedberg M, Reeves JA, Hussain SJ, Zaghloul KA, Wassermann EM (2020): Identifying site- and stimulation-specific TMS-evoked EEG potentials using a quantitative cosine similarity metric. *PloS One* 15: e0216185.
53. Rocchi L, Di Santo A, Brown K, Ibáñez J, Casula E, Rawji V, *et al.* (2021): Disentangling EEG responses to TMS due to cortical and peripheral activations. *Brain Stimulat* 14: 4–18.
54. Michel CM, Murray MM (2012): Towards the utilization of EEG as a brain imaging tool. *NeuroImage* 61: 371–385.

55. Thielscher A, Antunes A, Saturnino GB (2015): Field modeling for transcranial magnetic stimulation: A useful tool to understand the physiological effects of TMS? *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 222–225.

56. Gramfort A, Papadopoulo T, Olivi E, Clerc M (2010): OpenMEEG: opensource software for quasistatic bioelectromagnetics. *Biomed Eng OnLine* 9: 45.

57. Kybic J, Clerc M, Abboud T, Faugeras O, Keriven R, Papadopoulo T (2005): A common formalism for the Integral formulations of the forward EEG problem. *IEEE Trans Med Imaging* 24: 12–28.

58. Brainstorm: a user-friendly application for MEG/EEG analysis: Computational Intelligence and Neuroscience: Vol 2011, No null (n.d.): Retrieved February 14, 2023, from https://dl-acm-org.stanford.idm.oclc.org/doi/abs/10.1155/2011/879716

59. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, *et al.* (2016): A multi-modal parcellation of human cerebral cortex [no. 7615]. *Nature* 536: 171–178.

60. Freedberg M, Reeves JA, Hussain SJ, Zaghloul KA, Wassermann EM (2020): Identifying site- and stimulation-specific TMS-evoked EEG potentials using a quantitative cosine similarity metric. *PloS One* 15: e0216185.

61. Rosanova M, Casali A, Bellina V, Resta F, Mariotti M, Massimini M (2009): Natural frequencies of human corticothalamic circuits. *J Neurosci Off J Soc Neurosci* 29: 7679–7685.

62. Ross JM, Sarkar M, Keller CJ (2022): Experimental suppression of transcranial magnetic stimulation-electroencephalography sensory potentials. *Hum Brain Mapp* 43: 5141–5153.

63. Califf RM (2018): Biomarker definitions and their applications. *Exp Biol Med Maywood NJ* 243: 213–221.

64. Casarotto S, Fecchio M, Rosanova M, Varone G, D'Ambrosio S, Sarasso S, *et al.* (2021): *The Rt-TEP Tool: Real-Time Visualization of TMS-Evoked Potential to*

Maximize Cortical Activation and Minimize Artifacts. Neuroscience. https://doi.org/10.1101/2021.09.15.460488

65. Kallioniemi E, Daskalakis ZJ (2022): Identifying novel biomarkers with TMS-EEG - Methodological possibilities and challenges. *J Neurosci Methods* 377: 109631.
66. Cao K-X, Ma M-L, Wang C-Z, Iqbal J, Si J-J, Xue Y-X, Yang J-L (2021): TMS-EEG: An emerging tool to study the neurophysiologic biomarkers of psychiatric disorders. *Neuropharmacology* 197: 108574.
67. Ross JM, Santarnecchi E, Lian SJ, Fong TG, Touroutoglou A, Cavallari M, *et al.* (2023): Neurophysiologic predictors of individual risk for post-operative delirium after elective surgery. *J Am Geriatr Soc* 71: 235–244.
68. Casula EP, Bertoldo A, Tarantino V, Maiella M, Koch G, Rothwell JC, *et al.* (2017): TMS-evoked long-lasting artefacts: A new adaptive algorithm for EEG signal correction. *Clin Neurophysiol* 128: 1563–1574.
69. Opie GM, Rogasch NC, Goldsworthy MR, Ridding MC, Semmler JG (2017): Investigating TMS–EEG Indices of Long-Interval Intracortical Inhibition at Different Interstimulus Intervals. *Brain Stimulat* 10: 65–74.
70. Rogasch NC, Daskalakis ZJ, Fitzgerald PB (2015): Cortical inhibition of distinct mechanisms in the dorsolateral prefrontal cortex is related to working memory performance: A TMS–EEG study. *Cortex* 64: 68–77.

Supplementary Material

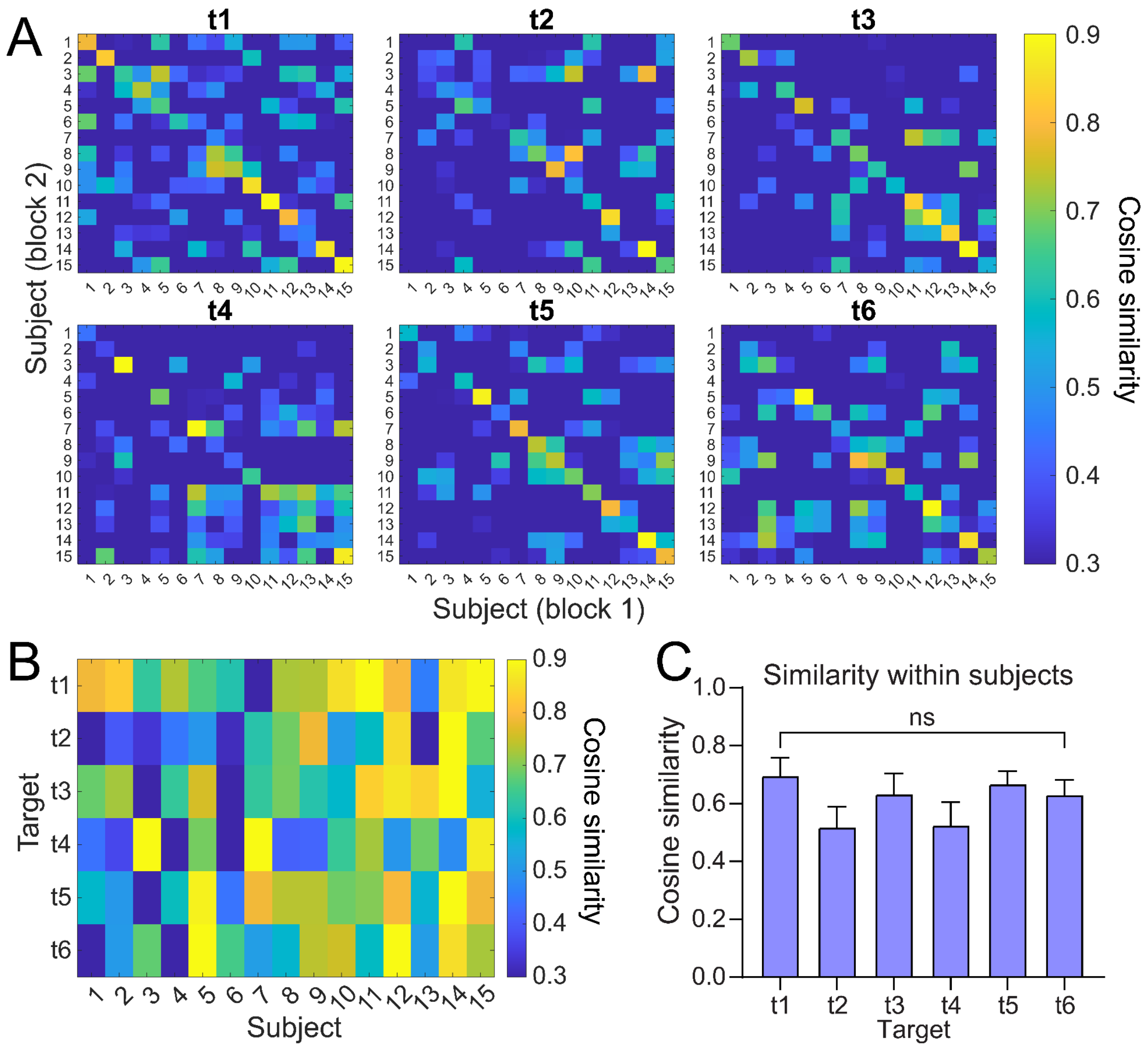


Figure S1 - Cosine similarity of EL-TEPs. Cosine similarity comparing the first to the second block, calculated within a 20-60 ms time window using all EEG channels. A) Cosine similarity across all subjects. B) Cosine similarity within subjects. C) Within-subject similarity for each target. Repeated measures ANOVA showed no significant differences between the stimulation targets ($F_{3.33,46.65}$=1.75, P=0.16). In C the bars denote mean cosine similarity across subjects ±SEM.

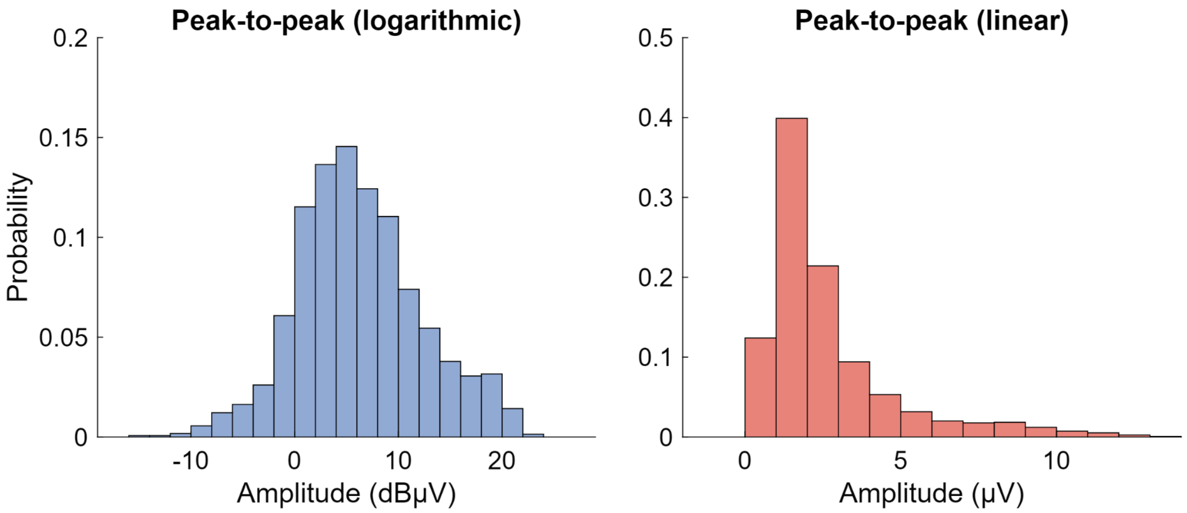


Fig S2 - Histograms of sensor space peak-to-peak EL-TEP amplitudes in logarithmic (dBμV) and linear (μV) values. Logarithmic distribution passed D'Agnostino & Pearson normality test (*P*=0.23), whereas the linear distribution did not (*P*<0.0001). Moreover, the linear distribution passed D'Agnostino & Pearson lognormality test (*P*=0.23)

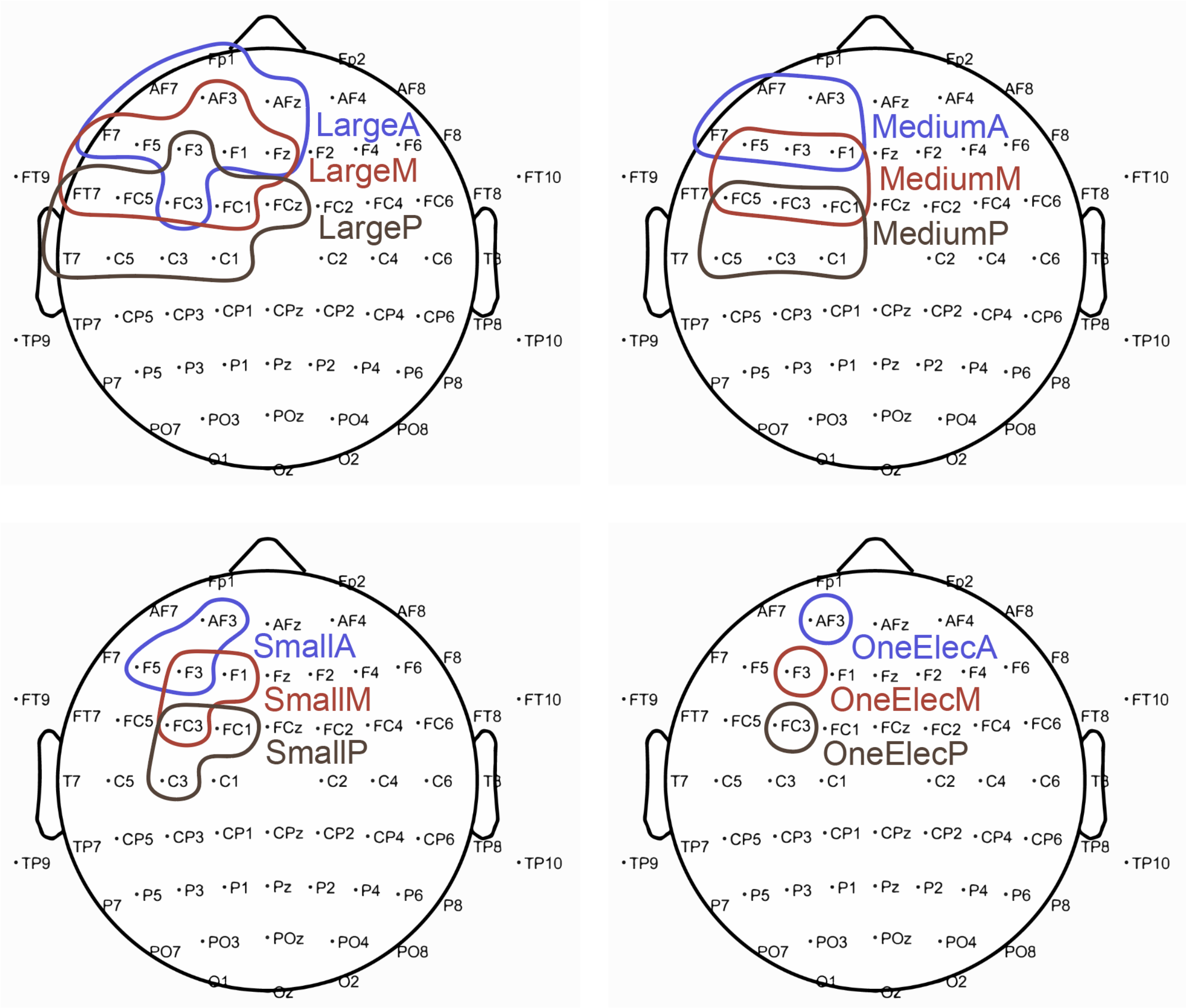


Figure S3 - Locations of the shifted ROIs.

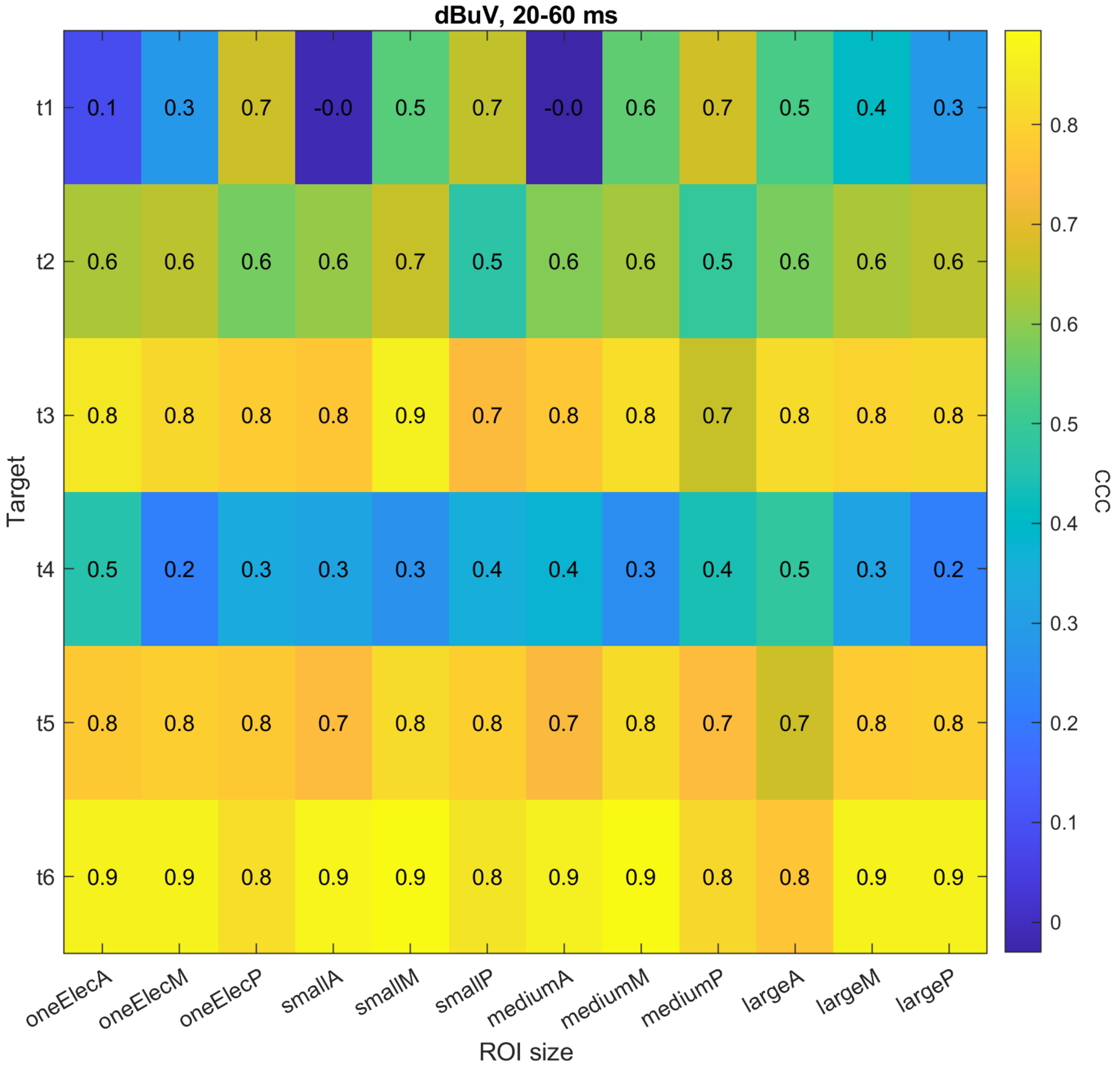


Figure S4 - Effect of shifting the ROI on EL-TEP reliability. To examine whether the spatial positioning of the ROIs affects the results, we additionally shifted each ROIs in AP direction (three different positions for each ROI size) and extracted the TEPs and CCCs for these shifted ROIs. The heatmap shows CCCs for all shifted ROIs (peak-to-peak amplitude, 20-60 ms time window).

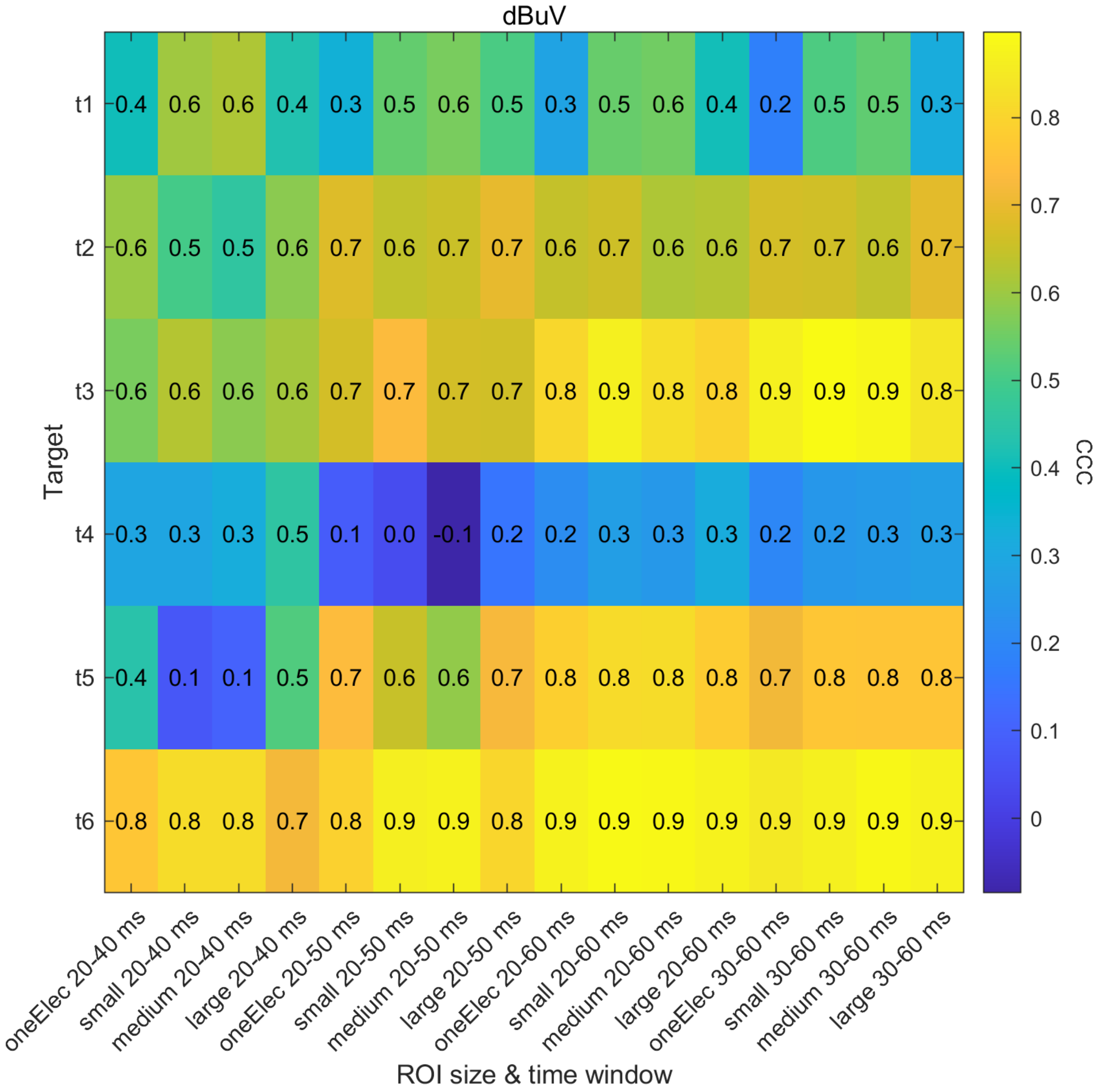


Fig S5 - Effect of time window and ROI size on reliability for peak-to-peak quantification (in dB units).

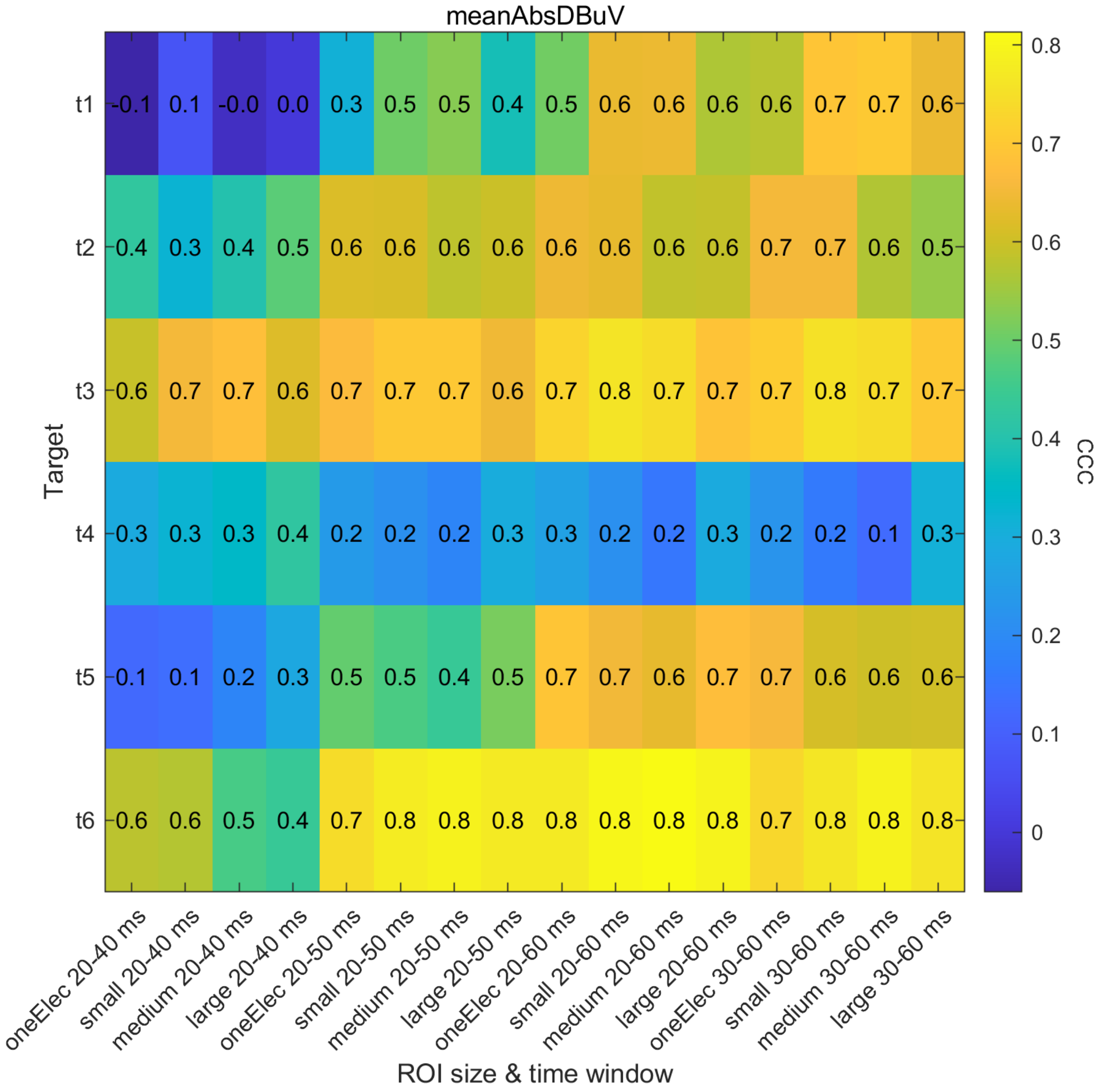


Fig S6 - Effect of time window and ROI size on reliability for mean(abs) quantification (in dB units).

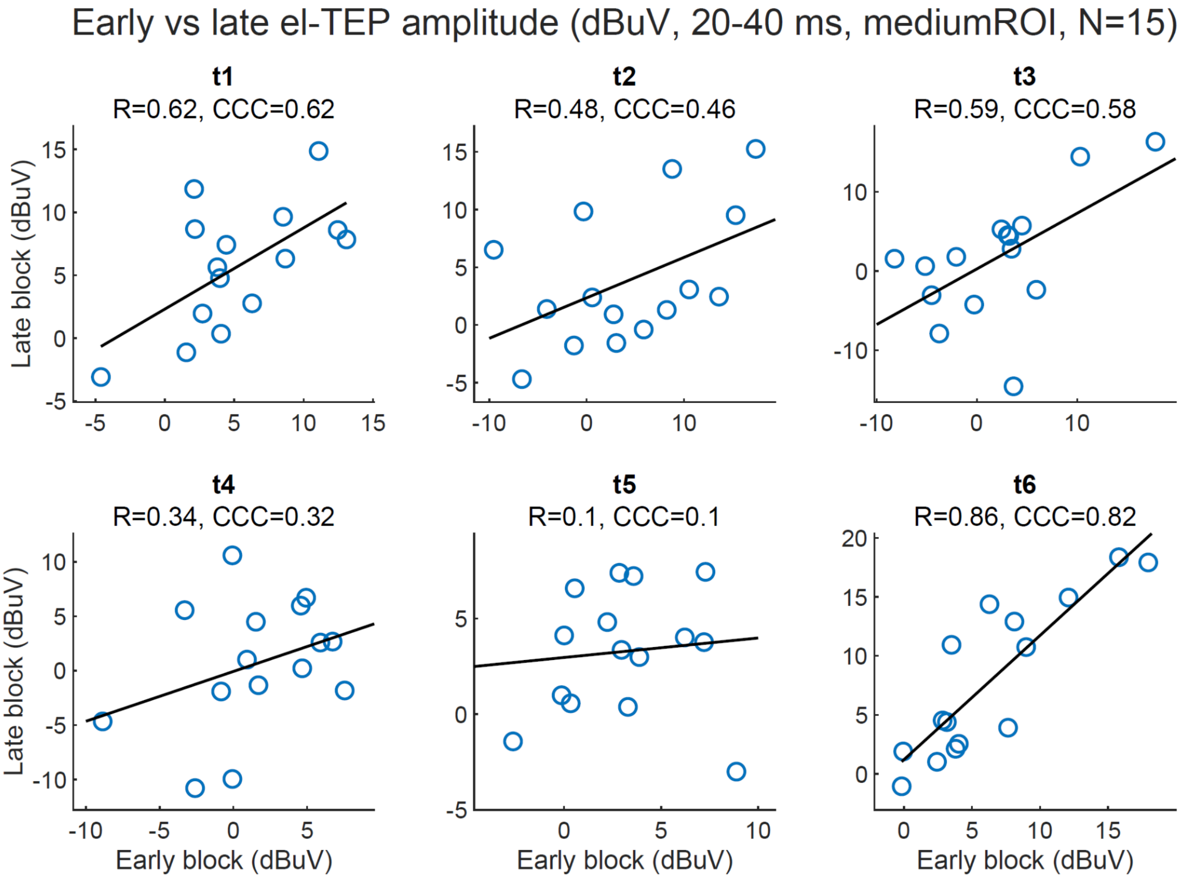


Figure S7 - EL-TEP amplitudes (medium ROI, 20-40 ms time window, peak-to-peak) of the first and the second experimental block. Each dot represents one subject (N=15).

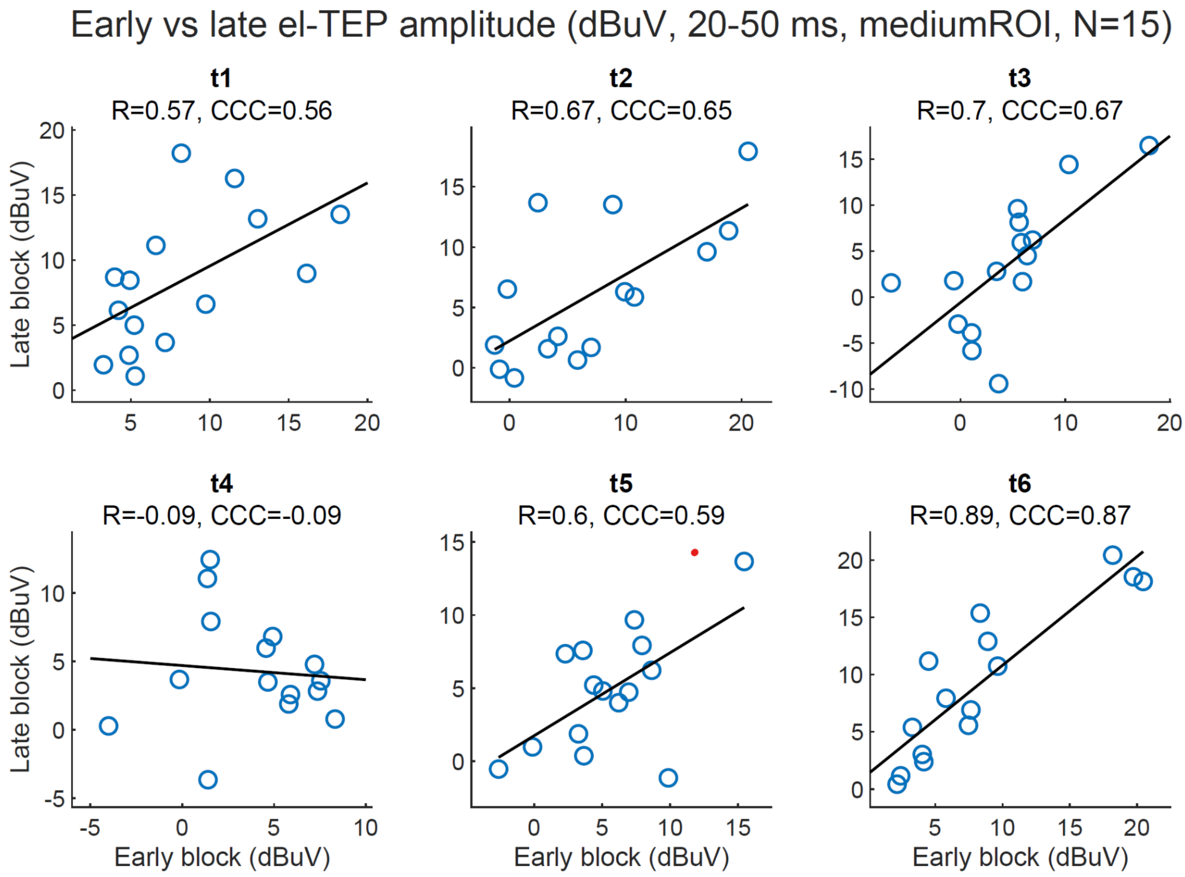


Figure S8 - EL-TEP amplitudes (medium ROI, 20-50 ms time window, peak-to-peak) of the first and the second experimental block. Each dot represents one subject (N=15).

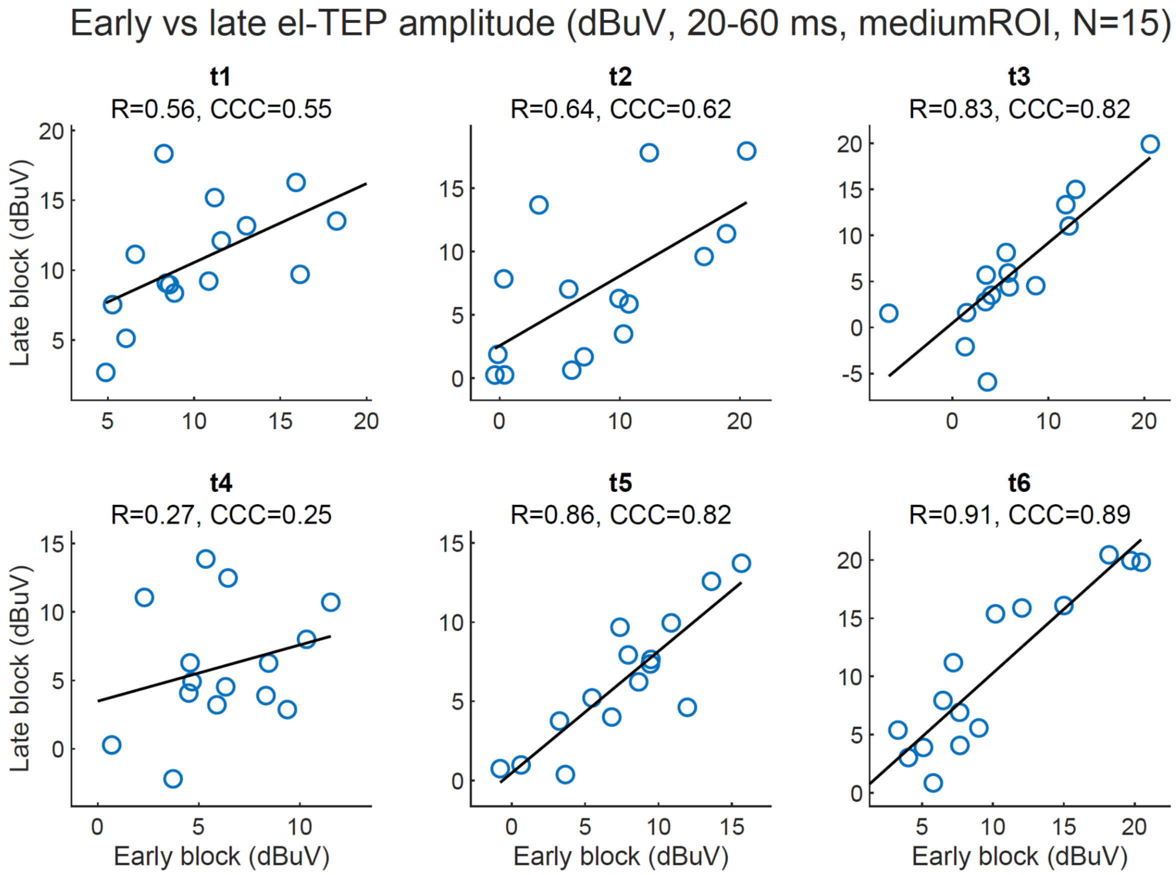


Figure S9 - EL-TEP amplitudes (medium ROI, 20-60 ms time window, peak-to-peak) of the first and the second experimental block. Each dot represents one subject (N=15).

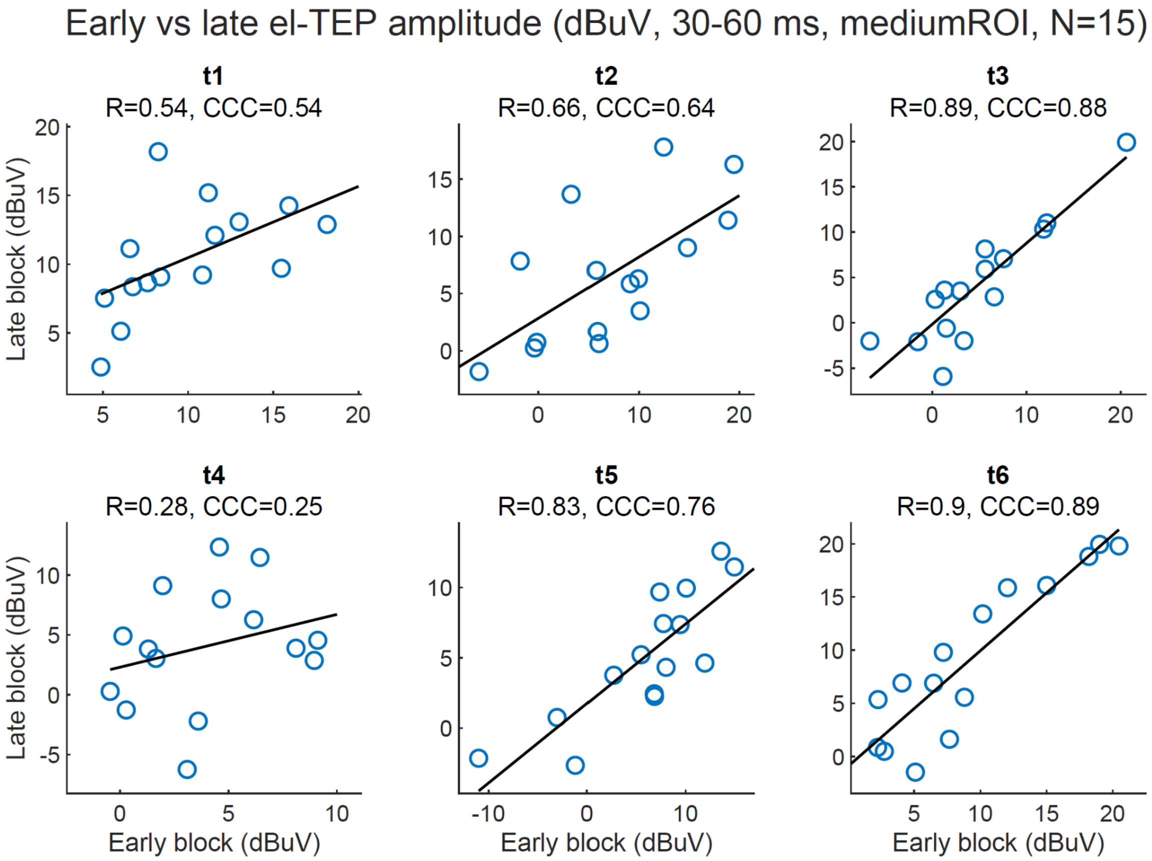


Figure S10 - EL-TEP amplitudes (medium ROI, 30-60 ms time window, peak-to-peak) of the first and the second experimental block. Each dot represents one subject (N=15).

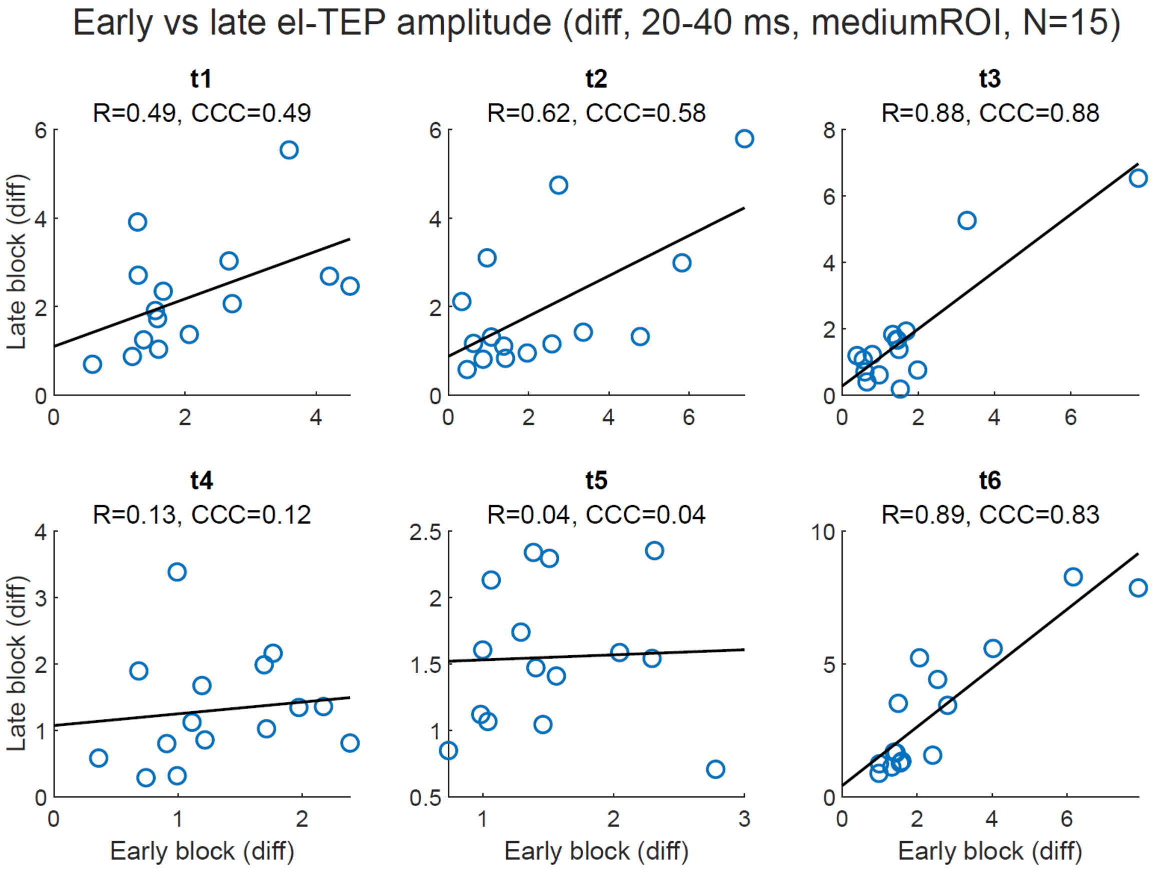


Figure S11 - EL-TEP amplitudes (medium ROI, 30-40 ms time window, peak-to-peak) of the first and the second experimental block using linear (µV) peak-to-peak values (instead of dBµV). Each dot represents one subject (N=15).

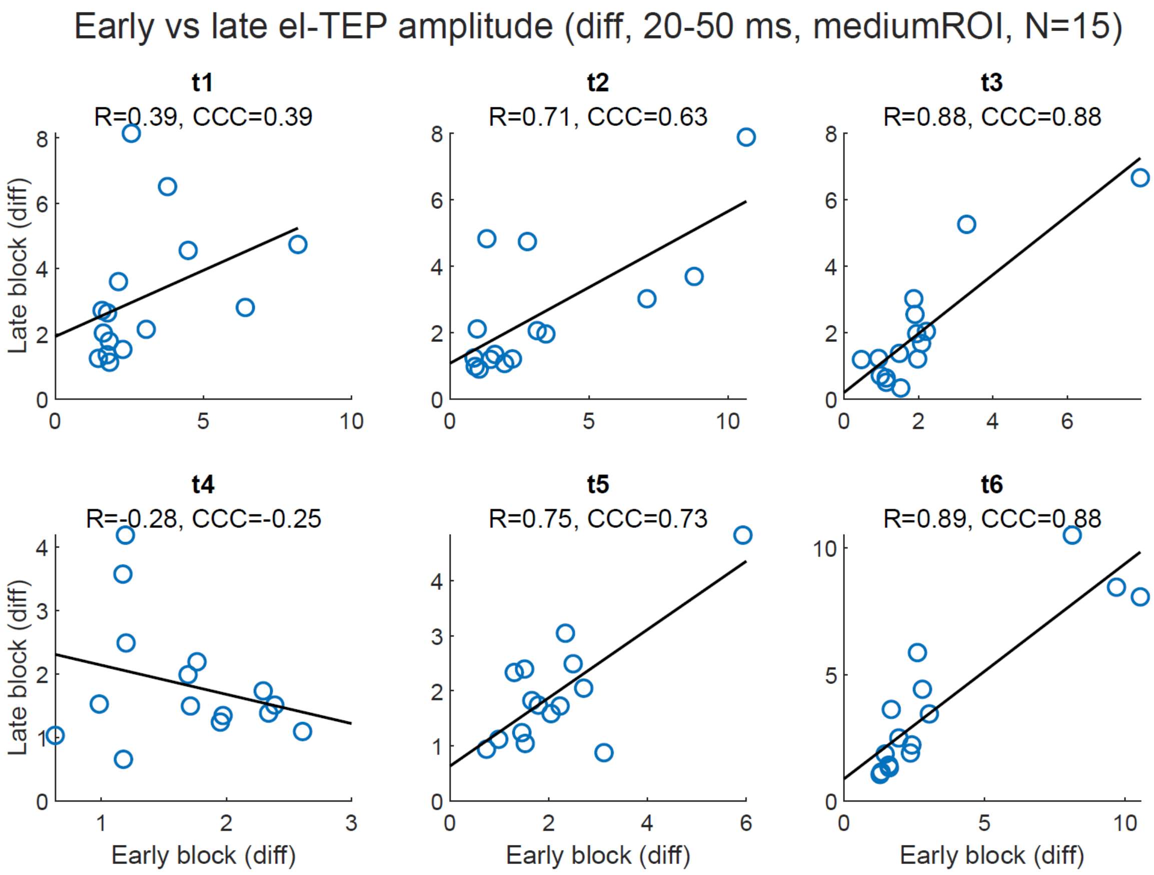


Figure S12 - EL-TEP amplitudes (medium ROI, 20-50 ms time window, peak-to-peak) of the first and the second experimental block using linear (μV) peak-to-peak values (instead of dBμV). Each dot represents one subject (N=15).

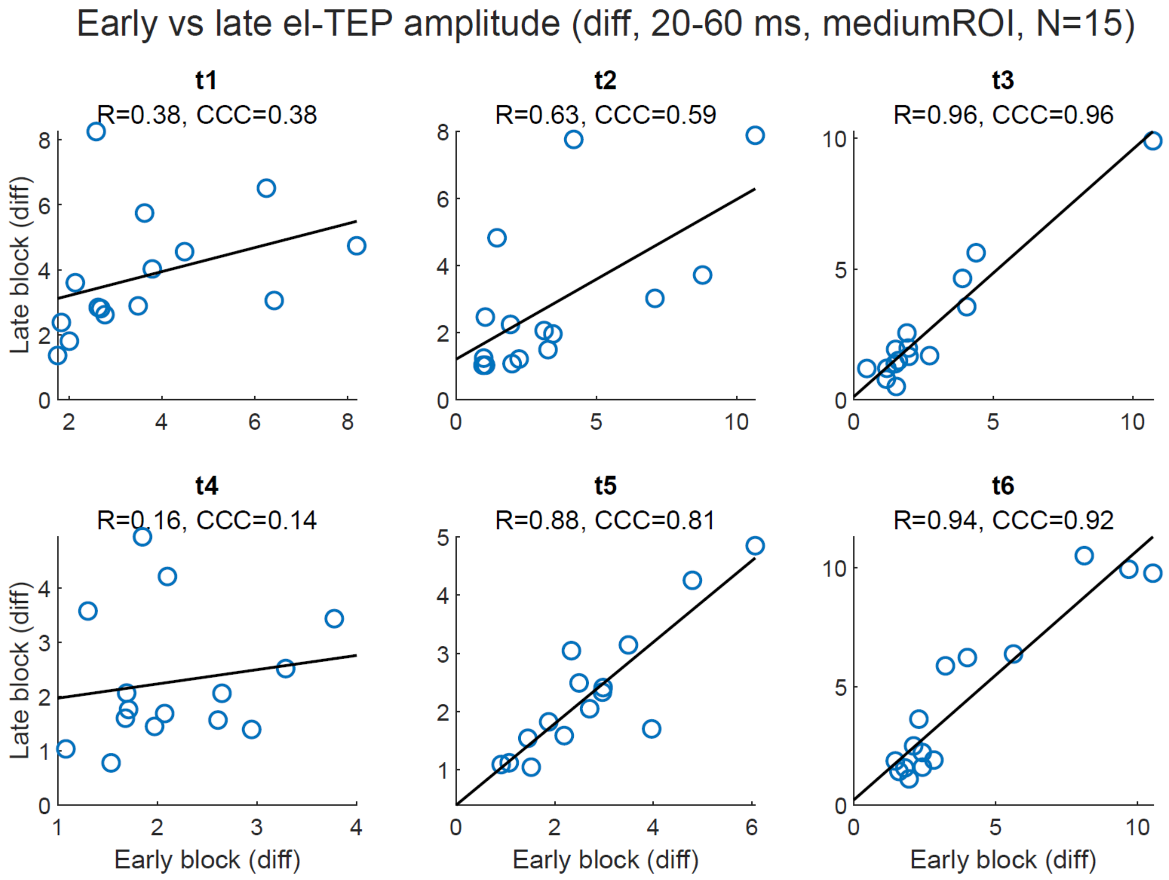


Figure S13 - EL-TEP amplitudes (medium ROI, 20-60 ms time window, peak-to-peak) of the first and the second experimental block using linear (µV) peak-to-peak values (instead of dBµV). Each dot represents one subject (N=15).

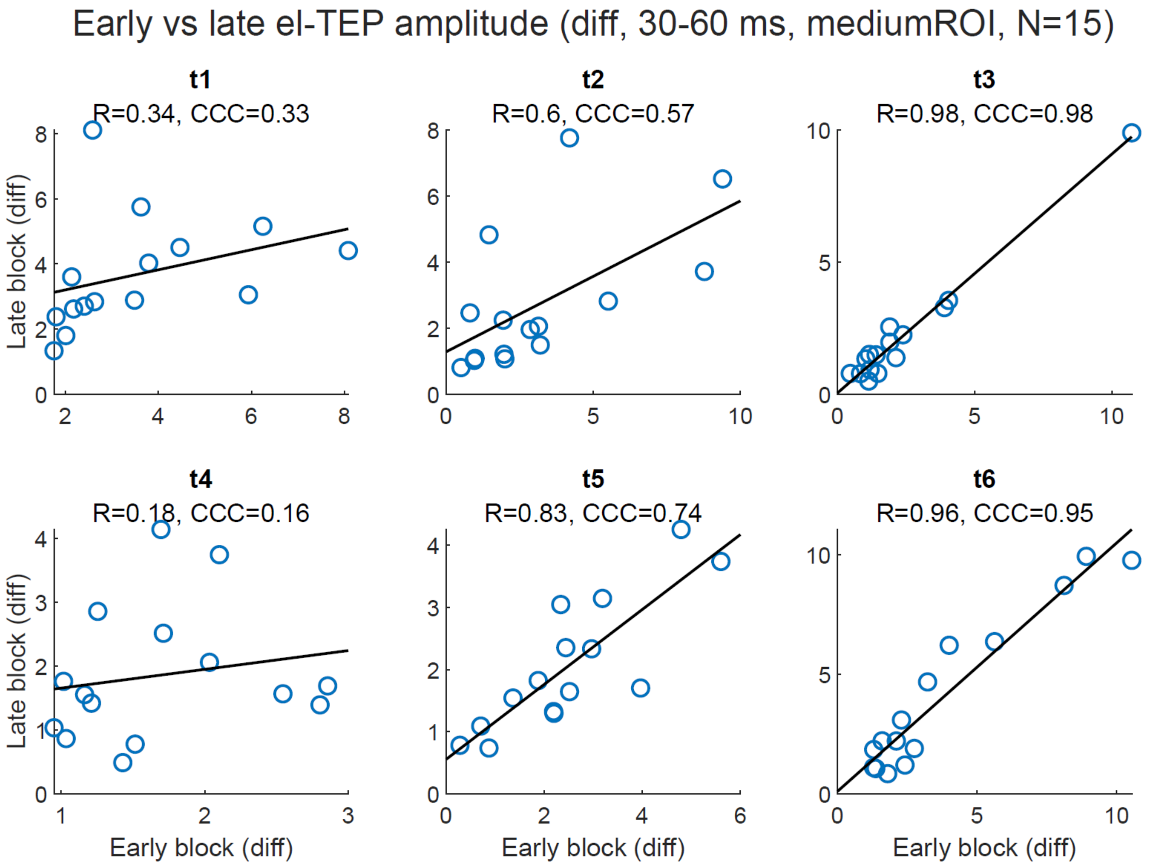


Figure S14 - EL-TEP amplitudes (medium ROI, 30-60 ms time window, peak-to-peak) of the first and the second experimental block using linear (µV) peak-to-peak values (instead of dBµV). Each dot represents one subject (N=15).

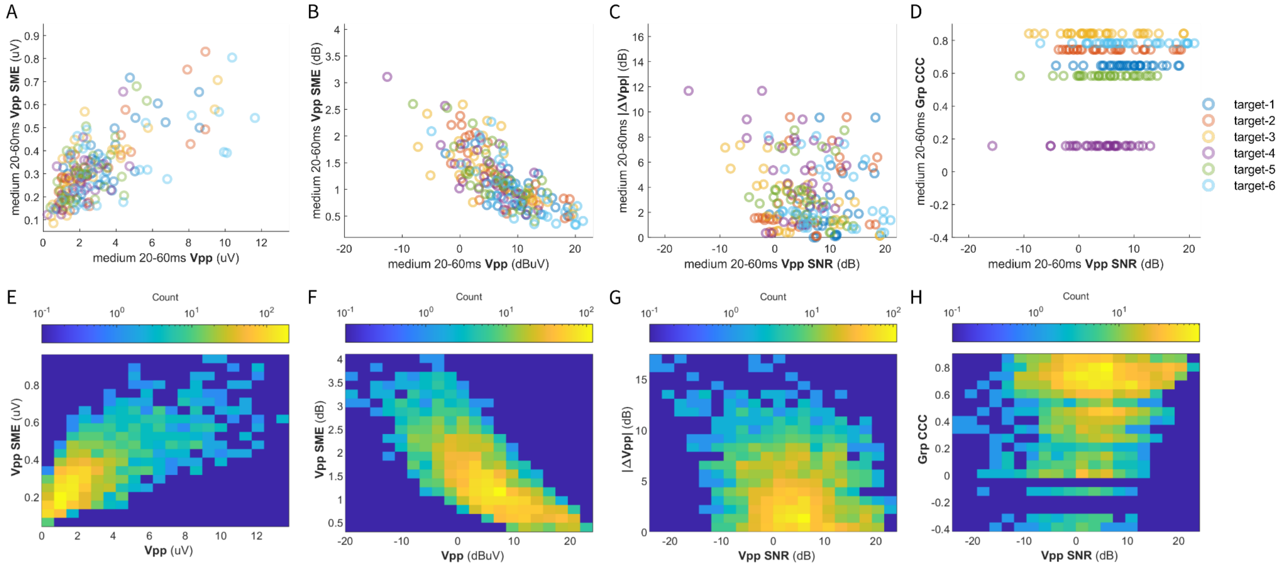


Fig S15 - Standardized measurement error (SME) analysis results. A-D) scatter plots comparing within-block and across-block amplitude, variance, and reliability metrics for the medium-size sensor-space ROI, 20-60 ms time window, and full set of 150 trials within each block. A) scatter plot of within-block mean linear response amplitude (linear scale) vs within-block linear SME. B) scatter plot of within-block mean logarithmic response amplitude vs within-block logarithmic SME. C) scatter plot of within-block logarithmic SNR vs across-block response amplitude change. D) scatter plot within-block logarithmic SNR vs. across-block group-level CCC. E-H) histograms for the same variables as (A-D) but pooled across all full-trial-count sensor-space analytical variable combinations.

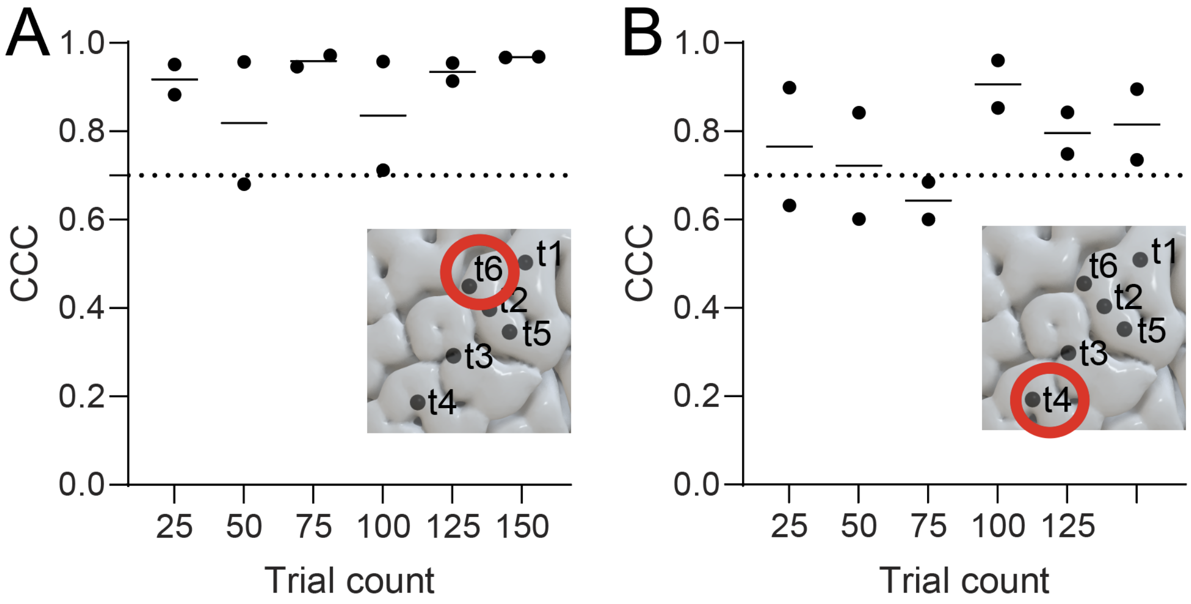


Fig S16 - Impact of automated preprocessing on EL-TEP reliability. Automated preprocessing was run twice for identical trials and CCCs were computed using the two preprocessing results. The dots in each trial count condition represent the first and the second stimulation block. Automated preprocessing introduces some variability in CCCs, particularly noticeable in target t4.

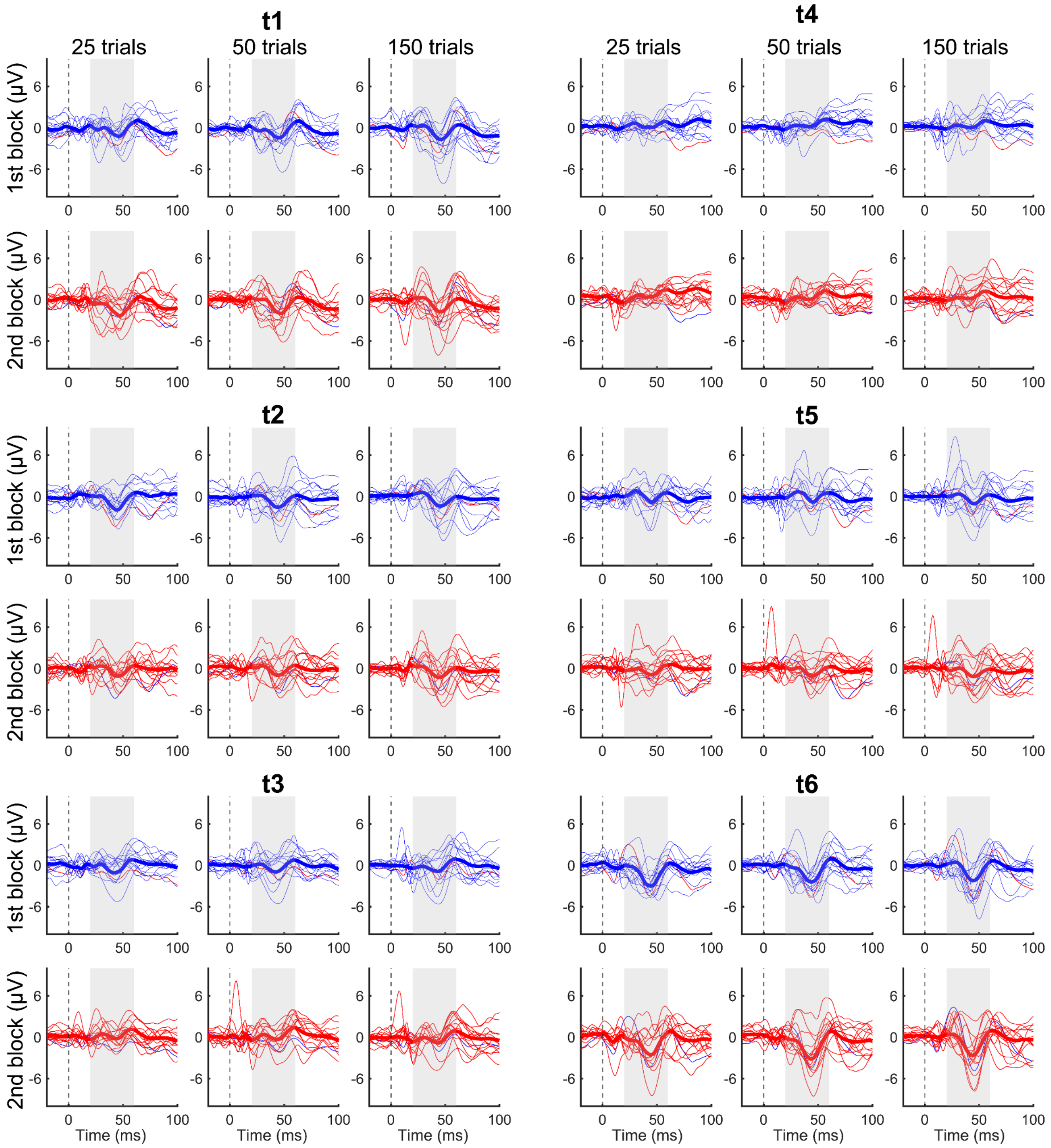


Fig S17 - TEP traces across targets for trial count analysis. TEP traces for 25, 50 and 150 trials for the first and second stimulation block are shown for each stimulation target. Bold lines represent average TEP response across subjects.