Clinical Neurophysiology 171 (2025) 133–145

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Bridging the gap to clinical use: A systematic review on TMS–EEG test-retest reliability

Giacomo Bertazzoli [a,b,*], Elisa Dognini [c], Peter J. Fried [a,b], Carlo Miniussi [d], Petro Julkunen [e,f], Marta Bortoletto [c]

[a] Berenson-Allen Center for Noninvasive Brain Stimulation, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA
[b] Department of Neurology, Harvard Medical School, Boston, MA, USA
[c] Neurophysiology Lab, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy
[d] Centre for Mind/Brain Sciences CIMeC, University of Trento, Rovereto, Italy
[e] Department of Clinical Neurophysiology, Kuopio University Hospital, Kuopio, Finland
[f] Department of Technical Physics, University of Eastern Finland, Kuopio, Finland

ARTICLE INFO

Keywords:
TMS–EEG
Reliability
TEPs
SDC
ICC

ABSTRACT

Background: Transcranial magnetic stimulation (TMS) combined with electroencephalography (EEG) can provide insight on cortical excitability and brain circuits. TMS-evoked potentials (TEPs) are phase-locked waveforms reflecting neural activity, with potential applications in psychiatry and neurology. However, the reliability of TEPs remains underexplored, hindering clinical standardization. This systematic review evaluates TEP reliability, focusing on commonly used measures and assessments.

Methods: A systematic review was conducted on PubMed for studies from 2002 to October 10, 2024, using keywords combining TMS, EEG, and reliability terms. Systematic reviews and non-English articles were excluded.

Results: Eighteen studies met inclusion criteria, mostly assessing young, healthy populations. Late TEP components demonstrated high relative reliability, while early components exhibited lower reliability and variability across sessions. Analytical methods like the intraclass and concordance correlation coefficients, and Pearson's correlations consistently favored late TEPs.

Discussion: Late TEPs exhibit higher reliability, while early components require further research. TMS artifacts complicate interpretation, in both late and early responses. Formal reliability assessments, standardized protocols, and diverse populations are essential for advancing TEP reliability for clinical application.

Conclusions: A more comprehensive reliability assessments is needed before the implementation of clinical applications.

1. Introduction

Transcranial magnetic stimulation (TMS) can be combined with neurophysiological and/or neuroimaging techniques to measure the state of the nervous system. The most common combination is with electromyography (EMG). In this case, TMS elicits motor-evoked potentials (MEPs) in contralateral muscles to investigate the excitability, conduction, and integrity of the cortico-spinal tract (Barker et al., 1985; Chen et al., 2008; Kobayashi and Pascual-Leone, 2003; Rossini et al., 2015). TMS-induced MEPs are an established measure used in clinical

Abbreviations: AD, Alzheimer's Disease; AG, Angular Gyrus; BIDS, Brain Imaging Data Structure; CCC, Concordance Correlation Coefficient; EMG, Electromyography; EPs, Evoked Potentials; ERPs, Event-Related Potentials; FAIR, Findable, Accessible, Interoperable, and Reusable; ICA, Independent Component Analysis; ICC, Intraclass Correlation Coefficient; IHB, Interhemispheric Balance; ISP, Interhemispheric Signal Propagation; LICI, Long-Interval Cortical Inhibition; M1, Primary Motor Cortex; MEPs, Motor-Evoked Potentials; mPFC, Medial Prefrontal Cortex; PEPs, Peripheral Evoked Potentials; SDC, Smallest Detectable Change; SEMeas, Standard Error of Measurement; SNR, Signal-to-Noise Ratio; SMA, Supplementary Motor Area; STE, Symbolic Transfer Entropy; TEPs, TMS-Evoked Potentials; TMS, Transcranial Magnetic Stimulation; VAR, Vector Autoregression.

* Corresponding author at: Berenson-Allen Center for Noninvasive Brain Stimulation, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA.

E-mail address: gbertazz@bidmc.harvard.edu (G. Bertazzoli).

https://doi.org/10.1016/j.clinph.2025.01.002
Accepted 3 January 2025
Available online 23 January 2025

practice to examine the functional state of corticospinal pathways in diseases involving motor dysfunction (Groppa et al., 2012).

TMS can also be combined with electroencephalography (EEG) to measure the direct cortical response to the TMS pulse (Ilmoniemi et al., 1997) and assess cortical excitability and effective connectivity, including in non-motor brain regions. The former refers to the strength at which the cortex responds to TMS (Bonato et al., 2006; Casula et al., 2022; Kähkönen et al., 2004), the latter refers to the activation of regions distant from the stimulation site and connected through structural and functional connections (Bortoletto et al., 2015, 2021; Momi et al., 2021a; Ozdemir et al., 2020; Zazio et al., 2022). Compared to other neuroimaging methods, the combination of TMS-EEG may offer a unique window into the study of the brain, as it possesses both the high temporal resolution of the EEG and the *causal* link to a specific region of cortex targeted by the TMS (Bortoletto et al., 2015).

Specifically, TMS-evoked potentials (TEPs) are complex phase-locked waveforms that reflect the cortical spreading of neural activity after TMS. TEPs capture the brain's complex response to stimulation and offer a range of analysis approaches, including examining peak amplitudes and latencies, decomposing and studying the signal in the time–frequency domain, and assessing intrinsic signal complexity. Many studies have shown the potential application of these TEPs to investigate alterations in cortical activity for psychiatric conditions such as major depressive disorder, bipolar disorder and schizophrenia (Canali et al., 2015; D'Agati et al., 2014; Kirkovski et al., 2016; Levit-Binnun et al., 2009; Naim-Feil et al., 2016; Noda et al., 2018b, Noda et al., 2018a), neurological conditions such as Alzheimer's disease (Bagattini et al., 2019; Casarotto et al., 2011; Ferreri et al., 2021, 2016; Julkunen et al., 2008; Koch et al., 2018; Kumar et al., 2017), disorders of consciousness (e.g., Arai et al., 2021; Bai et al., 2016; Bodart et al., 2017; Casarotto et al., 2016; Ferrarelli et al., 2010; Formaggio et al., 2016; Gosseries et al., 2014; Massimini et al., 2012; Ragazzoni et al., 2017, 2013; Rosanova et al., 2012, Rosanova et al., 2009; Sarasso et al., 2014; for review see Ragazzoni et al. 2017) and stroke (Bodart et al., 2017; Borich et al., 2016; Cipollari et al., 2015; Manganotti et al., 2015; Pellicciari et al., 2018).

Despite these studies prospect a clinical application of TMS-EEG, this technique is still far from achieving clinical standards. A systematic quantification of reliability of TEPs is needed to plan their exploitation in clinical contexts, e.g. in clinical trials (Julkunen et al., 2022). Here, we aim to (1) provide a working definition of reliability in the context of TMS-EEG, specifically focusing on types of test–retest reliability assessments and their specific meaning; (2) review the current literature on TEP reliability and (3) assess if TEPs can be considered reliable indices to be used for clinical applications.

Understanding the reliability of TEPs is crucial for their application in clinical contexts; however, a significant limitation in the field has been the lack of consistency and consensus regarding how reliability is defined. The following section aims to clarify these concepts to facilitate a more unified approach.

1.1. Consensus terminology and definition of reliability

When discussing the reliability of a biomarker or biological measures more generally, there are often inconsistent ideas of what reliability is. This is due to the use of different terminology to convey overlapping meaning, for example, reproducibility, validity, stability, and consistency. Moreover, some of these terms relate to different statistical concepts, leading to further misinterpretations (Mokkink et al., 2010; de Vet et al., 2006).

Adopting consensus terminology from Mokkink et al. (2010), the usefulness of a neurophysiological measure as a clinical or diagnostic tool depends on its *validity*, *responsiveness*, and *reliability*. In the context of neurophysiology, these domains cover the ability of a measure to (1) assess a neurophysiological process of interest (validity), (2) detect neurophysiological changes when they occur (responsiveness), and (3) remain stable in unchanging conditions (reliability) (see Table A2 in Mokkink et al. 2010).

Over the years, most of the debate on TMS–EEG has focused on its validity (Farzan and Bortoletto, 2022; Hernandez-Pavon et al., 2023; Parmigiani et al., 2023) and responsiveness (Julkunen et al., 2022; Napolitani et al., 2014; Tremblay et al., 2019), while TEPs reliability has been only partially established.

Given the ambiguity surrounding the definition of reliability, it is essential to clarify how reliability is operationalized, to ensure meaningful assessments.

1.2. Operationalization of reliability

Reliability is operationalized by breaking it down into two equally important parts, *relative* and *absolute* (Atkinson and Nevill, 1998; Beaulieu et al., 2017; McManus IC, 2012; Schambra et al., 2015; Weir, 2005).

Relative reliability refers to the degree to which unchanging individuals maintain their position relative to each other across repeated measures (Streiner and Norman, 2016; Terwee et al., 2007; de Vet et al., 2006; Weir, 2005). In other words, if we take the same measure is taken at two time points (e.g., T0 and T1) or by two raters in the same cohort of individuals, relative reliability refers to the degree to which the ranking of individuals is the same between the two time points or raters. Therefore, relative reliability depends both on within-subject variance and between-subject. This kind of reliability can be assessed with the intraclass correlation coefficient (ICC)

$$\textit{Relative Reliability} = ICC = \frac{\textit{between subject variance}}{\textit{between subject variance} + \textit{residual error}} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \quad (1)$$

where the between-subject variance and the residual error are estimated using mean square values. Based on the design of the experiment and the desired reliability outcome, the terms of the above equation can be estimated in different ways depending on the specific ICC model chosen. Each model accounts for different sources of variance, such as whether measurements are made by the same or different raters or whether raters are considered random or fixed effects. The choice of model impacts how the between-subject variance and residual error are calculated. However, a detailed discussion of these models is beyond the scope of this review (see Shrout and Fleiss, 1979; Weir, 2005 for details). The closer the ICC value is to 1, the higher the relative reliability. Values near 0 indicate poor to null relative reliability, while negative values are not theoretically meaningful but are possible in some cases (Giraudeau, 1996). When interpreting ICC values, many studies refer to Shrout (1998), who defined ranges of relative reliability with 0.00–0.10 as "virtually none", 0.11–0.40 as "slight", 0.41–0.60 as "fair", 0.61–0.80 as "moderate" and 0.81–1.00 as "substantial" (Shrout, 1998).

Absolute reliability, in contrast, captures the degree to which repeated measures for the same unchanged individual at different time points (e.g., T0 and T1) remain the same. Absolute reliability can thus be considered the absolute difference between measures taken at two different time points. If the absolute difference is small, the absolute reliability is high, while if the difference is high, the absolute reliability is low. Absolute reliability is often operationalized with the standard error of a measurement. To avoid confusion with the standard error of the mean, the abbreviation *SEMeas* is adopted here as in Schambra et al., 2015 (Beaulieu et al., 2017; Mokkink et al., 2010; Schambra et al., 2015; Terwee et al., 2007; de Vet et al., 2006). SEMeas typically captures the residual error (but in one of its variants, it can capture both residual and systematic errors) of a repeated measure (Hopkins, 2000). The smaller the SEMeas, the lower the measurement error across sessions and the more consistent the measure is:

$$Absolute\ reliability = SEMeas = \sqrt{residual\ error} = \sqrt{\sigma_e^2} \quad (2)$$

An important feature that differentiates the SEMeas from the ICC is that the SEMeas does not depend on the variability of the population in exam; hence, it is not influenced by between-subject variability. Rather, it captures the "typical error" of a measure, which mostly depends on the technique and is independent to the population (Hopkins, 2000). SEMeas is often used to derive a more immediate index, the *smallest detectable change* (*SDC*, also called *minimal difference*) (Beaulieu et al., 2017; Schambra et al., 2015; Terwee et al., 2007; Weir, 2005), which indicates the smallest change needed in a measure that can be considered a "true" change. For example, the smallest change in μV of a TEP component in a test–retest paradigm is considered a real effect of the manipulation rather than a random fluctuation.

In addition to the ICC, SEMeas, or SDC, other indices can be computed to assess reliability. It is worth mentioning the concordance correlation coefficient (CCC), which has often been used in the TMS–EEG literature. The CCC (Lin, 1989) evaluates the agreement between two variables, considering both the correlation and the accuracy of the data, making it a more comprehensive measure of agreement. Unlike the ICC, which focuses on variance components, the CCC takes into account both the precision (how well the data points follow a linear relationship) and accuracy (how closely they follow the line of perfect agreement) between measurements (Barnhart et al., 2007; Carrasco and Jover, 2003). While in some cases the CCC can yield results similar to the ICC, the inclusion of accuracy assessment makes it a useful complementary metric, particularly in clinical contexts where both the magnitude of differences and the direction of deviations between repeated measures matter. A high CCC indicates not only that the data points are correlated, but also that they align closely with the line of perfect agreement, providing a clearer picture of test–retest reliability. As a result, CCC has become a preferred choice in some TMS-EEG studies when precision and accuracy are equally important for assessing the reliability of derived measures across time. However, CCC should not be seen as a replacement for the combination of ICC and SEMeas. While CCC incorporates both precision and accuracy, SEMeas remains crucial for isolating measurement error independent of population variability (see Appendix 1 for an example of why the combination of relative and absolute reliability is crucial for reliability assessment).

2. Methods

In light of the abovementioned consensus terminology, here we review the current literature on TMS-EEG reliability. We included studies that used TMS and EEG as the primary research technique. Additionally, the studies had to focus on reliability of any TEP measure. We excluded

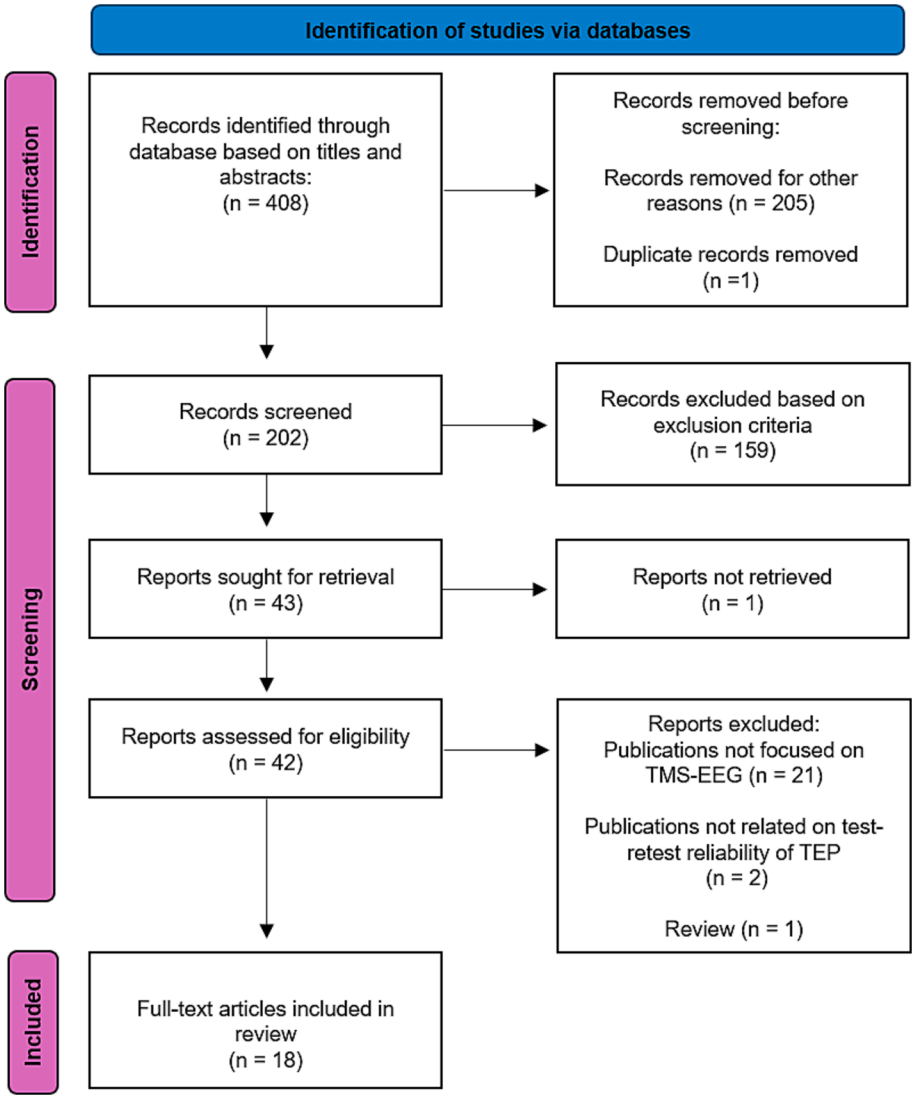

Fig. 1. PRISMA flowchart of studies selection.

systematic reviews and any articles written in a language other than English. The article search was conducted on the PubMed database and was concluded on October 10th, 2024. We applied restrictions on publication dates, considering only articles published from 2002 to the present, and sorted them by publication date from oldest to most recent. The keywords used to filter the search were ((TMS[Title/Abstract] AND EEG[Title/Abstract]) AND (reliability[Title/Abstract] OR reproducibility[Title/Abstract] OR replicability[Title/Abstract])) OR ((TEP [Title/Abstract]) AND (reliability[Title/Abstract] OR reproducibility [Title/Abstract] OR replicability[Title/Abstract])) OR ((TMS-evoked potentials[Title/Abstract]) AND (reliability[Title/Abstract] OR reproducibility[Title/Abstract] OR replicability[Title/Abstract])).

Of all the articles identified through the search (Fig. 1), two authors, GB and ED, independently reviewed the title of each, excluding articles that were not relevant to the topic of this review and any duplicates. The authors then proceeded to review the abstracts of the remaining articles, excluding studies that did not meet the inclusion criteria, that had any exclusion criteria, or whose main topic was not related to the topic of this review. Finally, the remaining articles were assessed for eligibility, and 18 articles were found to be eligible (Table 1).

3. Results

In order to assess the reliability of TEPs, two main approaches have been adopted: extracting TEP peak amplitude and latency or testing the continuous wave point-by-point.

Repeated measures of TEP peaks were first assessed by Lioumis et al., (2009), who applied TMS to the primary motor cortex (M1) and dorsolateral prefrontal cortex (DLPFC) at various intensities across two sessions one week apart. Pearson's correlation coefficients demonstrated high correlations between the amplitudes and latencies of the peaks across time points, generally exceeding 0.80. Paired t-tests revealed no significant differences in most peak amplitudes and latencies across sessions, supporting the reproducibility of TEPs. However, some differences were observed for specific peaks and intensities, particularly in DLPFC stimulation, indicating that reproducibility was high but not absolute.

Similarly, Kerwin et al., (2018) assessed the relative and absolute reliability of DLPFC TEP peaks across different time intervals and trial combinations using the CCC and SDC, respectively. They reported high relative and absolute reliability for the late peaks, N100 and P200, particularly in central and centroparietal regions. In contrast, mixed results were observed for the earlier peaks, N40 and P60.

de Goede et al., (2020) explored the relative reliability of M1 TEPs using the ICC. They found poor (0.37–0.49) to moderate (0.60–0.75) reliability for the N100 and P180 components, with variability depending on the stimulation protocol (single-pulse vs. paired-pulse TMS). The N100 and P180 showed higher reliability for single-pulse TMS, while paired-pulse TMS led to greater variability in these components.

Bertazzoli et al., (2021) tested the relative reliability of TEP peak amplitude and latency for the DLPFC and inferior parietal lobule (IPL) using the CCC. The results aligned with previous studies, showing increasing relative reliability from early to late TEP peaks in both regions. Late components, particularly the N100 and P200 peaks, showed higher reliability compared to earlier components, reflecting the typical robustness of late TEP peaks.

Guidali et al., (2023) evaluated the reliability of TEP measures, focusing on the M1-P15 component's amplitude and latency under different conditions. The study found good reliability (ICC > 0.75) for M1-P15 amplitude only in the condition using biphasic posterior-anterior stimulation when comparing blocks where participants contracted the hand ipsilateral to stimulation with blocks where the hand was relaxed (ICC $= 0.768$). However, for all other conditions—whether for amplitude or latency—poor reliability was observed, with ICC values below 0.5. Similar results were found for the CCC, which closely

Table 1
Literature review on TMS–EEG measures' reliability.

Publication	N	Target	Interval between tests	TMS-EEG measure	Reliability test
Lioumis et al. 2009	7	M1 (APB) – DLPFC	1 week	TEPs peaks amplitude and latency	T tests and Pearson's R
Casarotto et al. 2010	10	Left occipital, parietal, and frontal lobes	same day or 1 week	TEPs	DI and ROC
Farzan et al. 2010	36	M1 (APB) – DLPFC	1 week	LICI	Cronbach's alpha
Kerwin et al. 2018	16	DLPFC	2 recordings 5 min apart in the first session. Repeated 1 week after	TEPs peaks amplitude and latency	CCC – SDC
ter Braack et al. 2019	18	r-l M1	5 recordings in a day to assess effect of daytime on TEP. Same protocol repeated 1 week later for 3 subjs	Continuous TEPs and TEPs amplitude/ latency	T tests (cluster-based)
Ye et al., 2020	12	Oz	Subesquent sessions	VAR and STE	Spearman's R
Casula et al., 2021	50	r-l M1 (FDI)	3 weeks	IHB and ISP	ICC
de Goede et al. 2020	25	r-l M1 (ADM)	1 week	Continuous TEPs and TEPs amplitude/ latency	T tests/ICC (3,1)
Momi et al., 2021b	21	DMN and DAN	4 weeks	TEPs	Pearson's R
Ozdemir et al. 2021a	21	DMN and DAN	4 weeks	TEPs	Pearson's R
Mancuso et al. 2021	8	l-M1 – r-mPFC	1 week	TEPs	T tests – CCC
Ozdemir et al. 2021b	24	l-IPL, l-M1 (FDI) and l-DLPFC)	1 month	TEPs	SI
Bertazzoli et al. 2021	16	l-DLFPC – l-IPL	72.3 ± 35.8 days	TEPs	CCC – ANOVA – Spearman's R
Guidali et al. 2023	28	l-M1 (APB)	single session	M1-P15 and MEP amplitude & latency	rm-ANOVA, CCC, ICC
Song et al., 2024	24	m-PFC, l-AG, SMA	3 weeks	TEPs, PEPs	CCC
Gogulski et al., 2024	15	l-DLFPC	single session	TEPs	CCC
She et al., 2024	18	r-l M1 (APB/ ABP)	6 days	TEPs	CCC
Noda et al., 2024	20	l-DLFPC	Subesquent sessions	TEPs	CCC, SI

r = right; l = left; ADM: Abductor Digiti Minimi; AG: Angular Gyrus; ANOVA: Analysis of Variance; APB: Abductor Pollicis Brevis; CCC: Concordance Correlation Coefficient; DAN: Dorsal Attention Network; DMN: Default Mode Network; DLPFC: Dorsolateral Prefrontal Cortex; EEG: Electroencephalography; IHB: Interhemispheric Balance; ICC: Intraclass Correlation Coefficient; ISP:

Interhemispheric Signal Propagation; LICI: Long-Interval Cortical Inhibition; M1: Primary Motor Cortex; MEP: Motor Evoked Potential; m-PFC: Medial Prefrontal Cortex; P15: P15 component; PEPs: Peripheral evoked potentials; rm-ANOVA: Repeated Measures Analysis of Variance; SDC: Smallest Detectable Change; SI: Sensitivity Index; SMA: Supplementary Motor Area; STE: Symbolic Transfer Entropy; TEPs: TMS-evoked Potentials; TMS: Transcranial Magnetic Stimulation; VAR: Vector AutoRegression.

mirrored the ICC values.

Gogulski et al., (2024) assessed the reliability of early-latency TEPs in different subregions of the DLPFC. The study found that the medial DLPFC target produced the most reliable early-latency TEPs, with a CCC of 0.78, while the anterior DLPFC target had poor reliability, with a CCC of 0.24. The study also tested different analytical parameters including time window, quantification method, region of interest, sensor- vs. source-space, and number of trials and found that wider time windows (20–60 ms, CCC = 0.62) and later time windows (30–60 ms, CCC = 0.61) resulted in higher reliability than earlier and shorter windows. Notably, reliable early-latency TEPs (CCC up to 0.86) were achieved with as few as 25 to 50.

Similarly, Noda et al., (2024) investigated the optimal number of TMS pulses required to obtain reliable TEPs in the lateral DLPFC. The study compared four conditions using different numbers of TMS pulses—40, 80, 160, and 240—and evaluated their reliability using the CCC. The results indicated that both the 80-pulse and 160-pulse conditions showed high reliability, comparable to the 240-pulse condition. Specifically, CCC values for the 80- and 160-pulse conditions demonstrated strong agreement with the 240-pulse condition, while the 40-pulse condition showed only weak to moderate reliability, with significantly lower CCC values.

She et al., (2024) examined the stability of TEPs in pediatric epilepsy patients, specifically assessing the minimum number of pulses required to achieve stable TEPs in the motor cortex. The span between the test and retest sessions was two days. The study found that stable TEPs could be derived with fewer than 100 pulses, which is typically used in adult studies. The early segment of the TEP (15–80 ms) was less stable than the later segment (80–350 ms), and global mean field power (GMFP) showed less stability compared to local TEPs over the stimulated site.

Song et al., (2024) assessed the test–retest reliability of TEPs in spatial and temporal domains across repeated sessions for TMS targets (angular gyrus – AG, supplementary motor area – SMA, and medial prefrontal cortex – mPFC). High spatial CCCs (>0.8) were observed from 90 ms onward in both active and sham conditions. After removing peripheral evoked potentials (PEPs), the reliability of "cleaned" TEPs decreased. For AG, spatial CCCs remained fair to moderate (0.4–0.67) until 190 ms, with significant temporal CCCs up to 150 ms. SMA showed reliable spatial CCCs until 80 ms (0.2–0.6), while mPFC had lower CCCs, with spatial reliability until 80 ms and generally low temporal reliability.

Some studies, including Bertazzoli et al., (2021), ter Braack et al., (2019), de Goede et al., (2020), Mancuso et al., (2021), have used cluster-derived permutation t-tests to investigate differences in whole TEP responses across repeated sessions. These studies consistently reported no significant differences in TEPs across time points.

Other studies have explored the similarity of the TMS-EEG response across time points by testing correlations in both time and space dimensions. Correlation analyses were performed either spatially, by comparing responses across electrodes at each time point between sessions (spatial correlation), or temporally, by assessing the correlation of the signal at each electrode over time between sessions, sometimes divided into temporal segments (temporal correlation). These studies consistently found stronger correlations for late TEP components compared to early components, indicating the increased stability of late TEP peaks. For example, Bertazzoli et al., (2021) and Kerwin et al., (2018) reported higher correlations for late peaks (such as the N100 and P200) compared to early components. Similarly, Momi et al., 2021b and Ozdemir et al., (2021a) found that late latencies in the TMS-EEG response were more reliable across repeated sessions, further supporting the stability of late components over time.

Another strategy employed to assess the reliability of TEP responses is to synthesize the full waveform into a single index and test its reliability across multiple time points. For example, Casarotto et al., (2010) developed a nonparametric permutation-based Divergence Index (DI) to quantify the degree of divergence between two TEPs at different time points. The DI captures the percentage of significantly different samples across time and scalp locations, resulting in a single value. The receiver operating characteristic (ROC) analysis of this index allowed for the evaluation of TEP sensitivity and specificity. Casarotto et al. reported a 96.7 % accuracy, with 95.1 % sensitivity (true positive rate, akin to responsiveness in the Mokkink et al. 2010 framework) and 100 % specificity (true negative rate, akin to reliability). Their results led to two conclusions: first, that TEPs are highly sensitive to changes in stimulation site, coil angle, or intensity; and second, that TEPs are nonrandom and remain stable over time when stimulation parameters are held constant.

Ozdemir et al., (2021b) investigated the similarity of TEP responses at parietal, motor, and frontal sites by calculating the cosine similarity index between TEP matrices across different populations and time intervals. They generated a similarity matrix containing the cosine similarity values between participants' two separate visits. This matrix was used to compute several similarity metrics to assess reliability. The authors concluded that TEPs exhibit high within-subject reliability over time, but they also found substantial heterogeneity in TEPs between individuals, emphasizing the individual-specific nature of these responses.

Other studies have tested the reliability of TEP-derived measures such as TMS-EEG long-interval cortical inhibition (LICI), symbolic transfer entropy (STE), vector autoregression (VAR), and interhemispheric signal propagation (ISP) and balance (IHB). Farzan et al., (2010) demonstrated that LICI, when indexed using TMS-EEG, exhibits high test–retest reliability in both motor cortex and DLPFC, supporting its utility as a reliable measure of GABA-B-mediated cortical inhibition over time (Farzan et al., 2010). Ye et al., (2019) compared STE and VAR as methods to assess effective connectivity from TMS-EEG data. They found that STE consistently produced robust and reliable results across sessions, whereas VAR showed weaker reliability between sessions, highlighting STE as a more reliable method for detecting directional information flow in EEG signals induced by TMS (Ye et al., 2019). Casula et al., (2020) investigated the reliability of novel TMS-EEG indexes to assess interhemispheric signal propagation and balance. They reported high intra- and inter-subject reliability of these measures and their correlation with traditional measures of interhemispheric inhibition (IHI) using MEPs (Casula et al., 2020).

4. Discussion

The present review aimed to evaluate the current state TEPs' reliability. Our findings indicate that TEPs possess high relative reliability for late components (~>80 ms) and low relative reliability for early components. Absolute reliability, however, is not yet fully established for TEPs.

4.1. Reliability in early and late TEP latencies

Most studies report higher reliability for late TEP components (e.g., N100, P200) compared to earlier components (e.g., P15, N40), which suggests that later responses may offer more consistent response. Nevertheless, this estimate of reliability may be heavily dependent on TMS-related sensory activations. In fact, late TEPs are influenced by the superposition of sensory evoked-potentials induced by TMS (Farzan and Bortoletto, 2022; Rogasch et al., 2017). A strategy to address sensory artefacts is to include multiple control conditions in the experimental

design. Subtracting these control signals from the TEPs can help to minimize the contribution of unwanted signals and improve the SNR (Conde et al., 2019; Rocchi et al., 2021; Song et al., 2024). However, achieving a perfect control condition in TMS-EEG has been proven difficult due to the complex interplay of very different sources of noise. For example, for late-latency components, common artefacts are spontaneous blink, TMS-induced blinks, somatosensory evoked potentials, and auditory evoked potentials. All these artefacts are site- and intensity-dependent and contribute to masking the signal-of-interest, i.e., the direct cortical response to the TMS pulse. These artefacts complicate the *validity* of the results. However, their contribution to the *reliability* of TEPs is not straightforward as highly reliable artifacts would bias toward high reliability estimations while unstable artifact would have the opposite effect.

Early-latency TEP components tend to show lower reliability, likely due to their high frequency and focality but also for their superposition with high-amplitude TMS-induced artefacts such as TMS-muscular and decay artefacts (Farzan and Bortoletto, 2022; Hernandez-Pavon et al., 2022; Mutanen et al., 2016; Rogasch et al., 2017; Salo et al., 2020). These artefacts are removed or attenuated with sophisticated offline mathematical techniques. Most common methodologies employ independent component analysis (ICA) (Atluri et al., 2016; Rogasch et al., 2017; Wu et al., 2018). This technique divides the signal into independent components, allowing the researcher to separate those that capture artefacts from those that capture the signal of interest. However, ICA needs subjective choices to determine what part of the signal should remain and what part is an artefact to be removed. This operation is often not straightforward and may introduce experimenter variability (Hernandez-Pavon et al., 2022). In addition, many TMS-related artefacts are time-locked to the TMS pulse, similar to the genuine cortical response. This may disrupt the assumption of independence between the signal of interest and artefacts when using ICA to clean the EEG signal (Metsomaa et al., 2014). Other algorithms that require fewer subjective choices have recently been published (see the source–estimate–utilizing noise–discarding algorithm – SOUND (Mutanen et al., 2018) and signal–space projection–source–informed reconstruction – SSP-SIR (Mutanen et al., 2016)) to remove those artefacts and are a promising alternative to the more subjective ICA approach (Brancaccio et al., 2024). The variability in preprocessing strategies and in their effectiveness in removing these artefacts contribute to the difficulty of achieving high reliability and interpreting reliability measurements (Bertazzoli et al., 2021). Given the highly problematic nature of artefacts in TMS–EEG signals and the outlined limitations of preprocessing procedures, one strategy to reduce the impact of artefacts on TEP reliability may be to increase the signal-to-noise ratio (SNR) in the recording phase. A high SNR will reduce signal fluctuation due to noise, potentially increasing both relative and absolute reliability (Hämmerer et al., 2013). With this aim, Casarotto et al. (2022) suggested that visually checking the signal online and moving the coil based on a graphic-user interface allow stimulation parameters (e.g., site, coil orientation, intensity) to be identified that minimize artefacts and maximize the signal. This approach may reduce the impact of preprocessing on the final signal, but it is unclear whether it may introduce additional variability when measuring TEP components.

4.2. Relative and absolute reliability in TEP measures: Relevance for clinical applications

Most of the TMS-EEG reliability research focuses on relative reliability, such as the ICC and CCC. However, relying predominantly on relative reliability measures can introduce potential pitfalls, as these measures are highly influenced by between-subject variability within the studied population. Even if individuals' TEP values vary significantly between time points, the relative ranking of individuals may remain stable, resulting in a high ICC, which can falsely suggest high reliability. This is problematic in clinical contexts, where beyond ranking individuals, it is essential to detect within-subject changes over time — particularly for prognostic biomarkers, which track disease progression or treatment effects. Absolute reliability, assessed through metrics like the SEMeas or SDC, becomes critical for ensuring that TEP-derived measures can reliably capture subtle within-subject changes. Poor absolute reliability could result in clinically irrelevant fluctuations being interpreted as meaningful.

Relative and absolute reliability are particularly important in clinical research, depending on the type of biomarker being developed. High relative reliability is essential for diagnostic biomarkers, which distinguish between the presence and absence of pathology. Even with high validity and responsiveness, a diagnostic biomarker must consistently differentiate individuals across measurements i.e., high relative reliability. Without this, diagnoses would be unstable as the relative ranking of individuals would shift, undermining the accuracy of the diagnosis. Absolute reliability, on the other hand, is less critical for diagnostic purposes: as long as there is significant between-subject variability and high relative reliability, even large errors within the same subject do not affect the ability to distinguish individuals (Schambra et al., 2015).

For prognostic biomarkers, which assess disease progression within individuals, absolute reliability becomes the priority. These biomarkers must capture subtle changes within subjects over time to track disease progression accurately. If a measure has high ICC but poor SEMeas or SDC, it would struggle to detect small, meaningful changes due to high measurement error. In contrast, low ICC with optimal SEMeas or SDC might still be useful as a prognostic biomarker. Ideally, a biomarker with both high absolute and relative reliability could serve both diagnostic and prognostic purposes, detecting small changes while maintaining stable individual rankings.

4.3. Using t tests and correlations to assess reliability

Using ANOVAs and *t* tests to test for differences between measures taken from the same subjects in different sessions can be the first step in reliability assessment (Bertazzoli et al., 2021; ter Braack et al., 2019; Corneal et al., 2005; de Goede et al., 2020; Lioumis et al., 2009; Mancuso et al., 2021; Wolf et al., 2004). However, nonsignificant results do not imply good relative or absolute reliability because the underlying hypothesis of these analyses refers to the mean of a distribution rather than to the subject's relative ranking or residual error of a measure. As defined above, reliability (both absolute and relative) involves the ability of a measure to remain stable in unchanging individuals. A test for differences using a *t* test or an ANOVA is blind to the proportion to which a subject's measure changes with respect to other subjects after repeated measures. Moreover, high variance, resulting from substantial inter- or intra-individual variability, can raise the threshold for significance. This leads to the erroneous interpretation that a measure is stable and not changing when, in fact, the amount of change may be smaller than the variance present in the samples.

A demonstration of this case is provided in Table A2 (Panel B) and visualized in Fig. A2 (bottom right panel). In this situation, the responses change quite drastically within subjects between time points T0 and T1. In fact, the *ICC*, *SEMeas* and *SDC* suggest low reliability. However, the result of the ANOVA shows a p value of 0.08. Assuming a p threshold of 0.05 for statistical significance, we cannot conclude that the data are significantly different, nor can we assume that the measurement is reliable.

Another common strategy used to assess the reliability of TEP–derived measures is through the calculation of Pearson's or Spearman's R (Balslev et al., 2007) after repeated measurements of the same index (Bertazzoli et al., 2021; Lioumis et al., 2009; Momi et al., 2021b; Ozdemir et al., 2020; Ye et al., 2019). A common example is estimating the correlation of the latency of TEPs peaks to establish their reliability or the use of correlations to determine the topographical similarity in the same cohort in a test–retest fashion. Pearson's and Spearman's R test whether a linear relationship exists between two sets

of scores and can in some cases be used as measures of relative reliability, as they can mimick the behavior of the ICC (Portney and Watkins, 2015). However, Spearman's or Person's R mainly measures the linear relationship between test–retest scores without accounting for systematic errors or the agreement level between these scores (Bland and Altman, 2003, 1996; Lexell and Downham, 2005). It means that even a high Spearman's or Person's R value, doesn't guarantee that the scores agree closely (see Fig. A3 and Table A3 for a demonstration); it only indicates a strong linear relationship. This limitation can mislead into overestimating the agreement between measurements. ICC not only evaluates the consistency of individual performances over two or more tests but also considers changes in the group's average performance over time (Lexell and Downham, 2005). This makes it a more comprehensive measure for assessing relative reliability, especially in situations involving more than two sets of measurements or when systemic changes are expected (Atkinson and Nevill, 1998; Bland and Altman, 2003; Vaz et al., 2013). While Spearman's or Person's R is useful for assessing the strength of linear relationships, ICC provides a better understanding of test–retest reliability by accounting for both consistency and systematic changes, making it often a better choice for estimating relative reliability.

Note that correlations remain a valuable tool to infer similarity in the TMS–EEG field. Common is the use of topographic and spatial similarity assessment, to test the extent to which two topographies or time intervals are similar at a given instant in time or at given set of electrodes of the TEP wave. However, caution should be taken when interpreting high correlation as high relative reliability (Biabani et al., 2019).

4.4. Improving TMS–EEG reliability assessments

To fully comprehend the potential of TEPs for clinical implementation, an extensive assessment of the relative and absolute reliability of these measures is needed. For this, future studies should collect repeated measurements of TMS–EEG data in different target populations to test the relative (ICC) and absolute (SEMeas or SDC) reliability of the signal. Expanding our knowledge of TMS–EEG relative reliability in different populations, that is, the reliability measure that is mostly affected by intrinsic population variability, is fundamental for determining the diagnostic utility of TEP–derived measures. At the same time, more data on absolute reliability in different TMS–EEG experimental setups should be collected to establish the potential of TMS–EEG-driven measures as prognostic biomarkers. More data on TMS–EEG absolute reliability would also allow us to assess which TMS–EEG setup, recording procedure, apparatus and preprocessing method is able to yield the highest absolute reliability, helping the field move towards a standardization of TMS–EEG data collection.

ICC and SEMeas or SDC have usually been calculated when a single numeric value is extracted in each individual (e.g., peak latency), possibly due to high computational demands. However, reducing the TEPs to a series of peak amplitudes and latencies may add a further layer of variability since the process of peak extraction itself is not straightforward and the methodologies are not standardized (Luck, 2014). To address this issue, future studies could report the ICC and SEMeas or SDC computed at each time point and electrode. This strategy would allow us to follow the changes in reliability across time and electrodes, enabling a continuous evaluation of the most reliable intervals of the TMS–EEG response in both time and space.

In parallel, the establishment of normative values, akin to those used in conventional evoked potentials (EPs) and event-related potentials (ERPs) (Celesia et al., 1993; Duncan et al., 2009; Kappenman and Luck, 2011), would enable the reliable detection of neurophysiological changes, critical for both diagnostic and prognostic applications in clinical settings. Multicentric collaborations and data sharing initiatives, like the T4TE project (Bortoletto et al., 2022), are essential for improving reliability assessments, facilitating the collection of larger datasets, and promoting consistency across different labs (Pavlov et al., 2021; Weiner et al., 2017). Moreover, data sharing should follow FAIR principles (Gorgolewski et al., 2016; Pernet et al., 2019) to ensure transparency and enable the evaluation of the impact of preprocessing choices, which remains a major challenge for TMS–EEG reliability (Bertazzoli et al., 2021; Brancaccio et al., 2024; Rogasch et al., 2022). To support such efforts, the BIDS-extension proposal for noninvasive brain stimulation experiments (bids.neuroimaging.io/get_involved.html) aims at providing guidelines for harmonizing data structures across labs with the aim of facilitating data sharing. As in other neuroimaging modalities, harmonized preprocessing and analysis standards are needed to establish the clinical utility of TEP-derived measures.

5. Conclusions

In summary, TMS-EEG holds significant potential as a tool for assessing cortical excitability and connectivity, but its transition into clinical practice depends on addressing key methodological and reliability challenges. While existing studies have demonstrated high relative reliability for late latency TEP components, the overlapping sensory activation hiders the validity of the TMS response at those latencies. Early latencies of the TEP response show low relative reliability, probably due to their lower SNR and the overlap with big TMS-induced artifacts i.e., TMS-induced muscle and decay. Absolute reliability, which is crucial for detecting subtle within-subject changes, has been largely neglected in TMS-EEG.

To fully realize the potential of TMS-EEG as a clinical tool to develop diagnostic and prognostic biomarkers, future research must prioritize comprehensive reliability assessments that incorporate both relative and absolute reliability. Additionally, the development of standardized protocols and the inclusion of diverse demographic groups—beyond young, healthy Caucasian individuals—are essential.

By addressing these gaps through multicentric studies and data-sharing initiatives, TMS-EEG can progress from an experimental tool to a robust clinical biomarker capable of diagnosing and tracking neurological and psychiatric disorders with greater precision.

CRediT authorship contribution statement

Giacomo Bertazzoli: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Elisa Dognini:** Investigation, Writing – review & editing, Visualization. **Peter J. Fried:** Writing – review & editing, Supervision. **Carlo Miniussi:** Writing – review & editing, Supervision. **Petro Julkunen:** Writing – review & editing. **Marta Bortoletto:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Petro Julkunen shares an unrelated patent with and has received consulting fees from Nexstim Plc, Helsinki, Finland, manufacturer of navigated TMS systems. The other authors declare no competing interests].

Acknowledgements

Marta Bortoletto acknowledges funding from the Italian Ministry of Health ("Bando della ricerca finalizzata 2016 - Giovani Ricercatori" grant no: GR-2016-02364132 and Ricerca Corrente). Petro Julkunen acknowledges funding from the Academy of Finland (grant no: 322423).

Appendix A

Assessing relative or absolute reliability: an example of the relation between the ICC and SEMeas or SDC

Suppose we want to assess the reliability of one TEP component's latency in the same group of individuals across two time points (T0 and T1). Four sets of fictional data are provided in Table A2. For each set, we depict the latency of a TEP component in eight subjects and their analysis of variance (used to compute the ICC, SEMeas and SDC; example inspired by Weir, 2005).

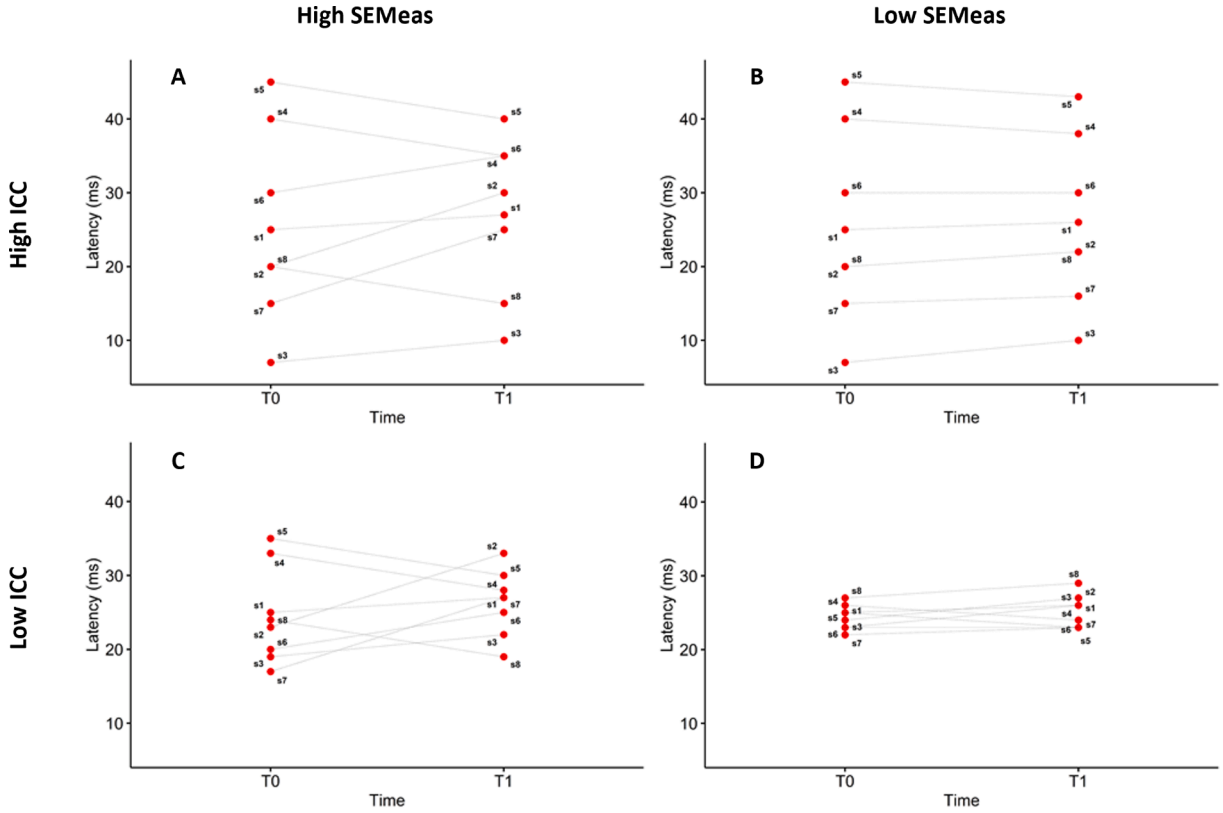

Fig. A2. Graphical representation of the test–retest fictional data presented in . Each red dot represents a fictional subject in a test (T0) and retest session (T1).The Y-axis represents the latency (ms) of a TEP component. The X-axis represents the test and retest sessions. ICC: intraclass correlation coefficient; SEMeas: standard error of the measurement.

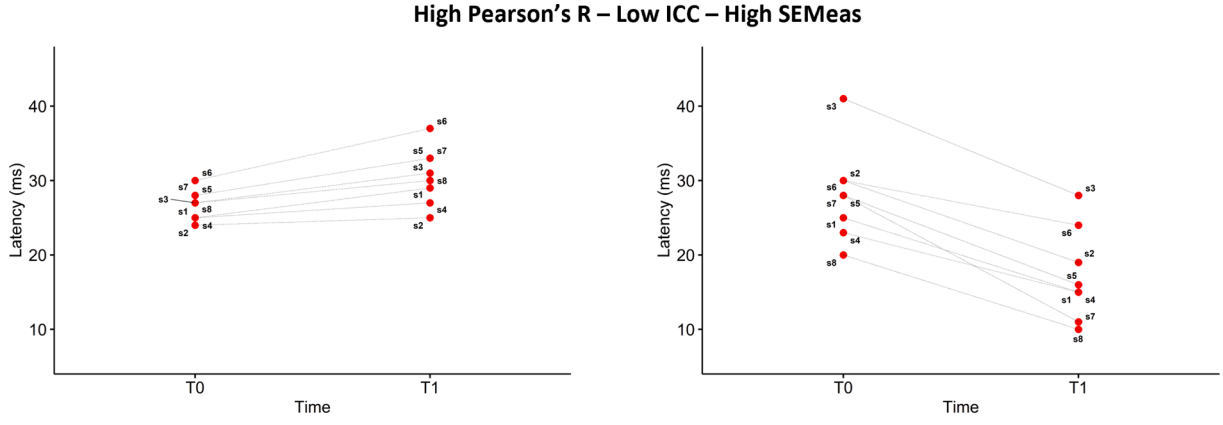

Fig. A3. Graphical representation of the test–retest fictional data presented in . Each red dot represents a fictional subject in a test (T0) and retest session (T1). The Y-axis represents the latency (ms) of a TEP component. The X-axis represents the test and retest sessions. From , the left panel corresponds to the left graph, and the right panel corresponds to the right graph. ICC: intraclass correlation coefficient; SEMeas: standard error of the measurement.

In Panel A, the ICC between the two sessions is 'substantial' (0.85) according to Shrout (1998), which suggests that the latency of this component has high relative reliability. However, the *SEMeas* and *SDC* are also high and thus suboptimal (4.42 ms and 12.25 ms, respectively). This SDC means that the change in the latency of that component could begin to be considered a "true" change (i.e., caused by an experimental manipulation) only when it exceeds 12.25 ms. In this example, this would be an average change in latency of more than 50 %, which reflects low absolute reliability. Why do the relative and absolute reliability estimations conflict? One explanation for the discrepancy between relative and absolute reliability is visualized in Fig. A2 panel A. The ICC is high because the relative position, i.e., the ranking, of the subjects between T0 and T1 remains the same for most subjects. Since the ICC tracks the ability of a measure to stably rank individuals across repeated measurements, the ICC remains high even if the change in absolute value between measures is high. Note also that the ICC is the proportion of between-subject variance to total variance, which, as reported in

the ANOVA (Panel A in Table A2), favours the between-subject variance. High between-subject variance could be due to many reasons, such as the measured phenomenon varying substantially between subjects in a specific population (e.g., when measuring a disease biomarker in a patient cohort). Conversely, the SEMeas and the SDC are sensitive only to the change in absolute value within subjects and between time points rather than the relative ranking of individuals between the two time points. Consequently, although the TEP component latency shows high relative reliability (high *ICC)*, the high *SEMeas* and *SDC* indicate that this measure is not stable within subjects across time and is thus suboptimal if the goal is to detect small but meaningful changes within subjects when they occur.

Table A2
Fictional data for demonstrating the relation between ICC and SEMeas/SDC.

A

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	27	2	**Between**	1723.93	7	246.27	12.10	0.002	3.78
2	20	30	10	**Within**	156.5	8	19.56			
3	7	10	3	**Time**	14.06	1	14.06	0.69	0.43	5.59
4	40	35	−5	**Error**	142.43	7	20.34			
5	45	40	−5	**Total**	1880.43	15				
6	30	35	5	**ICC**	0.85					
7	15	25	10	**SEMeas**	4.42					
8	20	15	−5	**SDC**	12.26					
Mean	25.25	27.12	1.87							
Std	11.85	9.63	6.38							

B

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	26	1	**Between**	1948.43	7	278.34	163.21	<0.001	3.78
2	20	22	2	**Within**	13.5	8	1.68			
3	7	10	3	**Time**	1.56	1	1.56	0.91	0.37	5.59
4	40	38	−2	**Error**	11.95	7	1.70			
5	45	43	−2	**Total**	1961.93	15				
6	30	30	0	**ICC**	0.98					
7	15	16	1	**SEMeas**	1.30					
8	20	22	2	**SDC**	3.60					
Mean	25.25	25.87	0.63							
Std	11.85	10.22	1.85							

C

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	27	2	**Between**	41	7	5.85	2.98	0.09	3.78
2	23	33	10	**Within**	16	8	2			
3	19	22	3	**Time**	2.25	1	2.25	1.14	0.32	5.59
4	33	28	−5	**Error**	13.75	7	1.96			
5	35	30	−5	**Total**	57	15				
6	20	25	5	**ICC**	0.49					
7	17	27	10	**SEMeas**	1.41					
8	24	19	−5	**SDC**	3.92					
Mean	24.5	26.37	1.88							
Std	6.04	4.12	6.38							

D

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	26	1	**Between**	41	7	5.85	2.98	0.086	41
2	24	27	3	**Within**	16	8	2			16
3	23	26	3	**Time**	2.25	1	2.25	1.14	0.32	2.25
4	26	24	−2	**Error**	13.75	7	1.96			13.75
5	25	23	−2	**Total**	57	15				
6	23	23	0	**ICC**	0.49					
7	22	23	1	**SEMeas**	1.41					
8	27	29	2	**SDC**	3.92					
Mean	24.37	25.12	0.75							
Std	1.57	2.08	1.98							

SS = sum of squares; df = degrees of freedom; MS = Mean sum of squares; F = F statistic; Sig = significance probability; F crit = threshold F value for significance; T0 = test session; T1 = retest session; Diff = difference in score; Std = standard deviation; ICC = intraclass correlation coefficient; SEMeas = standard error of the measurement; SDC = smallest detectable change;

Table A3
Fictional data for demonstrating the relation between ICC and SEMeas/SDC.

A – High Pearson R, Low ICC

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	29	4	**Between**	114.93	7	16.41	9.24	0.004	3.78
2	24	25	1	**Within**	72.5	8	9.06			
3	27	31	4	**Time**	60.06	1	60.06	33.80	0.001	5.59
4	25	27	2	**Error**	12.43	7	1.77			
5	28	33	5	**Total**	187.43	15				
6	30	37	7	**ICC**	0.28					
7	28	33	5	**SEMeas**	3.01					
8	27	30	3	**Pearson's R**	0.97					
Mean	26.75	30.62	3.88	**SDC**	8.34					
Std	1.85	3.53	1.88							

B – High Pearson R, Low ICC

Example data				ANOVA						
Subject	**T0**	**T1**	**Diff**	**Source of variance**	**SS**	**df**	**MS**	**F**	**Sig**	**F crit**
1	25	15	−10	**Between**	503.93	7	71.99	13.11	0.002	3.78
2	30	19	−11	**Within**	511.5	8	63.93			
3	41	28	−13	**Time**	473.06	1	473.06	86.15	<0.001	5.59
4	23	15	−8	**Error**	38.43	7	5.49			
5	28	16	−12	**Total**	1961.93	15				
6	30	24	−6	**ICC**	0.05					
7	28	11	−17	**SEMeas**	7.99					
8	20	10	−10	**Pearson's R**	0.85					
Mean	28.12	17.25	−10.88	**SDC**	22.16					
Std	5.86	5.78	3.31							

SS = sum of squares; df = degrees of freedom; MS = Mean sum of squares; F = F statistic; Sig = significance probability; F crit = threshold F value for significance; T0 = test session; T1 = retest session; Diff = difference in score; Std = standard deviation; ICC = intraclass correlation coefficient; SEMeas = standard error of the measurement; SDC = smallest detectable change;

Table A2 (Panel C) and Fig. A2 panel C describe an example similar to the one described above. The only difference is represented by the between-subject variance that, in this case, is reduced. Note that the differences in scores within subjects are identical to those in Panel A (Table A2). However, the shrinking of the between-subject variance caused the *ICC* to decrease to 0.35, which would be described as 'virtually none' by Shrout (1998), while the *SEMeas* and *SDC* remain the same. This highlights the population dependency of the ICC: when using the same measure (i.e., the TEP component latency) in another population with different between-subject variance, the ICC varies. However, the *SEMeas* and *SDC* remain the same because of their dependency on the technique instead of the population.

As another example, Panel B of Table A2 depicts a scenario in which the between-subject variability is similar to that found in Panel A, but the difference in measure scores between the two sessions is reduced. In this case, the high between-subject variance yields a high ICC of 0.98, and the low residual error yields a low (optimal) SEMeas and SDC of 1.30 ms and 3.60 ms, respectively. This is visualized in Fig. A2 panel B, which shows how the relative ranking of the subjects does not change (high *ICC*), while the change between time points within subjects remains low (low *SEMeas* and *SDC*).

Finally, we may face a scenario such as the one depicted in Panel D of Table A2, which is similar to Panel B but with reduced between-subject variance. In this case, the lower between-subject variance prevents the relative ranking of the subjects from remaining stable across repeated measures, yielding a low ICC of 0.49. However, the low residual error between measures within subjects maintains a low (optimal) SEMeas and SDC of 1.41 ms and 3.92 ms, respectively (Fig. A2 panel D).

The *ICC* can only be interpreted as a reflection of the examined population because there are cases in which the *ICC* is low (suboptimal) even when the absolute differences between repeated measures are low (Table A2 Panel D) or cases in which the ICC is high even when the differences between repeated measures are high (Table A2 Panel A). On the other hand, the SEMeas and SDC show how reliable a measure remains across populations as long as the same technique is employed (Beaulieu et al. 2017; Schambra et al. 2015; Weir 2005). Therefore, complementing the ICC with the SEMeas or the SDC (i.e., complementing relative reliability with absolute reliability estimates) provides a more complete picture of a measure's reliability. This is useful not only for determining the clinical potential of measures (in this case, TEP-derived measures) but also for avoiding erroneous conclusions based on one of the two reliability estimates.

A feature that distinguishes measures of relative reliability (e.g., the ICC) from measures of absolute reliability (here, the SEMeas and SDC) is also their ease of interpretation. A high or low ICC is fairly easy to define, with 1 reflecting perfect relative reliability and 0 reflecting no relative reliability. In addition, predefined interpretations and scales, such as those from Shrout (1998), can be used to standardize ICC interpretations and interpret intermediate values. In contrast, the interpretation of absolute reliability indices such as SEMeas and SDC is less straightforward. For instance, whether an SEMeas and SDC value should be considered high (suboptimal) or low (optimal) depends entirely on the technique employed and the phenomenon studied.

References

Arai, N., Nakanishi, T., Nakajima, S., Li, X., Wada, M., Daskalakis, Z.J., et al., 2021. Insights of neurophysiology on unconscious state using combined transcranial magnetic stimulation and electroencephalography: a systematic review. Neurosci. Biobehav. Rev. 131, 293–312. https://doi.org/10.1016/J.NEUBIOREV.2021.09.029.

Atkinson, G., Nevill, A.M., 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 26, 217–238. https://doi.org/10.2165/00007256-199826040-00002.

Atluri, S., Frehlich, M., Mei, Y., Garcia Dominguez, L., Rogasch, N.C., Wong, W., et al., 2016. TMSEEG: A MATLAB-based graphical user interface for processing electrophysiological signals during transcranial magnetic stimulation. Front. Neural Circuits 10, 78. https://doi.org/10.3389/fncir.2016.00078.

Bagattini, C., Mutanen, T.P., Fracassi, C., Manenti, R., Cotelli, M., Ilmoniemi, R.J., et al., 2019. Predicting Alzheimer's disease severity by means of TMS–EEG coregistration.

Neurobiol. Aging 80, 38–45. https://doi.org/10.1016/j.neurobiolaging.2019.04.008.
Bai, Y., Xia, X., Kang, J., Yin, X., Yang, Y., He, J., et al., 2016. Evaluating the effect of repetitive transcranial magnetic stimulation on disorders of consciousness by using TMS-EEG. Front. Neurosci. 10. https://doi.org/10.3389/FNINS.2016.00473.
Balslev, D., Braet, W., McAllister, C., Miall, R.C., 2007. Inter-individual variability in optimal current direction for transcranial magnetic stimulation of the motor cortex. J. Neurosci. Methods 162, 309–313. https://doi.org/10.1016/j.jneumeth.2007.01.021.
Barker, A.T., Jalinous, R., Freeston, I.L., 1985. Non-invasive magnetic stimulation of human cortex. Lancet 325, 1106–1107.
Barnhart, H.X., Lokhnygina, Y., Kosinski, A.S., Haber, M., 2007. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. J. Biopharm. Stat. 17, 721–738. https://doi.org/10.1080/10543400701329497.
Beaulieu, L.D., Flamand, V.H., Massé-Alarie, H., Schneider, C., 2017. Reliability and minimal detectable change of transcranial magnetic stimulation outcomes in healthy adults: a systematic review. Brain Stimulat. 10, 196–213. https://doi.org/10.1016/j.brs.2016.12.008.
Bertazzoli, G., Esposito, R., Mutanen, T.P., Ferrari, C., Ilmoniemi, R.J., Miniussi, C., et al., 2021. The impact of artifact removal approaches on TMS-EEG signal. Neuroimage 239, 118272. https://doi.org/10.1016/j.neuroimage.2021.118272.
Biabani, M., Fornito, A., Mutanen, T.P., Morrow, J., Rogasch, N.C., 2019. Characterizing and minimizing the contribution of sensory inputs to TMS-evoked potentials. Brain Stimulat 12, 1537–1552. https://doi.org/10.1016/j.brs.2019.07.009.
Bland, J.M., Altman, D.G., 1996. Statistics Notes: Measurement error and correlation coefficients. BMJ 313, 41–42. https://doi.org/10.1136/bmj.313.7048.41.
Bland, J.M., Altman, D.G., 2003. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet. Gynecol. 22, 85–93. https://doi.org/10.1002/uog.122.
Bodart, O., Gosseries, O., Wannez, S., Thibaut, A., Annen, J., Boly, M., et al., 2017. Measures of metabolism and complexity in the brain of patients with disorders of consciousness. NeuroImage Clin. 14, 354–362. https://doi.org/10.1016/J.NICL.2017.02.002.
Bonato, C., Miniussi, C., Rossini, P.M., 2006. Transcranial magnetic stimulation and cortical evoked potentials: A TMS/EEG co-registration study. Clin. Neurophysiol. 117, 1699–1707. https://doi.org/10.1016/j.clinph.2006.05.006.
Borich, M.R., Wheaton, L.A., Brodie, S.M., Lakhani, B., Boyd, L.A., 2016. Evaluating interhemispheric cortical responses to transcranial magnetic stimulation in chronic stroke: A TMS-EEG investigation. Neurosci. Lett. 618, 25–30. https://doi.org/10.1016/J.NEULET.2016.02.047.
Bortoletto M, Bonzano L, Zazio A, Ferrari C, Pedull L, Gasparotti R, et al. Asymmetric transcallosal conduction delay leads to finer bimanual coordination 2021. Doi: 10.1016/j.brs.2021.02.002.
Bortoletto, M., Veniero, D., Thut, G., Miniussi, C., 2015. The contribution of TMS-EEG coregistration in the exploration of the human cortical connectome. Neurosci. Biobehav. Rev. 49, 114–124. https://doi.org/10.1016/j.neubiorev.2014.12.014.
Bortoletto, M., Veniero, D., Julkunen, P., Hernandez-Pavon, J.C., Mutanen, T.P., Zazio, A., et al., 2022. T4TE: Team for TMS−EEG and improve reproducibility through an open collaborative initiative. Brain Stimul. Basic Transl. Clin. Res. Neuromodulation. https://doi.org/10.1016/J.BRS.2022.12.004.
Brancaccio, A., Tabarelli, D., Zazio, A., Bertazzoli, G., Metsomaa, J., Ziemann, U., et al., 2024. Towards the definition of a standard in TMS-EEG data preprocessing. Neuroimage 301, 120874. https://doi.org/10.1016/j.neuroimage.2024.120874.
Canali, P., Sarasso, S., Rosanova, M., Casarotto, S., Sferrazza-Papa, G., Gosseries, O., et al., 2015. Shared reduction of oscillatory natural frequencies in bipolar disorder, major depressive disorder and schizophrenia. J. Affect. Disord. 184, 111–115. https://doi.org/10.1016/j.jad.2015.05.043.
Carrasco, J.L., Jover, L., 2003. Estimating the generalized concordance correlation coefficient through variance components. Biometrics 59, 849–858. https://doi.org/10.1111/J.0006-341X.2003.00099.X.
Casarotto, S., Lauro, L.J.R., Bellina, V., Casali, A.G., 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. https://doi.org/10.1371/journal.pone.0010281.
Casarotto, S., Määttä, S., Herukka, S.K., Pigorini, A., Napolitani, M., Gosseries, O., et al., 2011. Transcranial magnetic stimulation-evoked EEG/cortical potentials in physiological and pathological aging. Neuroreport 22, 592–597. https://doi.org/10.1097/WNR.0B013E328349433A.
Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., et al., 2016. Stratification of unresponsive patients by an independently validated index of brain complexity. Ann. Neurol. 80, 718–729. https://doi.org/10.1002/ANA.24779.
Casarotto, S., Fecchio, M., Rosanova, M., Varone, G., D'Ambrosio, S., Sarasso, S., et al., 2022. The rt-TEP tool: real-time visualization of TMS-Evoked Potentials to maximize cortical activation and minimize artifacts. J. Neurosci. Methods 370. https://doi.org/10.1016/J.JNEUMETH.2022.109486.
Casula, E.P., Maiella, M., Pellicciari, M.C., Porrazzini, F., D'Acunto, A., Rocchi, L., et al., 2020. Novel TMS-EEG indexes to investigate interhemispheric dynamics in humans. Clin. Neurophysiol. 131, 70–77. https://doi.org/10.1016/j.clinph.2019.09.013.
Casula, E.P., Pellicciari, M.C., Bonnì, S., Spanò, B., Ponzo, V., Salsano, I., Giulietti, G., Martino Cinnera, A., Maiella, M., Borghi, I., Rocchi, L., Bozzali, M., Sallustio, F., Caltagirone, C., Koch, G., 2021. Evidence for interhemispheric imbalance in stroke patients as revealed by combining transcranial magnetic stimulation and electroencephalography. Hum Brain Mapp. 42 (5), 1343–1358. https://doi.org/10.1002/hbm.25297.
Casula, E.P., Tieri, G., Rocchi, L., Pezzetta, R., Maiella, M., Pavone, E.F., et al., 2022. Feeling of ownership over an embodied avatar's hand brings about fast changes of fronto-parietal cortical dynamics. J. Neurosci. 42, 692–701. https://doi.org/10.1523/JNEUROSCI.0636-21.2021.
Celesia, G.G., Bodis-Wollner, I., Chatrian, G.E., Harding, G.F.A., Sokol, S., Spekreijse, H., 1993. Recommended standards for electroretinograms and visual evoked potentials. Report of an IFCN committee. Electroencephalogr. Clin. Neurophysiol. 87, 421–436. https://doi.org/10.1016/0013-4694(93)90157-Q.
Chen, R., Cros, D., Curra, A., Di Lazzaro, V., Lefaucheur, J.P., Magistris, M.R., et al., 2008. The clinical diagnostic utility of transcranial magnetic stimulation: Report of an IFCN committee. Clin. Neurophysiol. 119, 504–532. https://doi.org/10.1016/j.clinph.2007.10.014.
Cipollari, S., Veniero, D., Razzano, C., Caltagirone, C., Koch, G., Marangolo, P., 2015. Combining TMS-EEG with transcranial direct current stimulation language treatment in aphasia. Expert Rev. Neurother. 15, 833–845. https://doi.org/10.1586/14737175.2015.1049998.
Conde, V., Tomasevic, L., Akopian, I., Stanek, K., Saturnino, G.B., 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. https://doi.org/10.1016/j.neuroimage.2018.10.052.
Corneal, S.F., Butler, A.J., Wolf, S.L., 2005. Intra- and intersubject reliability of abductor pollicis brevis muscle motor map characteristics with transcranial magnetic stimulation. Arch. Phys. Med. Rehabil. 86, 1670–1675. https://doi.org/10.1016/j.apmr.2004.12.039.
D'Agati, E., Hoegl, T., Dippel, G., Curatolo, P., Bender, S., Kratz, O., et al., 2014. Motor cortical inhibition in ADHD: modulation of the transcranial magnetic stimulation-evoked N100 in a response control task. J. Neural Transm. Vienna Austria 1996 (121), 315–325. https://doi.org/10.1007/S00702-013-1097-7.
de Goede, A.A., Cumplido-Mayoral, I., van Putten, M.J.A.M., 2020. Spatiotemporal dynamics of single and paired pulse TMS-EEG responses. Brain Topogr. 33, 425–437. https://doi.org/10.1007/S10548-020-00773-6.
de Vet, H.C.W., Terwee, C.B., Knol, D.L., Bouter, L.M., 2006. When to use agreement versus reliability measures. J. Clin. Epidemiol. 59, 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015.
Duncan, C.C., Barry, R.J., Connolly, J.F., Fischer, C., Michie, P.T., Näätänen, R., et al., 2009. Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clin. Neurophysiol. 120, 1883–1908. https://doi.org/10.1016/J.CLINPH.2009.07.045.
Farzan, F., Barr, M.S., Levinson, A.J., Chen, R., Wong, W., Fitzgerald, P.B., et al., 2010. Evidence for gamma inhibition deficits in the dorsolateral prefrontal cortex of patients with schizophrenia. Brain 133, 1505–1514.
Farzan, F., Bortoletto, M., 2022. Identification and verification of a "true" TMS evoked potential in TMS-EEG. J. Neurosci. Methods 378, 109651. https://doi.org/10.1016/J.JNEUMETH.2022.109651.
Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B.A., Angelini, G., et al., 2010. Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. PNAS 107, 2681–2686. https://doi.org/10.1073/pnas.0913008107.
Ferreri, F., Vecchio, F., Vollero, L., Guerra, A., Petrichella, S., Ponzo, D., et al., 2016. Sensorimotor cortex excitability and connectivity in Alzheimer's disease: A TMS-EEG Co-registration study. Hum. Brain Mapp. 37, 2083–2096. https://doi.org/10.1002/hbm.23158.
Ferreri, F., Guerra, A., Vollero, L., Ponzo, D., Määttä, S., Könönen, M., et al., 2021. TMS-EEG biomarkers of amnestic mild cognitive impairment due to Alzheimer's disease: a proof-of-concept six years prospective study. Front. Aging Neurosci. 13. https://doi.org/10.3389/FNAGI.2021.737281.
Formaggio, E., Cavinato, M., Storti, S.F., Tonin, P., Piccione, F., Manganotti, P., 2016. Assessment of event-related EEG power after single-pulse TMS in unresponsive wakefulness syndrome and minimally conscious state patients. Brain Topogr. 29, 322–333. https://doi.org/10.1007/S10548-015-0461-3.
Giraudeau, B., 1996. Negative values of the intraclass correlation coefficient are not theoretically possible. J. Clin. Epidemiol. 49, 1205–1206. https://doi.org/10.1016/0895-4356(96)00053-4.
Gogulski, J., Cline, C.C., Ross, J.M., Parmigiani, S., Keller, C.J., 2024. Reliability of the TMS-evoked potential in dorsolateral prefrontal cortex. Cereb. Cortex 34, bhae130. https://doi.org/10.1093/cercor/bhae130.
Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., et al., 2016.1–9.. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 2016 (31), 3. https://doi.org/10.1038/sdata.2016.44.
Gosseries, O., Thibaut, A., Boly, M., Rosanova, M., Massimini, M., Laureys, S., 2014. Assessing consciousness in coma and related states using transcranial magnetic stimulation combined with electroencephalography. Ann. Fr. Anesth. Reanim. 33, 65–71. https://doi.org/10.1016/J.ANNFAR.2013.11.002.
Groppa, S., Oliviero, A., Eisen, A., Quartarone, A., Cohen, L.G., Mall, V., et al., 2012. A practical guide to diagnostic transcranial magnetic stimulation: report of an IFCN committee. Clin. Neurophysiol. 123, 858–882. https://doi.org/10.1016/J.CLINPH.2012.01.010.
Guidali, G., Zazio, A., Lucarelli, D., Marcantoni, E., Stango, A., Barchiesi, G., et al., 2023. Effects of transcranial magnetic stimulation (TMS) current direction and pulse waveform on cortico-cortical connectivity: A registered report TMS-EEG study. Eur. J. Neurosci. 58, 3785–3809. https://doi.org/10.1111/ejn.16127.
Hämmerer, D., Li, S., Völkle, M., Müller, V., Lindenberger, U., 2013. A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. Psychophysiology 50, 111–123. https://doi.org/10.1111/j.1469-8986.2012.01476.x.

Hernandez-Pavon, J.C., Kugiumtzis, D., Zrenner, C., Kimiskidis, V.K., Metsomaa, J., 2022. Removing artifacts from TMS-evoked EEG: A methods review and a unifying theoretical framework. J. Neurosci. Methods 376. https://doi.org/10.1016/J.JNEUMETH.2022.109591.

Hernandez-Pavon, J.C., Veniero, D., Bergmann, T.O., Belardinelli, P., Bortoletto, M., Casarotto, S., et al., 2023. TMS combined with EEG: Recommendations and open issues for data collection and analysis. Brain Stimulat 16, 567–593. https://doi.org/10.1016/j.brs.2023.02.009.

Hopkins, W.G., 2000. Measures of reliability in sports medicine and science. Sports Med Auckl NZ 30, 1–15. https://doi.org/10.2165/00007256-200030010-00001.

Ilmoniemi, R.J., Virtanen, J., Ruohonen, J., Karhu, J., Aronen, H.J., Näätänen, R., et al., 1997. Neuronal responses to magnetic stimulation reveal cortical reactivity and connectivity. Neuroreport 8, 3537–3540. https://doi.org/10.1097/00001756-199711100-00024.

Julkunen, P., Jauhiainen, A.M., Westerén-Punnonen, S., Pirinen, E., Soininen, H., Könönen, M., et al., 2008. Navigated TMS combined with EEG in mild cognitive impairment and Alzheimer's disease: a pilot study. J. Neurosci. Methods 172, 270–276. https://doi.org/10.1016/J.JNEUMETH.2008.04.021.

Julkunen, P., Kimiskidis, V.K., Belardinelli, P., 2022. Bridging the gap: TMS-EEG from lab to clinic. J. Neurosci. Methods 369, 109482. https://doi.org/10.1016/J.JNEUMETH.2022.109482.

Kähkönen, S., Wilenius, J., Komssi, S., Ilmoniemi, R.J., 2004. Distinct differences in cortical reactivity of motor and prefrontal cortices to magnetic stimulation. Clin. Neurophysiol. 115, 583–588. https://doi.org/10.1016/J.CLINPH.2003.10.032.

Kappenman, E.S., Luck, S.J., 2011. The oxford handbook of event-related potential components. Oxf Handb Event-Relat Potential Compon 1–664. https://doi.org/10.1093/OXFORDHB/9780195374148.001.0001.

Kerwin, L.J., Keller, C.J., Wu, W., Narayan, M., Etkin, A., 2018. Test-retest reliability of transcranial magnetic stimulation EEG evoked potentials. Brain Stimulat. 11, 536–544. https://doi.org/10.1016/j.brs.2017.12.010.

Kirkovski, M., Rogasch, N.C., Saeki, T., Fitzgibbon, B.M., Enticott, P.G., Fitzgerald, P.B., 2016. Single pulse transcranial magnetic stimulation-electroencephalogram reveals no electrophysiological abnormality in adults with high-functioning autism spectrum disorder. J. Child Adolesc. Psychopharmacol. 26, 606–616. https://doi.org/10.1089/CAP.2015.0181.

Kobayashi, M., Pascual-Leone, A., 2003. Transcranial magnetic stimulation in neurology. Lancet Neurol. 2, 145–156. https://doi.org/10.1016/S1474-4422(03)00321-1.

Koch, G., Bonnì, S., Pellicciari, M.C., Casula, E.P., Mancini, M., Esposito, R., et al., 2018. Transcranial magnetic stimulation of the precuneus enhances memory and neural activity in prodromal Alzheimer's disease. Neuroimage 169, 302–311. https://doi.org/10.1016/J.NEUROIMAGE.2017.12.048.

Kumar, S., Zomorrodi, R., Ghazala, Z., Goodman, M.S., Blumberger, D.M., Cheam, A., et al., 2017. Extent of dorsolateral prefrontal cortex plasticity and its association with working memory in patients with Alzheimer disease. JAMA Psychiat. 74, 1266–1274. https://doi.org/10.1001/JAMAPSYCHIATRY.2017.3292.

Levit-Binnun, N., Litvak, V., Pratt, H., Moses, E., Zaroor, M., Peled, A., 2009. Differences in TMS-evoked responses between schizophrenia patients and healthy controls can be observed without a dedicated EEG system. Clin. Neurophysiol. 121, 332–339.

Lexell, J.E., Downham, D.Y., 2005. How to assess the reliability of measurements in rehabilitation. Am. J. Phys. Med. Rehabil. 84, 719. https://doi.org/10.1097/01.phm.0000176452.17771.20.

Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255. https://doi.org/10.2307/2532051.

Lioumis, P., Kičić, D., Savolainen, P., Mäkelä, J.P., Kähkönen, S., 2009. Reproducibility of TMS - evoked EEG responses. Hum. Brain Mapp. 30, 1387–1396. https://doi.org/10.1002/hbm.20608.

Luck, S.J., 2014. An introduction to the event-related potential technique. MIT Press.

Mancuso, M., Sveva, V., Cruciani, A., Brown, K., Ibáñez, J., Rawji, V., et al., 2021. Transcranial evoked potentials can be reliably recorded with active electrodes. Brain Sci. 11, 1–16. https://doi.org/10.3390/BRAINSCI11020145.

Manganotti, P., Acler, M., Masiero, S., Del Felice, A., 2015. TMS-evoked N100 responses as a prognostic factor in acute stroke. Funct. Neurol. 30, 125–130. https://doi.org/10.11138/FNEUR/2015.30.2.125.

Massimini, M., Ferrarelli, F., Sarasso, S., Tononi, G., 2012. Cortical mechanisms of loss of consciousness: insight from TMS/EEG studies. Arch. Ital. Biol. 150, 44–55. https://doi.org/10.4449/AIB.V150I2.1361.

McManus, I.C., 2012. The misinterpretation of the standard error of measurement in medical education: A primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. Med. Teach. 34, 569–576. https://doi.org/10.3109/0142159X.2012.670318.

Metsomaa, J., Sarvas, J., Ilmoniemi, R.J., 2014. Multi-trial evoked EEG and independent component analysis. J. Neurosci. Methods 228, 15–26. https://doi.org/10.1016/j.jneumeth.2014.02.019.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., et al., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63, 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Momi, D., Ozdemir, R.A., Tadayon, E., Boucher, P., Di Domenico, A., Fasolo, M., et al., 2021a. Perturbation of resting-state network nodes preferentially propagates to structurally rather than functionally connected regions. Sci. Rep. 11. https://doi.org/10.1038/S41598-021-90663-Z.

Momi, D., Ozdemir, R.A., Tadayon, E., Boucher, P., Di Domenico, A., Fasolo, M., et al., 2021b. Perturbation of resting-state network nodes preferentially propagates to structurally rather than functionally connected regions. Sci. Rep. 11, 12458. https://doi.org/10.1038/s41598-021-90663-z.

Mutanen, T.P., Kukkonen, M., Nieminen, J.O., Stenroos, M., Sarvas, J., Ilmoniemi, R.J., 2016. Recovering TMS-evoked EEG responses masked by muscle artifacts. Neuroimage 139, 157–166. https://doi.org/10.1016/j.neuroimage.2016.05.028.

Mutanen, T.P., Metsomaa, J., Liljander, S., Ilmoniemi, R.J., 2018. Automatic and robust noise suppression in EEG and MEG: The SOUND algorithm. Neuroimage 166, 135–151. https://doi.org/10.1016/j.neuroimage.2017.10.021.

Naim-Feil, J., Bradshaw, J.L., Rogasch, N.C., Daskalakis, Z.J., Sheppard, D.M., Lubman, D.I., et al., 2016. Cortical inhibition within motor and frontal regions in alcohol dependence post-detoxification: a pilot TMS-EEG study. World J. Biol. Psychiatry 17, 547–556. https://doi.org/10.3109/15622975.2015.1066512.

Napolitani, M., Bodart, O., Canali, P., Seregni, F., Casali, A., Laureys, S., et al., 2014. Transcranial magnetic stimulation combined with high-density EEG in altered states of consciousness. Brain Inj. 28, 1180–1189. https://doi.org/10.3109/02699052.2014.920524.

Noda, Y., Barr, M.S., Zomorrodi, R., Cash, R.F.H., Rajji, T.K., Farzan, F., et al., 2018a. Reduced short-latency afferent inhibition in prefrontal but not motor cortex and its association with executive function in schizophrenia: A combined TMS-EEG study. Schizophr. Bull. 44, 193–202.

Noda, Y., Zomorrodi, R., Vila-Rodriguez, F., Downar, J., Farzan, F., Cash, R.F.H., et al., 2018b. Impaired neuroplasticity in the prefrontal cortex in depression indexed through paired associative stimulation. Depress. Anxiety 35, 448–456. https://doi.org/10.1002/da.22738.

Noda, Y., Takano, M., Wada, M., Mimura, Y., Nakajima, S., 2024. Validation of the number of pulses required for TMS-EEG in the prefrontal cortex considering test feasibility. Neuroscience 554, 63–71. https://doi.org/10.1016/j.neuroscience.2024.07.011.

Ozdemir, R.A., Tadayon, E., Boucher, P., Momi, D., Karakhanyan, K.A., Fox, M.D., et al., 2020. Individualized perturbation of the human connectome reveals reproducible biomarkers of network dynamics relevant to cognition. PNAS 117, 8115–8125. https://doi.org/10.1073/pnas.1911240117.

Ozdemir, R.A., Boucher, P., Fried, P.J., Momi, D., Jannati, A., Pascual-Leone, A., et al., 2021a. Reproducibility of cortical response modulation induced by intermittent and continuous theta-burst stimulation of the human motor cortex. Brain Stimulat. 14, 949–964. https://doi.org/10.1016/j.brs.2021.05.013.

Ozdemir, R.A., Tadayon, E., Boucher, P., Sun, H., Momi, D., Ganglberger, W., et al., 2021b. Cortical responses to noninvasive perturbations enable individual brain fingerprinting. Brain Stimulat. 14, 391–403. https://doi.org/10.1016/J.BRS.2021.02.005.

Parmigiani, S., Ross, J.M., Cline, C.C., Minasi, C.B., Gogulski, J., Keller, C.J., 2023. Reliability and validity of transcranial magnetic stimulation–electroencephalography biomarkers. Biol. Psychiatry Cogn. Neurosci. Neuroimag. 8, 805–814. https://doi.org/10.1016/j.bpsc.2022.12.005.

Pavlov, Y.G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C.S.Y., Beste, C., et al., 2021. #EEGManyLabs: investigating the replicability of influential EEG experiments. Cortex J. Devoted Study Nerv. Syst. Behav. 144, 213–229. https://doi.org/10.1016/J.CORTEX.2021.03.013.

Pellicciari, M.C., Bonnì, S., Ponzo, V., Cinnera, A.M., Mancini, M., Casula, E.P., et al., 2018. Dynamic reorganization of TMS-evoked activity in subcortical stroke patients. Neuroimage 175, 365–378. https://doi.org/10.1016/j.neuroimage.2018.04.011.

Pernet, C.R., Appelhoff, S., Gorgolewski, K.J., Flandin, G., Phillips, C., Delorme, A., et al., 2019. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. Sci. Data 6, 1–5. https://doi.org/10.1038/s41597-019-0104-8.

Portney LG, Watkins MP. Foundations of Clinical Research: Applications to Practice. F.A. Davis Company; 2015.

Ragazzoni, A., Pirulli, C., Veniero, D., Feurra, M., Cincotta, M., Giovannelli, F., et al., 2013. Vegetative versus minimally conscious states: a study using TMS-EEG, sensory and event-related potentials. PLoS One 8. https://doi.org/10.1371/JOURNAL.PONE.0057069.

Ragazzoni, A., Cincotta, M., Giovannelli, F., Cruse, D., Young, B., Miniussi, C., et al., 2017. Clinical neurophysiology of prolonged disorders of consciousness: from diagnostic stimulation to therapeutic neuromodulation. Clin. Neurophysiol. 128, 1629–1646. https://doi.org/10.1016/j.clinph.2017.06.037.

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. https://doi.org/10.1016/j.brs.2020.10.011.

Rogasch, N.C., Sullivan, C., Thomson, R.H., Rose, N.S., Bailey, N.W., Fitzgerald, P.B., et al., 2017. Analysing concurrent transcranial magnetic stimulation and electroencephalographic data: a review and introduction to the open-source TESA software. Neuroimage 147, 934–951. https://doi.org/10.1016/j.neuroimage.2016.10.031.

Rogasch, N.C., Biabani, M., Mutanen, T.P., 2022. Designing and comparing cleaning pipelines for TMS-EEG data: a theoretical overview and practical example. J. Neurosci. Methods 371, 109494. https://doi.org/10.1016/j.jneumeth.2022.109494.

Rosanova, M., Casali, A., Bellina, V., Resta, F., Mariotti, M., Massimini, M., 2009. Natural frequencies of human corticothalamic circuits. J. Neurosci. 29, 7679–7685. https://doi.org/10.1523/JNEUROSCI.0445-09.2009.

Rossini, P.M., Burke, D., Chen, R., Cohen, L.G., 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. https://doi.org/10.1016/j.clinph.2015.02.001.

Salo, K.S.T., Mutanen, T.P., Vaalto, S.M.I., Ilmoniemi, R.J., 2020. EEG Artifact removal in TMS studies of cortical speech areas. Brain Topogr. 33, 1–9. https://doi.org/10.1007/S10548-019-00724-W.

Sarasso, S., Rosanova, M., Casali, A.G., Casarotto, S., Fecchio, M., Boly, M., et al., 2014. Quantifying cortical EEG responses to TMS in (un)consciousness. Clin. EEG Neurosci. 45, 40–49. https://doi.org/10.1177/1550059413513723.

Schambra, H.M., Ogden, R.T., Martínez-Hernández, I.E., Lin, X., Chang, Y.B., 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. https://doi.org/10.3389/fncel.2015.00335.

She, X., Nix, K.C., Cline, C.C., Qi, W., Tugin, S., He, Z., et al., 2024. Stability of transcranial magnetic stimulation electroencephalogram evoked potentials in pediatric epilepsy. Sci. Rep. 14, 9045. https://doi.org/10.1038/s41598-024-59468-8.

Shrout, P.E., 1998. Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. 7, 301–317. https://doi.org/10.1177/096228029800700306.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037//0033-2909.86.2.420.

Song, Y., Gordon, P.C., Metsomaa, J., Rostami, M., Belardinelli, P., Ziemann, U., 2024. Evoked EEG responses to TMS targeting regions outside the primary motor cortex and their test-retest reliability. Brain Topogr. 37, 19–36. https://doi.org/10.1007/s10548-023-01018-y.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use (5th edition). Aust N Z J Public Health 2016;40:294–5. Doi: 10.1111/1753-6405.12484.

ter Braack, E.M., de Goede, A.A., van Putten, M.J.A.M., 2019. Resting motor threshold, MEP and TEP variability during daytime. Brain Topogr. 32, 17–27. https://doi.org/10.1007/S10548-018-0662-7.

Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A.W.M., Knol, D.L., Dekker, J., et al., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

Tremblay, S., Rogasch, N.C., Premoli, I., Blumberger, D.M., Casarotto, S., Chen, R., et al., 2019. Clinical utility and prospective of TMS–EEG. Clin. Neurophysiol. 130, 802–844. https://doi.org/10.1016/j.clinph.2019.01.001.

Vaz, S., Falkmer, T., Passmore, A.E., Parsons, R., Andreou, P., 2013. The case for using the repeatability coefficient when calculating test–retest reliability. PLoS One 8. https://doi.org/10.1371/journal.pone.0073990 e73990.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., et al., 2017. The Alzheimer's disease neuroimaging initiative 3: continued innovation for clinical trial improvement. Alzheimers Dement J. Alzheimers Assoc 13, 561–571. https://doi.org/10.1016/J.JALZ.2016.10.006.

Weir, J.P., 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. 19, 231–240. https://doi.org/10.1519/15184.1.

Wolf, S.L., Butler, A.J., Campana, G.I., Parris, T.A., Struys, D.M., Weinstein, S.R., et al., 2004. Intra-subject reliability of parameters contributing to maps generated by transcranial magnetic stimulation in able-bodied adults. Clin. Neurophysiol. 115, 1740–1747. https://doi.org/10.1016/j.clinph.2004.02.027.

Wu, W., Keller, C.J., Rogasch, N.C., Longwell, P., Shpigel, E., Rolle, C.E., et al., 2018. ARTIST: A fully automated artifact rejection algorithm for single-pulse TMS-EEG data. Hum. Brain Mapp. 39, 1607–1625. https://doi.org/10.1002/hbm.23938.

Ye, S., Kitajo, K., Kitano, K., 2019. Information-theoretic approach to detect directional information flow in EEG signals induced by TMS. Neurosci. Res. https://doi.org/10.1016/j.neures.2019.09.003.

Ye, S., Kitajo, K., Kitano, K., 2020. Information-theoretic approach to detect directional information flow in EEG signals induced by TMS. Neurosci Res 156, 197–205. https://doi.org/10.1016/j.neures.2019.09.003.

Zazio, A., Barchiesi, G., Ferrari, C., Marcantoni, E., Bortoletto, M., 2022. M1-P15 as a cortical marker for transcallosal inhibition: a preregistered TMS-EEG study. Front. Hum. Neurosci. 16. https://doi.org/10.3389/FNHUM.2022.937515.

Further reading

Rosanova M, Gosseries O, Casarotto S, Lanie Boly M, Casali AG, Lie Bruno M-A, et al. Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. J Neurol n.d. Doi: 10.1093/brain/awr340.