The reliability of repeated TMS measures in older adults and in patients with subacute and chronic stroke

Heidi M. Schambra[1]*, R. Todd Ogden[2], Isis E. Martínez-Hernández[1], Xuejing Lin[2], Y. Brenda Chang[2], Asif Rahman[3], Dylan J. Edwards[4†] and John W. Krakauer[5†]

[1] Motor Performance Laboratory, Department of Rehabilitation and Regenerative Medicine, Columbia University, New York, NY, USA, [2] Department of Biostatistics, Columbia University, New York, NY, USA, [3] Neural Engineering Group, Department of Biomedical Engineering, City College of New York, NY, USA, [4] Non-Invasive Brain Stimulation and Human Motor Control Laboratory, Burke-Cornell Medical Research Institute, White Plains, NY, USA, [5] Brain, Learning, Animation, and Movement Lab, Department of Neurology, Johns Hopkins University, Baltimore, MD, USA

The reliability of transcranial magnetic stimulation (TMS) measures in healthy older adults and stroke patients has been insufficiently characterized. We determined whether common TMS measures could reliably evaluate change in individuals and in groups using the smallest detectable change (SDC), or could tell subjects apart using the intraclass correlation coefficient (ICC). We used a single-rater test-retest design in older healthy, subacute stroke, and chronic stroke subjects. At twice daily sessions on two consecutive days, we recorded resting motor threshold, test stimulus intensity, recruitment curves, short-interval intracortical inhibition, and facilitation, and long-interval intracortical inhibition. Using variances estimated from a random effects model, we calculated the SDC and ICC for each TMS measure. For all TMS measures in all groups, SDCs for single subjects were large; only with modest group sizes did the SDCs become low. Thus, while these TMS measures cannot be reliably used as a biomarker to detect individual change, they can reliably detect change exceeding measurement noise in moderate-sized groups. For several of the TMS measures, ICCs were universally high, suggesting that they can reliably discriminate between subjects. TMS measures should be used based on their reliability in particular contexts. More work establishing their validity, responsiveness, and clinical relevance is still needed.

Keywords: TMS, reliability, standard error of the measurement, measurement error, smallest detectable change, ICC, biomarker

Introduction

Over the past three decades, increasing numbers of studies have used transcranial magnetic stimulation (TMS) to examine neurophysiology in pathology and health. Single- and paired-pulse TMS techniques probe the motor cortex and its connections, but do not themselves alter overall brain excitability (Kujirai et al., 1993; Nakamura et al., 1997; Chen, 2000, 2004; Chen and Udupa, 2009). It is generally assumed that change in the average motor evoked potential (MEP) amplitude, resulting from the same stimulation intensities, reflects true biological change in the corticospinal

tract (CST), and/or intracortical circuitry (Rothwell, 2010). Although TMS has great potential for establishing physiological biomarkers, a rigorous appraisal of the quality of TMS as a measurement instrument has never been sufficiently undertaken. The instrument of TMS encapsulates the stimulation and recording setup and the operator. As a field, we have not established the degree to which TMS measurements are precise, accurate, or clinically relevant. Without this foundational knowledge, we risk ascribing meaningful neurophysiological mechanisms to meaningless TMS changes.

There have been hundreds of studies since the early 1990s using TMS to investigate neurophysiological change in healthy adults and stroke patients. To our knowledge, none placed their results in the context of the measurement error associated with TMS, instead judging results on the statistical assessment of differences associated with an intervention or time. The pervasive use of TMS, without questioning the quality of the measurements it produces, has perhaps led to a false sense of security that TMS is a superior measurement instrument. Work to prove this point, however, is largely lacking.

How do we judge if TMS is a good instrument for assessing neurophysiology? A useful measurement instrument is reliable and valid, producing data that are accurate and meaningful (Portney and Watkins, 2009). To gauge an instrument's utility, we must know how it fares in three main *domains* of instrument quality: reliability, validity, and responsiveness (Mokkink et al., 2010). In other words, TMS measures should produce stable measurements in unchanging subjects, the measures should tell us something about neurophysiology, and the measures should be able to detect change in neurophysiology. These instrument qualities should furthermore be characterized for different subject populations. In this study, we focused on characterizing the first quality domain of TMS—reliability—in older healthy subjects and stroke patients.

The lack of a comprehensive instrument assessment is not unique to TMS. Across health outcomes research, a major barrier to evaluating an instrument's utility has been disagreement about the terminology, definitions, and computations associated with each quality domain (Mokkink et al., 2010). For example, the term "reliability" has been used interchangeably with variability, consistency, reproducibility, precision, repeatability, agreement, and stability; depending on context and application, some of these terms represent distinct concepts operationalized by different mathematical formulae (Mokkink et al., 2010). The lack of a coherent vocabulary and taxonomy has led to a considerable misunderstanding of concepts, misuse of calculations, and misapplication of results for judging instrument quality, with TMS being no exception. In light of the confusion, we adopt the language developed by experts in clinimetrics, the methodological discipline focusing on the quality and use of measures in clinical medicine (de Vet et al., 2003; Mokkink et al., 2010).

Reliability is the degree to which repeated measures in unchanging individuals provide similar results (de Vet et al., 2006). A reliable instrument produces measurements that are consistent and error-free (Portney and Watkins, 2009; Mokkink et al., 2010). In stable individuals, reliability is the degree to which

these measurements are the same over time (i.e., test-retest), as assessed by the same rater at different times (i.e., intra-rater), or as assessed by different raters at the same time (i.e., inter-rater) (Portney and Watkins, 2009; Mokkink et al., 2010).

For measurements taken at test-retest, reliability can address two different questions: how unchanging measurements are within individuals, or how unchanging individuals are relative to others. The domain of reliability is thus an umbrella term for distinct subtypes, so-called "measurement properties," of reliability. The first measurement property of reliability is called measurement error, which assesses how good the agreement is between repeated measurements in an individual. It tells us how close measurements are with repeated testing in stable individuals (de Vet et al., 2006).

The second measurement property of reliability is called reliability$_{MP}$, which assesses how well individuals can be distinguished from one another. It tells us how unchanging the positions of stable individuals are relative to each other at repeated testing (de Vet et al., 2006; Streiner and Norman, 2008). [N.B.: The duplicate use of the term "reliability" for both the *domain* of reliability and one of its *measurement properties* is undeniably confusing; despite debate, however, this duplication was upheld in light of historical usage (Mokkink et al., 2010). For clarity, we specify the *measurement property* of reliability with the subscript "MP."].

When deciding which reliability measurement property to use for evaluating an instrument, the instrument's intended application needs to be considered (Guyatt et al., 1987; de Vet et al., 2006). If TMS is to be used for evaluation (e.g., "did Patient X's measurements change after an intervention?"), a small measurement error is needed (Guyatt et al., 1987; de Vet et al., 2006; Terwee et al., 2007). If TMS is instead to be used for diagnosis or staging (e.g., "is Patient X sicker than Patient Y?"), a high reliability$_{MP}$ is needed (de Vet et al., 2006, 2011).

The majority of research with TMS uses it in an evaluative manner, for example, measuring neurophysiological change associated with time or an intervention. Knowing the measurement error associated with TMS is therefore critical for assessing its reliability. Knowing the smallest detectable change (SDC), derived from the measurement error, is its pragmatic extension. The SDC is the amount of change necessary to exceed measurement error (Beckerman et al., 2001). When the SDC is unknown, one cannot know if a change in measurements reflects a true change or simply measurement noise. An observed change less than the threshold value of SDC would be classified as measurement error.

Although generally not used in this manner, TMS also has diagnostic potential, for example, grading (i.e., staging) the severity of a lesion in the corticospinal tract. For this, reliability$_{MP}$ should be known. Of note, most of the studies evaluating "reliability" in TMS have documented only reliability$_{MP}$, despite TMS's primarily evaluative use.

We are not the first to undertake the important task of characterizing TMS reliability, but previous attempts in our view are questionable with respect to their methodologies and interpretations. For example, Pearson's correlation coefficient (r) has been used to identify the strength of association

between test-retest values (Balslev et al., 2007). While correlation reflects the association between two measurements, it does not reveal the amount of sameness (i.e., agreement) between them. For instance, in a sample whose measurements are twice the magnitude at retest, correlation between measurements is perfect (i.e., $r = 1$), but there is no agreement. Others have used a Cronbach's alpha to document test-retest reliability (Farzan et al., 2010). Cronbach's alpha characterizes the consistency of multiple distinct measures probing the same construct, for example, how well the scores of the first half of questions correlate with the second half in a questionnaire (Mokkink et al., 2010; de Vet et al., 2011). Cronbach's alpha is not appropriate for assessing agreement between individual measurements from a single TMS measure over time.

Another common approach has been the use of inferential statistics to draw conclusions about measurement agreement. T-tests and ANOVAs have been used to test for statistical differences between repeated TMS measures in younger healthy subjects (McMillan et al., 1998; Boroojerdi et al., 2000; Maeda et al., 2002; Uy et al., 2002; Corneal et al., 2005), older healthy subjects (Wolf et al., 2004), and chronic stroke subjects (Liepert et al., 2000a; Butler et al., 2005). Based on non-significant differences between mean measurements, investigators have concluded that the measures are stable and reliable. Carl Sagan's warning is appropriate here: "the absence of evidence is not evidence of absence" (Sagan, 1995). Not detecting significant differences between measurements does not prove they are the same. Two or more distributions may show no statistical difference, but may be composed of pairs of measurements with no agreement. Thus, a non-significant difference between measurements does not imply high agreement or low measurement error.

A final problematic approach has been the widespread misunderstanding and misapplication of reliability$_{MP}$. A high reliability$_{MP}$ estimate is commonly misinterpreted as signifying low measurement error, with investigators deeming the TMS measure appropriate for evaluative use (Mortifee et al., 1994; Carroll et al., 2001; Kamen, 2004; Malcolm et al., 2006; Christie et al., 2007; Koski et al., 2007; Cacchio et al., 2009, 2011; Wheaton et al., 2009; Hoonhorst et al., 2014; Liu and Au-Yeung, 2014). This is simply wrong; the reliability measurement properties are not interchangeable and each implies a distinct concept. Additionally, reliability$_{MP}$ is highly influenced by the dispersion of subjects in a sample. A sample with large between-subject variability will produce a high reliability$_{MP}$ estimate for a measure, even despite a sizable within-subject measurement error. The influence of the sample's dispersion on a measure's reliability$_{MP}$ thus constrains its generalizability: the reliability$_{MP}$ estimate is appropriate for use only in samples with similar heterogeneity, a point that is rarely underscored.

In this study, we sought to establish the reliability of common TMS measures taken repeatedly from the FDI representation of healthy older adults and subacute and chronic stroke subjects. For TMS's evaluative and diagnostic applications, we estimated measurement error and reliability$_{MP}$, respectively. We

additionally derive the SDC for TMS measures, for practical use in future evaluative studies. We endeavored to make this manuscript operate as a primer for how one goes about assessing the qualities of a TMS measure, which requires detailed descriptions of TMS techniques and rigorous understanding of reliability assessments. It is our hope that with its transparency and detail, the manuscript can serve as a template for future reliability studies in TMS.

Materials and Methods

General Approach

Four identical testing sessions were used to assess test-retest reliability of MEP recordings from TMS applied to bilateral primary motor cortices. Subjects were studied on two consecutive days, in morning and afternoon sessions, 3.5–5 h apart. Sessions were performed at the same time each day within subject. Each session lasted 1.5–2 h. No experimental interventions were given between sessions.

Subject testing was conducted in the Motor Performance Laboratory at Columbia University and in the Non-Invasive Brain Stimulation and Human Motor Control Laboratory at Burke Rehabilitation Center. The study was approved by each facility's Institutional Review Board. The laboratories used identical hardware, software, equipment, and supplies in their neurophysiology setups, and differed only in the subject's chair and window view. A single operator (HS) conducted the assessments and data analysis. Each subject was tested at one site only.

Subjects

Three separate groups of adults (total $n = 62$) were studied: healthy older adults ($n = 21$), subacute stroke subjects ($n = 20$), and chronic stroke subjects ($n = 21$). Healthy older adults were included because they often serve as stroke study controls, given the increased incidence of stroke with age (Sacco et al., 1997), and because their neurophysiology is different from that of younger adults (Rossini et al., 1992; Peinemann et al., 2001; McGinley et al., 2010). Subacute and chronic stroke groups were used because both recovery epochs have been probed with TMS (Bütefisch et al., 2003, 2008; Hummel et al., 2005; Hummel and Cohen, 2005; Liepert, 2006; Swayne et al., 2008; Khedr et al., 2009, 2013; Takechi et al., 2014). All subacute strokes, 5 chronic stroke, and 6 healthy older subjects were tested at Burke; all other chronic stroke and healthy older subjects were tested at Columbia.

Subjects were included if they were ≥40 years old, were able to give informed consent, and if stroke patients, had a single ischemic stroke resulting in paresis <6 months (subacute stroke group) or ≥6 months previously (chronic stroke group). Because we wanted to obtain TMS outcomes in both hemispheres, only subjects who could at least marginally abduct their paretic index finger (MRC $\geq$ 1) and who had a recordable TMS-evoked response were included. Psychoactive medications (SSRI, SNRI, or bupropion) were permitted as long as they were taken consistently. Stroke subjects were excluded for preserved motor strength in the upper extremity (i.e., full strength on

manual motor testing including no pronator drift, orbiting, or reduced finger individuation) or receptive aphasia; all subjects were excluded for any major medical, psychiatric, or non-stroke neurological condition that could interfere with motor function assessment or participation; history of seizure, neurosurgery, traumatic brain injury, or substance abuse; or thoracic or intracranial metal objects, implants, or devices, except for dental work. All subjects gave written informed consent to participate, in accordance with the Declaration of Helsinki.

Psychophysical and Clinical Characteristics

Following each session, subjects reported levels of alertness during the session and excitement to participate on a scale from 1 to 10, with 10 as maximum. They also reported quantity of pre-testing sleep, exercise, and caffeine intake. Following session 3, subject demographics, clinical history, and handedness [Edinburgh Handedness Inventory; +1 and −1 indicate dominance for right and left hand, respectively (Oldfield, 1971)] were obtained. A neurological examination of the upper extremities, including the assessment of bilateral upper extremity strength by Medical Research Council (MRC) scale (Medical Research Council of U.K, 1978), was also performed at that time by a neurologist (HS). Stroke risk factors (hypertension, atrial fibrillation, coronary artery disease, diabetes, hypercholesterolemia, and tobacco use) and current medication use were documented.

TMS Measures

The two neurophysiology laboratories had an identical setup and equipment. Subjects were seated comfortably in an office chair with their forearms relaxed on a lap pillow. Arms were consistently positioned across sessions. Frameless stereotaxic equipment (Brainsight, Rogue Research, Canada), used to co-register the subject's scalp positions with a phantom MRI brain image, ensured stimulation accuracy during and across sessions. Co-registration errors to the phantom's surface landmarks were matched to ≤3 mm at each follow-up session.

Surface EMG was obtained from bilateral first dorsal interosseous (FDI) muscles, with electrodes taped in a belly-tendon orientation (SX230-100 and K800; Biometrics Ltd, UK). The integrated electrode contains two poles at a fixed distance of 2 cm. To ensure consistent electrode placement across sessions, electrodes were outlined with permanent ink on the skin and subjects were advised not to scrub the area. The EMG signal was sampled at 1000 Hz, amplified 1000x, band-pass filtered at 15–450 Hz, and saved for offline analysis. All assessments were taken at rest, and EMG activity was monitored online to ensure muscle relaxation.

TMS was delivered to the cortical hand representation of the motor cortex (M1), using Magstim BiStim[2] and a 70-mm figure-of-eight remote control coil (Magstim Company Ltd, UK). Stimulation intensity determinations and data acquisition were performed in BiStim mode. Pulses were generated using specialized software (Signal; Cambridge Electronic Devices, UK) and a 1401 microprocessor (Cambridge Electronic Devices, UK). The TMS coil was held tangentially to the skull with the coil handle pointed 45° posterior-laterally to the sagittal plane, which

orients the coil approximately perpendicular to the central sulcus and induces a posterior-to-anterior current direction (Pascual-Leone et al., 1994; Ruohonen and Ilmoniemi, 1999). A search in a ∼1 cm-step grid pattern at ∼50 percent of the maximum stimulator output (%MSO) was conducted to grossly identify the area producing the largest amplitude MEP (the "hotspot") for the contralateral first dorsal interosseous (FDI) muscle. Resting motor threshold (rMT), the %MSO eliciting at least 5 out of 10 MEPs ≥50 μV, was obtained at that location. Using this %MSO and centering on this preliminary hotspot, a repeated grid search and new rMT determination refined the localization of the hotspot. The hotspot was virtually marked on the phantom brain and used at subsequent sessions, though its position was confirmed physiologically each time. If a superior hotspot was found at retest, this position was marked and followed. Almost always, the hotspot did not change; occasionally, the coil position required a minor adjustment in roll or pitch.

In healthy subjects, the left FDI hotspot was always probed first. In stroke subjects, the non-lesioned FDI hotspot was always probed first. The order of the TMS measures was purposefully fixed so that an order effect, if present, would have a consistent influence on measurement variability. If an order effect were present, for example from fatigue, then randomization would introduce additional measurement noise, i.e., by testing a subject who is alert for the measure at one session and fatigued at the next. Our approach was also most representative of a typical pre-post testing design, in which the order of outcome measures is not varied across sessions or individuals. We chose TMS measures based on their common usage in TMS studies, not necessarily based on their potential validity.

All TMS stimuli were delivered at an inter-trial interval of 7 s. Recruitment curves (RC) were generated from 10 stimuli each given at ascending stimulation intensities of 100, 110, 130, 150, and 170% rMT, or until 100% MSO was reached. MEP amplitudes were fitted to a Boltzmann sigmoid function (Carroll et al., 2001)

$$y(x) = \frac{plateau}{1 + e^{\frac{S_{50} - x}{slope}}}$$

to estimate component RC parameters. Recruitment curve *plateau* is the maximum amplitude (mV), S_{50} is the stimulus intensity x (%MSO) required to evoke a response equal to half the maximum amplitude, and *slope* is the MEP amplitude increase with each percentage point of stimulator intensity increase (mV/%MSO). Model parameters ($\vec{\theta} = \{plateau, S_{50}, slope\}$) were estimated with a standard least squares curve-fitting algorithm (trust-region-reflective algorithm) in MATLAB 8.1 (MathWorks Inc., USA). The parameter estimation method minimizes the root mean square error

$$\vec{\theta} = \underset{\vec{\theta}}{\arg\min} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[d_i - y(x_i|\theta) \right]^2}$$

between the actual MEP amplitudes (*d*) and the estimated amplitude $y(x|\theta)$ given a set of parameters. The amplitude of *plateau* was bounded to ≤8 mV to provide a physiologically

plausible approximation of the maximum MEP amplitude. All other parameters were left unconstrained.

Short-interval intracortical inhibition (SICI) and intracortical facilitation (ICF) were generated with a conditioning stimulus (CS) delivered 2 or 10 ms prior to a test stimulus (TS), respectively (Kujirai et al., 1993). SICI and ICF were tested at CS intensities of 60 and 80% rMT to identify which CS intensity produced more reliable measurements; these measures are henceforth specified as $SICI_{60}$ and ICF_{60} or $SICI_{80}$ and ICF_{80}. For TS, the stimulator intensity (TS_{MSO}) was adjusted to produce an MEP (TS_{MEP}) ~1 mV in amplitude. If this size could not be achieved, particularly in stroke subjects, the TS_{MSO} was set to the stimulation intensity above which no further increases in TS_{MEP} amplitude could be found (Swayne et al., 2008). Ten trials each of TS_{MEP}, SICI, and ICF were recorded in repeating order for each CS intensity. Long-interval intracortical inhibition (LICI) was obtained with two stimuli separated by 100 ms (Nakamura et al., 1997). Both stimuli followed the same TS_{MEP} determination as above. Ten trials each of alternating TS_{MEP} and LICI were recorded. For SICI, ICF, and LICI, the average amplitude of conditioned MEPs were normalized to the average amplitude of unconditioned TS_{MEP} [i.e., (CS+TS)/TS], and are reported as a decimal fraction of TS_{MEP}.

For reliability assessments of TS_{MEP}, the MEP amplitudes of the 30 TS trials in each session (i.e., the three sets of 10 TS trials used to calculate $SICI_{60}/ICF_{60}$, $SICI_{80}/ICF_{80}$, and LICI separately) were averaged. For TS_{MSO}, only the first %MSO was used, in order to exclude stimulator intensity adjustments that may have occurred within session.

Following TMS assessment in a hemisphere, pinch force was obtained from its contralateral hand. Subjects sat with shoulder adducted, elbow flexed at 90°, forearm midway between pronation and supination, and wrist in ~15° extension. Subjects held the force transducer (P200; Biometrics, UK) between the pad of thumb and radial side of the flexed index finger (i.e., a lateral or key pinch). Three maximal voluntary contractions (MVC) were held for 3 s each, with 10–20 s rest between, and stored offline. The maximum voltage of the force was extracted with a custom-made Signal script, and a conversion of 11.34 kg/V was applied. The trials were averaged within and then across sessions for each hand.

TMS Measurement Analysis

For TMS measurements, peak-to-peak MEP amplitude was measured using a custom-made script (Gray, 2015). Trials were discarded if EMG activity exceeded 100 μV in the 250 ms prior to TMS stimulus delivery. Three additional stimuli to be discarded were built into the beginning of each assessment, to eliminate the influence of excessively large MEPs commonly seen with initial stimuli and to allow time for coil positioning.

For RC parameter estimations, the curve fits to the raw MEP amplitudes for each session were visually examined to confirm the appropriate fit of the model to the data. The hemisphere's RC parameter estimations were discarded for any session where the curve appeared exponential or the plateau value equaled the upper parameter boundary (plateau = 8 mV).

Reliability Measurement Properties
Variance Decomposition

Classical test theory (CTT) postulates that an observed measurement is composed of the true measurement and an error term of the measurement (Lord and Novick, 1968). By CTT assumptions, the variance of the observed measurement is composed of the variance in the true measurement and the variance in the error term (de Vet et al., 2011):

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error}$$

Since the available data involve multiple observations on the same subjects, it is possible to decompose the variances of measurements from each hemisphere into several components and, by fitting statistical models that include random effects, to obtain estimates for the variance of each component. In particular, the "error" component can consist of between-subjects variability of measurements made on different subjects, day-to-day variability of measurements made on the same subject, replication variability of measurements made on the same subject within the same day, and "residual" variance, which captures all other sources of variability.

Based on our analysis of the data, we did not include the replication variance (measurements made on the same subject within the same day) in our final models. Because there are only two replications per day per subject, there are insufficient data to obtain reliable estimates of a time effect. Furthermore, estimates of this variance component were consistently very small. Therefore, we omitted the random effect of time in the model, effectively combining that variance with the residual variance.

For the purposes of calculating reliability measurement properties, a random effects model was used. Fixed effects, if any, should be considered a part of measurement error (de Vet et al., 2006). In most experimental contexts, testing days, and raters are assumed to be randomly chosen and representative of their class, and are expected to have similar fixed effects in future studies (de Vet et al., 2006, 2011). Excluding systematic differences could be appropriate if ranking subjects over time is intended (Terwee et al., 2007), but this has never been an application of TMS measures.

In our model, the variance of our measures was decomposed for each hemisphere as:

$$\sigma^2_{observed} = \sigma^2_{subjects} + \sigma^2_{days} + \sigma^2_{residual}$$

where $\sigma^2_{subjects}$ is the between-subject variance, σ^2_{days} is the between-days variance, and $\sigma^2_{residual}$ is the error term that captures the remaining unexplained variability and includes interactions between all predictors (McGraw and Wong, 1996; de Vet et al., 2006).

Measurement Error: Standard Error of the Measurement (SEMeas) and Relative SEMeas (SEMeas%)

The decomposed variances were used to estimate measurement error. Measurement error, also known as agreement (de Vet

et al., 2006), is the spread of repeated measurements due to systematic and random error, not due to true changes in the construct being measured (Terwee et al., 2007; Mokkink et al., 2010). In other words, the measurement error estimates the "spread" or "noisiness" of repeated measures within stable individuals. The smaller the measurement error, the less variable and the more reliable the measure (Harvill, 1991; Tighe et al., 2010). The measurement error is particular to the measure and the population from which it is taken, not to the instrument universally. One would therefore expect the measurement errors of TMS measures to differ from one another, and for their measurement errors to change with pathology, age, and cortical muscle representation.

Measurement error is estimated by the standard error of the measurement (SEMeas; N.B.: it is generally abbreviated as SEM, but we chose to permute the acronym to avoid confusion with standard error of the mean.) SEMeas is the standard deviation of all within-subject sources of variance and excludes between-subject variance. It is calculated as:

$$SEMeas = \sqrt{\sigma_{day}^2 + \sigma_{residual}^2}$$

The SEMeas has some unique attributes that make it particularly helpful in assessing instrument reliability. First, SEMeas is uninfluenced by the heterogeneity of sample from which it was derived, because between-subject variability is not included in its estimation (Tighe et al., 2010). Thus, it is considered a concrete property of the measure for the sampled population (Nunnally and Bernstein, 1994; Weir, 2005). Second, the SEMeas is largely stable across the spectrum of measurements for a measure (Harvill, 1991; Nunnally and Bernstein, 1994). Finally, the SEMeas is expressed in the same metrics as the measurement, providing ease of interpretation (Nunnally and Bernstein, 1994; Wyrwich et al., 1999).

The relative SEMeas (SEMeas%) is similar in concept to coefficient of variation in that, normalizing to the measurement mean, it provides the relative size of a measure's measurement error (Lexell and Downham, 2005). It is calculated as:

$$SEMeas\% = \frac{SEMeas}{mean} * 100$$

We use it here solely to inspect the relative noisiness of various measures.

Smallest Detectable Change (SDC)

The SEMeas is used to estimate the SDC for the measure (Beckerman et al., 2001). Unlike the SEMeas, which estimates the potential scatter of repeated measurements, the SDC estimates amount of change an observation would need to exceed that expected scatter. The SDC is the smallest change in score that, with some degree of certainty, can be declared a real change above the measurement error (Beckerman et al., 2001; Terwee et al., 2007). If a measurement changes by less than the SDC, it is assumed to be measurement noise. Of note, "real change" does not imply that the change validly reflects a changing construct or is clinically meaningful, issues we discuss later. SDC is also known as the smallest real difference, true change,

minimal detectable difference and minimal detectable change, and is conceptually similar to the limits of agreement (Bland and Altman, 1986). The SDC can determined for an individual (SDC$_{indiv}$) or a group (SDC$_{group}$).

The SDC for an individual is the 95% confidence interval of the SEMeas of the change scores (Schuck and Zwingmann, 2003; Terwee et al., 2007), and is calculated as:

$$SDC_{indiv} = SEMeas * \sqrt{2} * 1.96$$

where $\sqrt{2}$ accounts for the variances associated with 2 independent sessions (e.g., pre, post) used to calculate the change score (see derivation in Schuck and Zwingmann, 2003), and 1.96 represents a 95% confidence interval, assuming normally distributed scores. The SDC for a group is based on the individual SDC (Terwee et al., 2007; de Vet et al., 2011), and is calculated as:

$$SDC_{group} = \frac{SDC_{indiv}}{\sqrt{n}}$$

Like the SEMeas, the SDC is stable across the spectrum of measurements for a measure, as long as change scores are not heteroscedastic. In other words, once assured of change score homoscedasticity, the absolute SDC can be applied to measurements regardless of where they lie in the spectrum.

Reliability$_{MP}$: Intraclass Correlation Coefficient (ICC)

The decomposed variances were also used to estimate reliability$_{MP}$. Reliability$_{MP}$ assesses how well the measure can tell subjects apart despite measurement error (de Vet et al., 2006, 2011; Terwee et al., 2007; Streiner and Norman, 2008). It gauges the consistency of an individual's position relative to others in a group assessed at test-retest or between raters (Weir, 2005; Streiner and Norman, 2008). Reliability$_{MP}$ is estimated using the Intraclass Correlation Coefficient (ICC), and can be calculated using several different formulae, depending on the intended interpretation and application of the results (Shrout and Fleiss, 1979; McGraw and Wong, 1996). Generally speaking, the ICC is the proportion of between-subject variance to all sources of variance (Shrout and Fleiss, 1979; Mokkink et al., 2010). For our study, where test-retest measurements were taken on separate days with a single TMS operator, the ICC is calculated as:

$$ICC_{agreement} = \frac{\sigma_{subjects}^2}{\sigma_{subjects}^2 + \sigma_{day}^2 + \sigma_{residual}^2}$$

This formula is identical to ICC (2, k) or (A,k) (Shrout and Fleiss, 1979; McGraw and Wong, 1996; Streiner and Norman, 2008).

ICC values range from 0 to 1, with 1 being highest possible reliability$_{MP}$. Guidelines for the interpretation of ICC suggest that a value >0.70 is acceptable reliability$_{MP}$, but acknowledge that this threshold was arbitrarily demarcated (Portney and Watkins, 2009). Because of the way the ICC is calculated, a defining feature is the influence of sample heterogeneity—the larger the between-subject variance, $\sigma_{subjects}^2$, the higher the ICC. This is not to say that measurement error (i.e., all other sources

of variance) has *no* influence on reliability$_{MP}$, just that the between-subject variance has relatively more (de Vet et al., 2006). Given the interpretation of reliability$_{MP}$, this calculation makes sense: the wider the dispersion of the subjects, the better the measure will be at telling subjects apart, almost regardless of the measurement noise within-subject. Conversely, if the sample's subjects are narrowly distributed, the instrument will have difficulty telling subjects apart, even with a small measurement error (de Vet et al., 2006). As such, reliability$_{MP}$ is not an intrinsic property of the measure, but is rather the combined property of the measure and the subject sample in which it was tested (Tighe et al., 2010).

Statistical Methods

Bland-Altman plots were used to inspect the homoscedasticity of change scores for the measures made on each hemisphere in each group (Bland and Altman, 1986). Differences between morning and afternoon sessions for each subject were plotted against the mean score of the two sessions. Heteroscedasticity, determined by visual inspection, was noted for RC slope and LICI. For these measures, group comparisons and assessment of reliably can be made, but only after appropriately transforming the observations. The most common transformation in such a situation is the log transform; thereafter, comparative differences, SEMeas, SDC, and ICC can still be computed and reported, but are interpreted for the logarithmically transformed values of the measure. Also, because log-transformed data are no longer on a ratio scale, we do not calculate ratios such as SEMeas% for LICI and RC slope.

For continuous demographic data, group means and standard deviations were obtained by averaging outcomes across subjects. For TMS measures, group means and standard deviations for each hemisphere were obtained by first averaging each subject's measurements for the 4 sessions, then averaging across the 20–21 subjects within group.

For the comparison of categorical demographic data, Fisher's exact test was used; for continuous demographic data, a one-way ANOVA with *post-hoc* group-wise comparisons was used, with a Bonferroni adjustment for multiple comparisons between subject groups.

For TMS data, mean outcomes from each hemisphere were compared within and across groups. The outcomes of the healthy older hemispheres were averaged to create a "healthy control" hemisphere, against which the lesioned and non-lesioned hemispheres of stroke subjects were separately compared. For the heteroscedastic measures (LICI, RC slope), comparisons were performed on the logarithmically transformed data. For TMS data, a one-way ANOVA with *post-hoc* group-wise comparisons was used, followed by a Bonferroni adjustment for multiple comparisons. Corrections were made for comparisons between hemispheres within-group (e.g., lesioned chronic vs. non-lesioned chronic), type of hemisphere in one stroke group vs. same type of hemisphere in the other stroke group (e.g., lesioned chronic vs. lesioned subacute), and hemisphere in one stroke group vs. the "healthy control" hemisphere (e.g., lesioned chronic vs. healthy control). Significance was set at $p < 0.05$.

Confidence intervals were obtained for the SEMeas, SDC, and ICC for each measure using a standard bootstrapping resampling algorithm. Standard errors of the reliability estimates rely very heavily on properties of the assumed normal distribution, while bootstrap-based intervals are non-parametic and are valid for any distribution (Efron and Tibshirani, 1993).

Results

Clinical and Psychometric Characteristics

Mean gender, handedness, and racial composition were not significantly different across groups (**Table 1**). Subacute stroke subjects were significantly older than chronic stroke subjects ($p < 0.01$) and trended older than healthy older subjects ($p = 0.09$). Both stroke groups had significantly more comorbidities than healthy older subjects (both $p < 0.005$), but psychoactive medication intake was similar across the groups. The paretic FDI of subacute and chronic stroke subjects was significantly weaker than right FDI of healthy older subjects (both $p < 0.01$) on the MRC scale but not by MVC. The side and the location of the strokes were not significantly different between subacute and chronic subjects. As expected, significantly more days since the stroke had elapsed in chronic than in subacute stroke subjects ($p < 0.01$).

Self-reported levels of alertness, excitement to participate, and caffeine intake were not significantly different across groups (**Table 2**). Chronic stroke subjects slept significantly more hours than healthy older subjects ($p < 0.05$). Chronic stroke subjects also spent significantly less time exercising in the hours before the testing session than subacute stroke or healthy older subjects (both $p < 0.05$).

TABLE 1 | Clinical characteristics.

Groups	Healthy older	Subacute stroke	Chronic stroke
N	21	20	21
Age (years)	64.7 ± 10.1	72.2 ± 12.7	62.0 ± 9.2
Gender	10M: 11F	11M: 9F	15M: 6F
Handedness	0.9 ± 0.4	0.9 ± 0.5	0.8 ± 0.5
Race (White: Black: Asian)	16W: 3B: 1A	15W: 5B: 1A	16W: 3B: 2A
Number of comorbidities	1.3 ± 1.4	3.1 ± 1.3	2.8 ± 1.3
Psychoactive meds (% taking)	20.0%	20.0%	23.8%
FDI abduction strength (MRC)	5.0 ± 0.0	4.4 ± 0.6	4.0 ± 1.0
Lateral pinch strength (kg)	4.66 ± 2.20	4.09 ± 1.33	4.44 ± 2.00
Lesion type (Subcortical: Mixed)	–	11S: 9M	12S: 9M
Lesioned hemisphere (Left: Right)	–	9L: 11R	15L: 6R
Time since stroke (days)	–	17.4 ± 9.8	2617.9 ± 3166.1

Age, Edinburgh handedness score, number of comorbidities, FDI strength, MVC, and time since stroke are mean ± SD. The proportion of the group taking standing psychoactive medications is listed. FDI MRC score and pinch strength are shown for the right hand of healthy subjects and the paretic hand of stroke subjects.

Missing and Excluded TMS Data

An accidental recording failure occurred for a single LICI and TS_{MEP} set in a chronic stroke subject and for a single $SICI_{80}$, ICF_{80}, and TS_{MEP} set for a healthy older subject (0.6% data missing from each group); all other data collections were complete. Due to an active pre-stimulus EMG baseline, 3.1% of single trials from healthy older, 6.3% from subacute, and 3.7% from chronic stroke subjects were removed. Inappropriate model fits prompted removal of 5.9% of healthy older, 9.4% of subacute stroke, and 12.7% of chronic stroke RC estimations.

TMS Measurements within and Across Groups

We averaged hemispheric data and compared within and across groups to verify similarity to measurements previously documented in the field (**Table 3**). Though RC slope and LICI are transformed for comparison of their means and estimation of their reliability measurement properties, means of their untransformed data are shown to facilitate comparison to extant results.

Within healthy older and subacute stroke groups, outcomes were not significantly different between hemispheres. In chronic

stroke, compared to the non-lesioned hemisphere, the lesioned hemisphere had a significantly lower RC plateau ($p < 0.05$) and TS_{MEP} ($p < 0.01$); it also had a significantly higher rMT ($p < 0.005$) and TS_{MSO} ($p < 0.005$). As outcomes were not significantly different across hemispheres in healthy older subjects, we combined their data to create a "healthy control" hemisphere against which the lesioned and non-lesioned hemispheres of stroke subjects were compared.

Compared to the healthy control hemisphere, $SICI_{80}$ was reduced in the lesioned hemispheres of subacute and chronic stroke (both $p < 0.05$; higher decimal fraction connoting disinhibition) and in the non-lesioned hemisphere of subacute stroke only ($p < 0.05$). Compared to healthy control hemisphere, RC plateau and TS_{MEP} were decreased (both $p < 0.01$), and TS_{MSO} was increased ($p < 0.05$) in the lesioned hemisphere of chronic stroke. Comparing stroke groups, outcomes were not significantly different between the two non-lesioned hemispheres or between the two lesioned hemispheres.

Reliability Measurement Properties

SEMeas and SDC_{indiv} are shown for each group's hemisphere and outcome (**Table 4**). For untransformed measures, SEMeas and SDC_{indiv} values are reported in units particular to the measure: rMT, TS_{MSO}, and RC S_{50} are in %MSO; RC plateau is in mV; and SICI and ICF are in the decimal fraction of TS_{MEP}. As an example of interpretation, take for instance the left hemisphere of the healthy older adult. Measurement error was 1.72%MSO points around an observed rMT, and a stimulator intensity change of at least 4.77%MSO points would be necessary to exceed measurement noise in an individual. Similarly, measurement error was a 0.22 decimal fraction of TS around an observed ICF_{80} value, and would have to change by at least 0.62 to be declared a real change exceeding noise in an individual.

For transformed measures, SEMeas and SDC_{indiv} values are unitless, but are based on data in their original units: mV/%MSO

TABLE 2 | Psychometric data.

Groups	Healthy older	Subacute stroke	Chronic stroke
Alertness	7.7 ± 1.3	7.9 ± 1.3	7.6 ± 1.1
Excitement	8.1 ± 1.1	8.1 ± 1.9	8.1 ± 1.7
Sleep duration (h)	6.4 ± 1.5	7.0 ± 1.5	7.6 ± 1.5
Caffeine intake (cups)	0.7 ± 0.6	0.4 ± 0.4	0.7 ± 0.5
Exercise duration (min)	23.6 ± 15.4	23.0 ± 39.0	12.4 ± 10.8

All self-reported outcomes are mean ± SD. Alertness and excitement are on a scale of 1–10 (10 is maximum). Sleep duration includes daytime naps.

TABLE 3 | TMS measures.

Groups	Healthy older		Subacute stroke		Chronic stroke	
Hemisphere	Left	Right	Lesioned	Non-lesioned	Lesioned	Non-lesioned
rMT (%MSO)	48.67 ± 9.54	50.35 ± 8.49	51.15 ± 12.65	47.22 ± 10.46	$55.89 \pm 10.62^*$	$46.14 \pm 7.39^*$
TS_{MSO} (%MSO)	67.52 ± 12.08	65.42 ± 10.52	67.48 ± 17.76	64.41 ± 17.46	$77.44 \pm 13.34^{*\ddagger}$	$62.90 \pm 11.56^*$
TS_{MEP} (mV)	1.34 ± 0.62	1.31 ± 0.43	0.94 ± 0.67	1.25 ± 0.42	$0.72 \pm 0.67^{*\ddagger}$	$1.29 \pm 0.44^*$
RC slope (mV/%MSO)	0.38 ± 0.37	0.50 ± 0.40	0.43 ± 0.70	0.28 ± 0.17	0.68 ± 1.28	0.46 ± 0.54
RC S_{50} (%MSO)	61.56 ± 13.35	61.46 ± 12.07	61.07 ± 13.44	61.57 ± 14.21	61.41 ± 12.04	60.38 ± 10.53
RC plateau (mV)	2.20 ± 1.30	2.42 ± 1.38	1.83 ± 1.56	2.64 ± 1.35	$1.08 \pm 1.23^{*\ddagger}$	$2.06 \pm 0.98^*$
$SICI_{60}$	0.76 ± 0.31	0.72 ± 0.26	0.82 ± 0.26	0.74 ± 0.33	0.82 ± 0.22	0.74 ± 0.28
$SICI_{80}$	0.41 ± 0.22	0.39 ± 0.18	$0.64 \pm 0.32^{\ddagger}$	$0.60 \pm 0.30^{\dagger}$	$0.61 \pm 0.26^{\ddagger}$	0.54 ± 0.36
ICF_{60}	1.20 ± 0.18	1.22 ± 0.24	1.32 ± 0.40	1.21 ± 0.24	1.31 ± 0.42	1.12 ± 0.25
ICF_{80}	1.39 ± 0.24	1.34 ± 0.31	1.55 ± 0.77	1.30 ± 0.39	1.60 ± 0.70	1.39 ± 0.36
LICI	0.29 ± 0.38	0.23 ± 0.32	0.12 ± 0.18	0.25 ± 0.36	0.36 ± 0.39	0.22 ± 0.29

Measurements are mean ± SD from the 4 sessions. All paired-pulse measurements are decimal percentage of the unconditioned stimulus.
*Significant within-group interhemispheric difference.
†Significant difference between the non-lesioned hemisphere and healthy control hemisphere.
‡Significant difference between lesioned hemisphere and healthy control hemisphere.

TABLE 4 | Standard error of the measurement (SEMeas) and smallest detectable change for individuals (SDC$_{indiv}$) with lower and upper 95% confidence intervals in parentheses.

Groups	Healthy older				Subacute stroke				Chronic stroke			
Hemisphere	Left		Right		Lesioned		Non-lesioned		Lesioned		Non-lesioned	
	SEMeas	SDC$_{indiv}$	SEMeas	SDC$_{indiv}$	SEMeas	SDC$_{indiv}$	SEMeas	SDC$_{indiv}$	SEMeas	SDC$_{indiv}$	SEMeas	SDC$_{indiv}$
rMT	1.72 (0.58, 2.68)	4.77 (1.61, 7.43)	1.29 (0.84, 1.75)	3.58 (2.32, 4.85)	2.41 (0.99, 3.64)	6.67 (2.75, 10.08)	1.04 (0.70, 1.32)	2.88 (1.94, 3.67)	2.07 (1.12, 2.68)	5.75 (3.11, 7.42)	1.17 (0.85, 1.36)	3.24 (2.36, 3.78)
TS$_{MSO}$	1.60 (0.85, 2.17)	4.42 (2.36, 6.00)	1.95 (1.22, 2.55)	5.41 (3.38, 7.08)	1.92 (0.95, 2.49)	5.33 (2.63, 6.91)	2.11 (1.20, 3.06)	5.85 (3.33, 8.47)	1.57 (1.26, 2.79)	4.36 (3.50, 7.74)	2.65 (1.57, 3.34)	7.34 (4.35, 9.25)
Ln(RC slope)	0.66 (0.43, 0.91)	1.83 (1.19, 2.53)	0.76 (0.59, 1.03)	2.10 (1.64, 2.87)	0.48 (0.29, 0.61)	1.33 (0.80, 1.68)	0.39 (0.27, 0.52)	1.08 (0.74, 1.45)	0.84 (0.51, 1.26)	2.32 (1.41, 3.50)	0.66 (0.40, 0.98)	1.83 (1.12, 2.72)
RC S$_{50}$	4.12 (2.54, 5.30)	11.42 (7.03, 14.69)	3.52 (2.03, 4.51)	9.76 (5.63, 12.49)	5.40 (3.22, 7.19)	14.98 (8.93, 19.92)	4.18 (2.84, 5.20)	11.58 (7.88, 14.41)	4.31 (2.58, 6.41)	11.95 (7.16, 17.76)	5.18 (2.47, 7.73)	14.37 (6.85, 21.41)
RC plateau	0.42 (0.29, 0.51)	1.16 (0.81, 1.40)	0.61 (0.30, 0.87)	1.68 (0.84, 2.42)	0.38 (0.14, 0.57)	1.06 (0.38, 1.58)	0.60 (0.37, 0.70)	1.66 (1.03, 1.95)	0.19 (0.12, 0.27)	0.54 (0.34, 0.76)	0.59 (0.35, 0.81)	1.63 (0.97, 2.24)
SICI$_{60}$	0.18 (0.13, 0.22)	0.51 (0.35, 0.62)	0.22 (0.11, 0.31)	0.60 (0.28, 0.87)	0.19 (0.11, 0.28)	0.53 (0.30, 0.77)	0.18 (0.11, 0.24)	0.50 (0.31, 0.67)	0.21 (0.11, 0.30)	0.58 (0.30, 0.82)	0.19 (0.12, 0.23)	0.52 (0.34, 0.63)
SICI$_{80}$	0.11 (0.08, 0.13)	0.31 (0.22, 0.37)	0.16 (0.11, 0.19)	0.44 (0.30, 0.53)	0.17 (0.08, 0.22)	0.46 (0.26, 0.60)	0.17 (0.11, 0.23)	0.48 (0.31, 0.63)	0.21 (0.13, 0.27)	0.59 (0.36, 0.75)	0.13 (0.08, 0.15)	0.35 (0.24, 0.42)
ICF$_{60}$	0.21 (0.16, 0.28)	0.59 (0.45, 0.77)	0.28 (0.14, 0.44)	0.79 (0.39, 1.21)	0.33 (0.14, 0.50)	0.90 (0.40, 1.38)	0.22 (0.15, 0.27)	0.61 (0.40, 0.74)	0.51 (0.19, 0.86)	1.42 (0.53, 2.38)	0.27 (0.20, 0.37)	0.74 (0.54, 1.02)
ICF$_{80}$	0.22 (0.18, 0.29)	0.62 (0.50, 0.80)	0.36 (0.28, 0.50)	1.00 (0.77, 1.40)	0.59 (0.16, 1.06)	1.63 (0.44, 2.93)	0.22 (0.14, 0.27)	0.60 (0.39, 0.74)	0.78 (0.48, 1.29)	2.17 (1.33, 3.58)	0.34 (0.22, 0.43)	0.95 (0.62, 1.20)
Ln(LICI)	0.55 (0.37, 0.71)	1.54 (1.01, 1.98)	0.54 (0.32, 0.69)	1.50 (0.88, 1.91)	0.42 (0.26, 0.50)	1.16 (0.71, 1.38)	0.57 (0.39, 0.78)	1.59 (1.08, 2.17)	0.60 (0.29, 0.87)	1.65 (0.79, 2.42)	0.53 (0.35, 0.63)	1.47 (0. 97, 1.74)

SEMeas is the measurement error of the measure, whereas SDC$_{indiv}$ delimits the smallest delta that is considered a real measurement change for an individual. SEMeas and SDC$_{indiv}$are in units particular to the measure for non-transformed data: rMT, TS$_{MSO}$, and RC S$_{50}$ are %MSO; RC plateau is mV; SICI and ICF are in decimal fraction of TS. The SEMeas and SDC for ln(RC slope) and ln(LICI) are dimensionless, but are based on original units of mV/%MSO (slope) and decimal fraction of TS (LICI). To calculate SDC$_{group}$ for a group of size n, divide SDC$_{indiv}$ by $\sqrt{n}$. SDC$_{group}$ is the smallest delta that is considered a real measurement change for a group of size n.

for RC slope and decimal fraction of TS_{MEP} for LICI. Staying with the left hemisphere of the healthy older adult, measurement error is 0.66 around the observed ln(RC slope) value, and a difference of at least 1.83 between two observed ln(RC slope) values would be needed to exceed measurement noise in an individual.

Importantly, SDC_{indiv} is provided to enable future investigators to generate SDC_{group} for their samples of size n. Given that measurement error changes with context, it is imperative that investigators calculate the SDC_{group} pertaining to each subject type, hemisphere, and sample size. As would be expected, SDC_{group} shrinks dramatically as group size increases. A graphical example is given for $SICI_{80}$ in the lesioned hemisphere of subacute stroke subjects (**Figure 1**). For $SICI_{80}$, changes greater than $SDC_{indiv} = 0.46$ would be required to exceed measurement error for an individual, but as n increases beyond 19, $SDC_{group} \leq 0.10$.

Across all measures and all groups, SDC_{indiv} were sizable. SEMeas% provides an impression of the relative noisiness of a measure; because transformed data are no longer on a ratio scale, LICI and RC slope are excluded (**Table 5**). Others have used SEMeas% < 10% as a cutoff for high measurement stability, though this threshold is arbitrary (Flansbjer et al., 2005). From **Table 5**, the measures with the lowest relative measurement error were rMT (2.2–4.7%), TS_{MSO} (2.0–4.7%), and RC S_{50} (5.7–8.9%), with RC plateau (18.1–28.4%), SICI (24.3–40.8%), and ICF (16.1–48.9%) having moderate measurement error. In general, measures were noisier in the right compared to the left hemisphere of healthy older subjects, and were marginally noisier in the lesioned compared to the non-lesioned hemispheres of subacute and chronic stroke subjects. SICI and ICF were slightly noisier when obtained with a CS of 80% rMT than with 60% rMT.

$Reliability_{MP}$ is shown for each group's hemisphere and outcome (**Table 6**). The ICC quantifies $reliability_{MP}$, the measure's ability to distinguish between subjects in a sample. As the ICCs of RC slope and LICI were based on logarithmically transformed data, their ICCs comment on how well the transformed values of the measure can discriminate between subjects in a sample. ICC > 0.70 is generally considered good $reliability_{MP}$ (Portney and Watkins, 2009). By this convention, rMT, TS_{MSO}, RC S_{50}, RC plateau, and ln(LICI) had generally good $reliability_{MP}$, whereas RC slope and all SICI and ICF measures had generally poorer $reliability_{MP}$. SICI and ICF had higher ICCs when obtained with a CS of 80% rMT compared with 60% rMT. For the most part, measures taken in healthy older subjects had marginally higher $reliability_{MP}$ in the left than the right hemisphere. In subacute stroke subjects, measures from the lesioned hemisphere had slightly higher $reliability_{MP}$ than the non-lesioned hemisphere; the opposite occurred in chronic stroke subjects. Generally measures taken in subacute stroke subjects had the highest $reliability_{MP}$.

Diagnostics

To insure that variability did not change over the course of the study (e.g., because drift in operator technique) we inspected the variances of each measure from the first and last 5 subjects of the three groups. We did not see a systematic pattern of variance change over time. Given the fixed testing order for the measures,
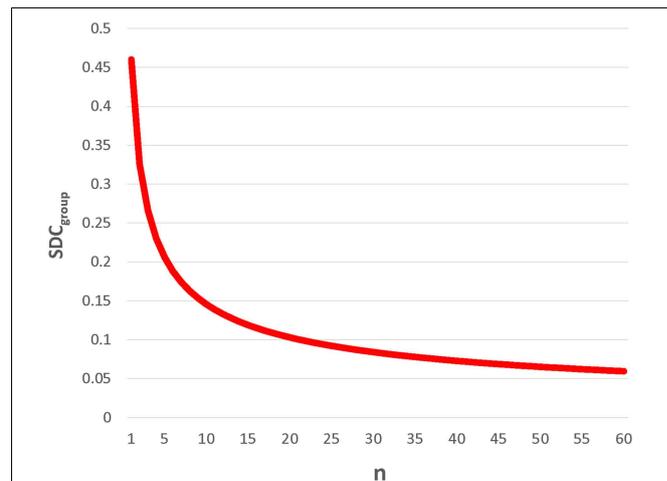

FIGURE 1 | SDC_{group} decreases dramatically with increasing n. This example, using $SICI_{80}$ in lesioned hemisphere of subacute stroke, demonstrates for that for an individual, the SDC is quite large, requiring a change greater than 0.45 in $SICI_{80}$ in the lesioned hemisphere to be deemed real. By calculating SDC_{group} (= $SDC_{indiv}/\sqrt{n}$), it is apparent that even with modest sample sizes of 10–20, changes exceeding measurement error could be conceivably detected, and the measure would be considered reliable.

we also inspected order effects on variance. We compared within-subjects variance of TS_{MEP} obtained with same TS_{MSO}, recorded earlier and later in the session (i.e., TS_{MEP} used for $SICI_{60}/ICF_{60}$ vs. LICI). We found no systematic pattern of change in TS_{MEP} variance in either hemisphere in any subject group.

Discussion

In the present study, we characterized the reliability of common TMS measures in healthy older subjects and stroke subjects, using variances estimated from a four-session test-retest paradigm. We distinguished between reliability measurement properties that inform the two main uses of TMS: (1) evaluation of change in subjects, for which measurement error must be considered; and (2) discrimination between subjects within a sample, for which $reliability_{MP}$ must be considered. We additionally used the measurement error to generate the SDC, which assists investigators in deciding, for a given sample size, whether an observed change is sufficiently large to exceed measurement error.

We found that SDCs for individuals were universally large across all TMS measures, precluding reasonable single-subject evaluative use; therefore these TMS measures *should not* be used as biomarkers for assessing individual change. For these same TMS measures, SDCs become sufficiently low in modest sample sizes to justify their evaluative use; therefore these measures *can* be used to detect group-level changes. $Reliability_{MP}$ was highly variable for many measures in our subject groups. However, for samples with similar measurement distributions as ours, rMT, TS_{MSO}, RC S_{50}, RC plateau, and ln(LICI) could be used for discriminatory purposes, for example, for patient staging or diagnosis.

TABLE 5 | Relative measurement error.

Groups	Healthy older		Subacute stroke		Chronic stroke	
Hemisphere	Left	Right	Lesioned	Non-lesioned	Lesioned	Non-lesioned
rMT	3.53	2.56	4.70	2.20	3.71	2.54
TS_{MSO}	2.36	2.98	2.85	3.28	2.03	4.21
RC S_{50}	6.69	5.73	8.85	6.78	7.02	8.58
RC plateau	18.96	25.05	20.85	22.63	18.07	28.44
$SICI_{60}$	24.28	30.18	23.49	24.49	25.68	25.25
$SICI_{80}$	26.81	40.83	25.91	29.12	34.76	23.22
ICF_{60}	17.71	23.17	24.75	18.12	39.22	23.71
ICF_{80}	16.13	26.70	38.06	16.70	48.90	24.54

SEMeas% is measurement error relative to the mean measurement of the outcome. It is used to inspect the relative noisiness of the measure. Measures with SEMeas% <10% may imply better measurement stability. RC slope and LICI are excluded because transformed data are no longer on ratio scale.

TABLE 6 | Intraclass correlation coefficient (ICC) with lower and upper 95% confidence intervals in parentheses.

Groups	Healthy older		Subacute stroke		Chronic stroke	
Hemisphere	Left	Right	Lesioned	Non-lesioned	Lesioned	Non-lesioned
rMT	0.97 (0.90, 0.99)	0.98 (0.96, 0.99)	0.96 (0.91, 0.99)	0.99 (0.98, 0.99)	0.96 (0.93, 0.98)	0.98 (0.96, 0.99)
TS_{MSO}	0.98 (0.96, 0.99)	0.97 (0.94, 0.99)	0.99 (0.97, 0.99)	0.99 (0.96, 0.99)	0.99 (0.96, 0.99)	0.95 (0.91, 0.98)
Ln(RC slope)	0.03 (0, 0.51)	0.07 (0, 0.70)	0.70 (0.35, 0.84)	0.53 (0.26, 0.77)	0.18 (0, 0.86)	0.23 (0, 0.71)
RC S_{50}	0.91 (0.80, 0.95)	0.92 (0.82, 0.96)	0.85 (0.72, 0.94)	0.92 (0.85, 0.96)	0.88 (0.77, 0.97)	0.78 (0.49, 0.92)
RC plateau	0.90 (0.82, 0.94)	0.82 (0.62, 0.94)	0.94 (0.87, 0.98)	0.82 (0.66, 0.90)	0.98 (0.95, 0.99)	0.71 (0.55, 0.84)
$SICI_{60}$	0.71 (0.52, 0.83)	0.51 (0.31, 0.70)	0.60 (0.33, 0.82)	0.74 (0.57, 0.90)	0.33 (0, 0.78)	0.64 (0.28, 0.80)
$SICI_{80}$	0.78 (0.62, 0.86)	0.44 (0.01, 0.63)	0.77 (0.57, 0.88)	0.71 (0.47, 0.86)	0.49 (0.20, 0.72)	0.88 (0.77, 0.92)
ICF_{60}	0.04 (0, 0.59)	0.23 (0, 0.73)	0.48 (0, 0.85)	0.43 (0.03, 0.66)	0.15 (0, 0.81)	0.00 (0, 0.69)
ICF_{80}	0.45 (0.17, 0.75)	0.32 (0, 0.79)	0.56 (0.30, 0.94)	0.73 (0.53, 0.86)	0.16 (0, 0.92)	0.38 (0, 0.70)
Ln(LICI)	0.88 (0.76, 0.94)	0.88 (0.76, 0.94)	0.89 (0.81, 0.95)	0.86 (0.73, 0.94)	0.87 (0.75, 0.97)	0.89 (0.79, 0.93)

ICC quantifies reliability_MP, which is how well a measure can distinguish between subjects in a sample. ICC > 0.70 is considered good reliability_MP.

TMS Measurement Outcomes in Healthy Aging and Stroke Subjects

Our TMS technique produced measurements that largely replicate the outcomes of other neurophysiology studies conducted on the FDI of healthy older subjects (Bütefisch et al., 2003, 2008; Delvaux et al., 2003; Werhahn et al., 2003; Fridman et al., 2004; Swayne et al., 2008; Takechi et al., 2014), subacute stroke subjects (Bütefisch et al., 2003, 2008; Delvaux et al., 2003; Liepert et al., 2005; Swayne et al., 2008; Prashantha et al., 2013), and chronic stroke subjects (Delvaux et al., 2003; Werhahn et al., 2003; Fridman et al., 2004; Liepert, 2006; Swayne et al., 2008; Takechi et al., 2014).

One exception was a relatively lower rMT observed in the lesioned hemisphere of our subacute stroke subjects compared to what some have recorded. This difference could arise from a common practice of assigning an MSO of 100% when there is an absent motor response (Bütefisch et al., 2008; Swayne et al., 2008; Takechi et al., 2014). This approach inflates group averages (Heald et al., 1993), and removal of non-responders from the group aggregate reveals a reduction in the group rMT (Bütefisch et al., 2008). Another source of dissimilarity may arise from equipment differences: absolute magnetic field strength associated with %MSO varies with stimulator brand, coil shape, and Bistim vs. single-stimulator mode for the Magstim device. A lower rMT in the lesioned hemisphere may also indicate better preserved corticospinal tract integrity in our patient sample, as may be inferred from their reasonably high level of function.

Comparing the magnitude of SICIs elicited by different CS intensities, we did not appreciate an inhibitory nadir with a CS at 60% rMT ($SICI_{60}$), as has been observed in the non-lesioned hemisphere (Bütefisch et al., 2003, 2008) and lesioned hemisphere (Bütefisch et al., 2008) of subacute stroke subjects. Rather, inhibition was stronger with a CS at 80% rMT ($SICI_{80}$) in all subject groups, similar to that seen in healthy younger subjects (Kujirai et al., 1993). Differences may result from our use of higher TS MEP amplitudes, as per the original SICI paradigm (Kujirai et al., 1993); perhaps there is a unique interaction between CS intensity and TS amplitude in stroke subjects, given each affects apparent SICI (Kujirai et al., 1993; Sanger et al., 2001). In keeping with other observations, $SICI_{80}$ was reduced in both hemispheres of acute stroke subjects (Liepert et al., 2000b; Bütefisch et al., 2003, 2008; Takechi et al., 2014) and in the lesioned hemisphere of chronic stroke subjects (Swayne et al., 2008) relative to healthy controls.

In summary, our TMS measurements are in line with the majority of previous studies in healthy older and stroke subjects, including observed similarities and differences between the healthy and stroke groups. As with any study, it is possible that despite correcting for multiple comparisons, our comparative findings may have arisen by chance alone. However, our replication of past observations supports the generalization of our reliability results to conventional TMS paradigms.

Reliability Measurement Properties

The following discussion of reliability results is restricted to previous studies focusing on FDI neurophysiology in older healthy subjects and stroke subjects. TMS measures do not behave the same with age (Rossini et al., 1992; Peinemann et al., 2001; McGinley et al., 2010) or in different muscle representations (Brasil-Neto et al., 1992; Malcolm et al., 2006; Menon et al., 2014), and therefore we do not assume that their variances or derived reliability measurement properties are similar.

Measurement Error and Smallest Detectable Change

Two previous reliability studies estimated the SEMeas of TMS measures in FDI of chronic stroke subjects (Koski et al., 2007; Liu and Au-Yeung, 2014), although differences in paradigm, analysis, and measures preclude comparison to the first study (Koski et al., 2007). Our findings were comparable to work by Liu and colleagues (Liu and Au-Yeung, 2014), who found rMT SEMeas in chronic stroke subjects was 1.9 MSO points in the lesioned hemisphere and 2.02 in the non-lesioned hemisphere for rMT, similar to our 2.07 and 1.17, respectively. To our knowledge, SEMeas has not been previously reported for other TMS measures in the FDI of chronic stroke, or for any measure in the FDI of subacute stroke or healthy older subjects. SDC has not been estimated for any measures in the FDI of healthy older adults or stroke subjects. A recent study used the Limits of Agreement (LOA), similar in concept to the SDC, to estimate test-retest reliability of total motor conduction time in subacute stroke subjects (Hoonhorst et al., 2014).

Just how noisy are the TMS measures? Are the SEMeas and SDC small enough to encourage the use of the measure within individuals, i.e., as a biomarker in clinical practice? There are no universally accepted norms for an acceptably low SEMeas or SDC. One approach is to use a relative measurement error (SEMeas%) <10% as a cutoff for high measurement stability (Flansbjer et al., 2005). SEMeas% in our study ranged from very low (e.g., 2.2% for subacute non-lesioned rMT) to moderately high (e.g., 48.9% for chronic lesioned ICF_{80}). The measures with generally the lowest relative measurement error were rMT, TS_{MSO}, and RC S_{50}, with the highest being SICI and ICF. In general, the relative measurement error was marginally lower in the left healthy older hemisphere and in the non-lesioned stroke hemispheres, suggesting a possible role for hemispheric dominance and non-pathology contributing to measurement stability.

In reality, however, whether a measure is sufficiently stable for practical individual use is determined by the amount of change that one could reasonably expect in the measurements

with a true clinical state change. Take by analogy the hematocrit, the percent of red blood cells in a sample. The hematocrit SDC is approximately 3% points at most institutions; that is, the hematocrit can change by 3 points and still be considered measurement noise. Clinicians tolerate this SDC because a real change in clinical state generally produces a hematocrit change exceeding 3 points. If instead the hematocrit SDC was 20 points, measurement error may produce a retest value that would be interpreted as a clinical change in a stable patient; in this case, hematocrit would be unreliable for measuring change in the individual. Thus, an overall sense of expected real change in the individual or group is helpful for determining whether the measure has a sufficiently low SDC to be considered a viable biomarker.

Across groups for most TMS measures, SDC_{indiv} exceeds what is likely to be observed in an individual changing due to a standard intervention or a clinical state change. Take the left hemisphere of a healthy subject for example: after intervention, it would be unusual to observe more than a 4.4%MSO adjustment to reproduce TS_{MEP}, a RC plateau change greater than 1.1 mV, a $SICI_{80}$ change greater than 0.31, a ICF_{80} change greater than 0.62, or an ln(LICI) change greater than 1.54 (equivalent to a change of $e^{1.54}$, or 466%, in LICI); the other hemisphere and groups generally require even larger changes (**Table 4**). We therefore advise against using TMS measures to track changes within the individual; SDC_{indiv} are too large to be used as individual biomarkers for the changes that can be reasonably expected with our current interventions. (The exception, of course, is if an intervention is particularly potent and induces marked change.)

With increasing n, the SDC_{group} becomes smaller, i.e., smaller average measurement changes are required to exceed measurement noise. One may use the SDC_{group} to ascertain whether a change exceeding measurement error has occurred in collected data. Inferential statistics (i.e., t-tests to test for the significance of the change) are complimentary, giving concordant information if the measurement variances of our group and the compared sample are similar. If an observed change exceeds SDC_{group} and is statistically significant, it is likely real measurement change; if the observed change does not exceed SDC_{group} and is non-significant, it is likely measurement noise. Given that our variances were derived from a large number of observations, to match the precision of our estimates, the compared sample would also need to be large.

Variances are prone to be dissimilar with small sample sizes. In scenarios where the change exceeds the SDC_{group} but it is not significant, or vice-versa, the discordance is explained by differences in variance estimates of the sample compared to ours. Thus, a non-significant change exceeding the SDC_{group} implies a higher variability in the compared sample, whereas a significant change less than the SDC_{group} implies a lower variability in the compared sample. Higher or lower variability may have arisen through differences in instrument technique or chance in the sample of subjects.

A change in excess of the SDC_{group}, even if statistically significant, does not guarantee that it is "true," i.e., that it reflects real change in the population (Button et al., 2013). It is critically important that the sample be large enough to ensure

robust statistical power—the ability to detect a true population-representative effect. A formal power analysis accounts for both measurement error and probability of detecting a significant change. If a study is known to be adequately powered, then when a statistical test is judged significant, there is a good likelihood that the detected change is in fact true. When underpowered studies find significant differences, these are likely detecting effects that are so inflated that they do not represent a true effect (Button et al., 2013). The highly precise constituent variances derived from our SEMeas and ICCs can and should be used for the appropriate power analyses.

Finally, the potential clinical applications of SDC$_{indiv}$ and SDC$_{group}$ are important to distinguish. For a given TMS measure, if an intervention tested in a group of subjects induces a measurement change in excess of the corresponding SDC$_{group}$, the intervention is shown to be effective (or at the minimum, to exceed measurement noise) at the group level. However, this does not mean that the same TMS measure can then be used to detect the intervention's effects in an individual. The SDC$_{indiv}$ is substantially larger than SDC$_{group}$, and only by rare chance (<5%) or by strong interventional potency might the individual show a TMS measurement change exceeding the SDC$_{indiv}$. What this implies clinically is that a TMS measure could be used to detect a change induced by an intervention in a group, but the same measure may not be used to assess individual efficacy. That is, the intervention will need to be given agnostic to individual TMS outcome, because the TMS measure cannot reliably detect change in the single subject.

Reliability$_{MP}$

We estimated reliability$_{MP}$ for our TMS measures, to evaluate the potential for using TMS measures to differentiate between subjects for staging or diagnosis. To our knowledge, TMS has not been applied this way in stroke, aside from the combination of a dichotomous presence/absence of an MEP with other clinical and radiographic features to help prognosticate recovery (Stinear et al., 2014). Although several studies have estimated ICC for TMS measures in young adults (Mortifee et al., 1994; Carroll et al., 2001; Kamen, 2004; Malcolm et al., 2006; Christie et al., 2007; Koski et al., 2007; Cacchio et al., 2009, 2011; Wheaton et al., 2009; Hoonhorst et al., 2014; Liu and Au-Yeung, 2014), few have characterized reliability$_{MP}$ in the FDI of older subjects and subjects with stroke. Again we focus on the estimation of reliability$_{MP}$ for measures taken in the FDI of elderly subjects and stroke subjects, because ICCs between intrinsic hand muscles or between hand and forearm muscles are not the same (Malcolm et al., 2006) and are not expected to be similar for young and old adults.

Guidelines suggest that ICCs > 0.70 indicate acceptable reliability$_{MP}$ to distinguish between subjects (Nunnally and Bernstein, 1994). Of those that did investigate reliability$_{MP}$ in subacute and chronic stroke subjects, generally ICCs > 0.70 were noted for motor thresholds, linearly-derived recruitment curve parameters, silent periods, and total motor conduction time (Koski et al., 2007; Hoonhorst et al., 2014; Liu and Au-Yeung, 2014). Our ICCs for rMT in chronic stroke subjects were comparable to those found by Liu and colleagues; importantly

for this comparison, their sample's rMT variances were similar to ours. ICCs were 0.97 in the lesioned hemisphere and 0.95 in the non-lesioned hemisphere, very similar to our 0.96 and 0.98, respectively (Liu and Au-Yeung, 2014). To our knowledge, there have been no studies evaluating reliability$_{MP}$ of any TMS measure in FDI of older healthy subjects, or of paired-pulse measures in subacute or chronic stroke subjects.

In our study, ICCs ranged widely across subject groups and outcomes, with rMT and TS$_{MSO}$ having very high ICCs, and ln(RC slope) and ICF generally low ICCs. Measurements taken in subacute stroke subjects tended to have the highest reliability$_{MP}$. These findings are not counterintuitive. The ICC more predominantly reflects the spread of measurements between subjects and less the measurement error within subjects. For example, the relative measurement errors for SICI$_{60}$ in subacute stroke were 23.5% and 24.5% in the lesioned and non-lesioned hemisphere, respectively. Despite these essentially equivalent SEMeas%, ICCs were 0.60 and 0.74, respectively; the higher ICC was simply due to a larger between-subject variance in the latter. It is for this reason that subacute stroke subjects generally have higher ICCs—not because measurement error is markedly lower in subacute stroke, but because the measurements are more widely dispersed across individuals.

Our findings underscore the direct relationship between the magnitude of the between-subject variance and the magnitude of the ICC (Streiner and Norman, 2008), and how a moderately large measurement error can be obscured by an even larger spread of subjects. It is therefore vital for investigators to understand exactly what the ICC means, how it should be applied, and the constraints on its generalizability.

These three issues are rarely discussed in TMS reliability$_{MP}$ studies for various muscles (Mortifee et al., 1994; Carroll et al., 2001; Kamen, 2004; Malcolm et al., 2006; Christie et al., 2007; Koski et al., 2007; Cacchio et al., 2009, 2011; Wheaton et al., 2009; Hoonhorst et al., 2014; Liu and Au-Yeung, 2014). First, nearly all erroneously interpreted high reliability$_{MP}$ to mean low measurement error. Measures with a high ICC were incorrectly endorsed for evaluative use (Mortifee et al., 1994; Carroll et al., 2001; Kamen, 2004; Malcolm et al., 2006; Christie et al., 2007; Koski et al., 2007; Wheaton et al., 2009; Hoonhorst et al., 2014). Second, none explained the practical application of ICC, which is how well the measure can tell subjects apart, not how well it can detect individual or group change over time (Mortifee et al., 1994; Carroll et al., 2001; Kamen, 2004; Christie et al., 2007; Wheaton et al., 2009; Cacchio et al., 2011; Liu and Au-Yeung, 2014). Only one study pointed out the influence of between-subject variability on ICC estimations (Koski et al., 2007). The frequent observation of a lower reliability$_{MP}$ in a non-lesioned hemisphere or in healthy subjects speaks more to measurement homogeneity, not poor paradigmatic quality—an issue rarely raised.

Finally, no prior TMS reliability$_{MP}$ study has advised that ICCs should be generalized only to future samples with similar measurement spread. Reliability$_{MP}$ is not an intrinsic characteristic of a measurement instrument, but is rather born from the instrument and the sample (Streiner and Norman, 2008). As ICCs tend to be a function of *who* is being measured, special attention needs to be paid to the sample's measurement

distribution. A high ICC estimated from a sample with widely dispersed measurements will not be valid for a sample with a narrower distribution. This is not statistical dogmatism, but is simply a reflection of reality: "it *is* more difficult to tell people apart if they are relatively similar (i.e., homogeneous) than if they are very different" (Streiner and Norman, 2008). ICCs must not be assumed to be suitable for new samples without first ensuring that measurement distributions are similar. Our samples' standard deviations (**Table 3**) are expressly provided for this purpose.

Keeping these points in mind, we found generally high ICCs for rMT, TS_{MSO}, RC S_{50}, RC plateau, and ln(LICI) in our samples of healthy older adults and stroke subjects. In future samples with similar measurement distributions, these measures could reliably be used to distinguish subjects from one another, i.e., for staging or diagnostic purposes.

Quality of the Study

Studies of test-retest reliability require sufficient sample sizes to estimate variance of change and to make inferences about the relevant population. Sample sizes of 15–50 subjects have been suggested, equating to 30–100 single observations (Fleiss, 1999; Hopkins, 2000; Terwee et al., 2007). In lieu of a larger sample size, we measured each subject in each hemisphere 4 times, each of which was the average of 10 observations, for a total of 80–84 averaged observations per measure per hemisphere. This approach not only gave us a more precise estimate of the overall variance but also allowed us to estimate the various components of the overall variance. We also provide 95% confidence intervals to provide transparency about the level of uncertainty around our estimations.

Studying test-retest reliability assumes that the time between tests is generally long enough to avoid subject learning or carry-over, but short enough that there has not been a clinical change (Terwee et al., 2007). TMS neurophysiology probes are generally believed to be uninfluenced by prior non-modulatory sessions. Neuroplasticity leading to true changes in neurophysiology, particularly for subacute stroke subjects, was of greater concern. We thus chose intervals between tests that were sufficiently short to minimize plasticity-associated neurophysiological change.

Our paradigm called for a fixed testing order, but we do not think this amplified the variability of some measures over others. We noted that LICI had universally large measurement errors and was always the last TMS measure tested in each hemisphere (**Table 5**). Because direct comparison of LICI and other measures would not disambiguate between timing- and measure-related differences in variance, we inspected the variances of TS_{MEP} obtained with identical TS_{MSO} at separate times over the session. We found no systematic increase or decrease in TS_{MEP} variance in any subject group, suggesting that LICI is inherently noisy.

Generalizability of the Study

It is paramount that any reliability study explicitly delineate the extent to which its results can be generalized. Can our results be extrapolated to other laboratories and operators of TMS? Yes, but with qualification. Reliability is not a unique feature of the measurement instrument, but depends on the sources of variance and the study population (Beckerman et al., 2001). Several key features should thus be considered: the sample, the paradigm, the TMS setup, and the operator.

Measures were obtained solely in the FDI muscles of older healthy subjects and subacute and chronic stroke subjects—and our reliability measurement properties can be used only for the same. We do not assume identical magnitudes in reliability measurement properties in other muscles (Malcolm et al., 2006), in other populations, or in other TMS measures. This work remains to be undertaken.

We purposefully included only stroke subjects whose MEPs could be obtained in both hemispheres, to mirror the partially recovered patient commonly recruited for upper extremity interventional studies. Because of this, most of our subjects had reasonable paretic hand function. Therefore, our samples and their reliability measurement properties are not representative of the universal stroke population. Furthermore, it cannot be assumed that the non-lesioned hemispheres of well-recovered and poorly recovered subjects behave similarly; although both produce MEPs, their neurophysiology and reliability may differ (Manganotti et al., 2008).

Our groups were balanced for race, but all samples were predominantly white. Differences in TMS measurements have been noted for different races (Yi et al., 2014), but differential reliability has not been investigated. The predominance of white subjects in our samples should be kept in mind when extrapolating our results globally. Our samples were also balanced for gender. As our females were postmenopausal and did not take hormone replacement therapy, we do not expect gender-related differences in measurement stability. However, cortical excitability varies with fluctuating ovarian hormone levels in premenopausal subjects (Smith et al., 1999), and measurement error should be separately calculated for test-retest intervals spanning the follicular and luteal phases of the menstrual cycle.

Can the SDC from a short-term assessment (collected over days) be used for long-term test-retest data (collected over weeks or months)? Assuming that subjects are stable on both time scales, the short-term measurement error can be applied to longer-term data (de Vet et al., 2011).

We chose techniques and measures conventionally used in TMS studies. Our TMS setup uses stimulators and recording devices that are commonly used and commercially available. Assuming technique and setups are comparable, variability related to the technology should be similar across laboratories. Ours is the first reliability study in TMS to use neuronavigation to ensure spatial stimulation stability within and between sessions (Schönfeldt-Lecuona et al., 2005), and we strongly recommend the same in studies investigating neurophysiological change. Although marked scalps or swimming caps can direct TMS coil location and yaw, these approaches do not guide coil pitch and roll. It is expected that free-handed positioning of the coil would significantly increase measurement error, as it allows spatial drift (Julkunen et al., 2009) and diminishes the consistency and strength of voltage delivered to a target (Cincotta et al., 2010).

Similarly, electrode positions were outlined directly on the skin to reproduce exact placement over sessions. Given that hand intrinsic muscles have a high density of motor units and large corticospinal representations relative to other muscles

(Phillips and Porter, 1977; Brasil-Neto et al., 1992), small variations in electrode placement could lead to the probing of different corticomotoneuronal pathways (Brasil-Neto et al., 1992; Malcolm et al., 2006). To minimize the introduction of additional measurement error by faulty electrode placement, we recommend recording its precise location by marking directly on the skin for short testing intervals, or measuring electrode position relative to bony landmarks for longer intervals. Documenting position with digital photographs or tracings may also be helpful (Butler et al., 2005; Malcolm et al., 2006).

One TMS operator (HS) was used to limit rater-related sources of variability. At the advent of the study, the operator had 4 years of neurophysiology experience in healthy and stroke populations. Over the duration of data collection, variability did not systematically diminish, implying that over some criterion amount of competency, experience may not further influence an investigator's contribution to measurement noise. We assume that our reliability measurement properties would be less applicable to novice TMS investigators. For generalization to all TMS operators, future reliability studies would need to include multiple operators with varying levels of experience.

Validity and Responsiveness

Once reliability has been defined for TMS measures in specific populations, can we freely use the TMS measures to tell us about neurophysiology in these groups? The answer hinges on their validity, how well they actually measure the construct they are purporting to measure (Mokkink et al., 2010), and responsiveness, how well they detect true changes in the construct (Mokkink et al., 2010; de Vet et al., 2011). Like reliability, validity and responsiveness are fundamental quality domains of measurement instruments, and high validity and responsiveness are required to justify their use. Reliability is necessary but not sufficient to determine validity and responsiveness (Portney and Watkins, 2009). Though not the focus of this study, meeting the requirements necessary to deem a TMS measure valid and responsive is a formidable but critical next step for the field.

Interpretability of Measurements in the Clinical Context

Once the reliability has been established, and a change in an appropriately powered group is observed to exceed the SDC, one must decide whether it has any real clinical meaning. This is the measure's interpretability: "the degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to an instrument's quantitative scores or change in scores" (Mokkink et al., 2010).

The minimal clinically important difference (MCID) is the smallest change in outcome that has clinical value to the stakeholder (e.g., the patient, clinician, caretaker, society) (Wyrwich and Wolinsky, 2000; Eisen et al., 2007). For a measure to be useful for evaluation, its SDC must be smaller than the MCID (Hébert et al., 1997; Beckerman et al., 2001; de Vet et al., 2006). As a field, we have not established a MCID for neurophysiologic outcomes. Establishing an MCID would require the linking of TMS measures to clinical outcomes of interest, which is different from linking them to their mechanistic underpinnings (i.e., validation).

Additional discussion about TMS validity, responsiveness, and interpretability can be found in Supplementary Materials.

Conclusion

In the present study, we assessed the reliability of common TMS measures obtained from the FDI, with a focus on their potential evaluative and diagnostic applications in healthy older adults and subacute and chronic stroke patients. To determine whether TMS measures could be used to reliably evaluate change, we estimated their SEMeas and SDC. In all subject groups and for all measures, we found that SDCs at the single-subject level were prohibitively high. Thus, these TMS measures cannot be reliably used as a biomarker to assess individual change. However, SDCs become reasonably low with modest sample sizes, justifying use of the measures to detect group-level change. We provide measures' SDC_{indiv} and instructions for calculating SDC_{group}, expressly so that future investigators can estimate the SDC needed for pre-post testing in their particular sample. To determine whether TMS measures could be used to reliably discriminate between patients, i.e., to diagnose or stage them, we estimated the measures' reliability$_{MP}$. Although most measures' ICCs were variable across groups, they were universally high for a subset of measures [rMT, TS_{MSO}, RC S_{50}, RC plateau, and ln(LICI)]; in samples with similar heterogeneity as ours, these TMS measures can reliably discriminate between patients. TMS measures should thus be used based on their reliability in particular contexts. More work establishing their validity, responsiveness, and clinical relevance is still needed. TMS, despite its longstanding use, remains to be fully vetted to have a place at the health outcomes table.

Author Contributions

Author contributions were as follows: study conception and design (HS, JK, DE); data acquisition (HS); data analysis (HS, RO, IM, XL, YC, AR); data interpretation (HS, RO, XL, YC); and drafting/revising the manuscript (HS, RO, DE, JK).

Acknowledgments

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fncel.2015.00335

References

Balslev, D., Braet, W., McAllister, C., and 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. doi: 10.1016/j.jneumeth.2007.01.021

Beckerman, H., Roebroeck, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., and Verbeek, A. L. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Qual. Life Res.* 10, 571–578. doi: 10.1023/A:1013138911638

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310. doi: 10.1016/S0140-6736(86)90837-8

Boroojerdi, B., Kopylev, L., Battaglia, F., Facchini, S., Ziemann, U., Muellbacher, W., et al. (2000). Reproducibility of intracortical inhibition and facilitation using the paired-pulse paradigm. *Muscle Nerve* 23, 1594–1597. doi: 10.1002/1097-4598(200010)23:10<1594::AID-MUS19>3.0.CO;2-3

Brasil-Neto, J. P., McShane, L. M., Fuhr, P., Hallett, M., and Cohen, L. G. (1992). Topographic mapping of the human motor cortex with magnetic stimulation: factors affecting accuracy and reproducibility. *Electroencephalogr. Clin. Neurophysiol.* 85, 9–16. doi: 10.1016/0168-5597(92)90095-S

Bütefisch, C. M., Netz, J., Wessling, M., Seitz, R. J., and Hömberg, V. (2003). Remote changes in cortical excitability after stroke. *Brain* 126(Pt 2), 470–481. doi: 10.1093/brain/awg044

Butefisch, C. M., Wessling, M., Netz, J., Seitz, R. J., and Hömberg, V. (2008). Relationship between interhemispheric inhibition and motor cortex excitability in subacute stroke patients. *Neurorehabil. Neural Repair* 22, 4–21. doi: 10.1177/1545968307301769

Butler, A. J., Kahn, S., Wolf, S. L., and Weiss, P. (2005). Finger extensor variability in TMS parameters among chronic stroke patients. *J. Neuroeng. Rehabil.* 2:10. doi: 10.1186/1743-0003-2-10

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Cacchio, A., Cimini, N., Alosi, P., Santilli, V., and Marrelli, A. (2009). Reliability of transcranial magnetic stimulation-related measurements of tibialis anterior muscle in healthy subjects. *Clin. Neurophysiol.* 120, 414–419. doi: 10.1016/j.clinph.2008.11.019

Cacchio, A., Paoloni, M., Cimini, N., Mangone, M., Liris, G., Aloisi, P., et al. (2011). Reliability of TMS-related measures of tibialis anterior muscle in patients with chronic stroke and healthy subjects. *J. Neurol. Sci.* 303, 90–94. doi: 10.1016/j.jns.2011.01.004

Carroll, T. J., Riek, S., and Carson, R. G. (2001). Reliability of the input-output properties of the cortico-spinal pathway obtained from transcranial magnetic and electrical stimulation. *J. Neurosci. Methods* 112, 193–202. doi: 10.1016/S0165-0270(01)00468-X

Chen, R. (2000). Studies of human motor physiology with transcranial magnetic stimulation. *Muscle Nerve Suppl.* 9: S26–S32. doi: 10.1002/1097-4598(2000)999:9<::AID-MUS6>3.0.CO;2-I

Chen, R. (2004). Interactions between inhibitory and excitatory circuits in the human motor cortex. *Experimental Brain Research* 154, 1–10. doi: 10.1007/s00221-003-1684-1

Chen, R., and Udupa, K. (2009). Measurement and modulation of plasticity of the motor system in humans using transcranial magnetic stimulation. *Motor Control* 13, 442–453.

Christie, A., Fling, B., Crews, R. T., Mulwitz, L. A., and Kamen, G. (2007). Reliability of motor-evoked potentials in the ADM muscle of older adults. *J. Neurosci. Methods* 164, 320–324. doi: 10.1016/j.jneumeth.2007.05.011

Cincotta, M., Giovannelli, F., Borgheresi, A., Balestrieri, F., Toscani, L., Zaccara, G., et al. (2010). Optically tracked neuronavigation increases the stability of hand-held focal coil positioning: evidence from transcranial magnetic stimulation-induced electrical field measurements. *Brain Stimul.* 3, 119–123. doi: 10.1016/j.brs.2010.01.001

Corneal, S. F., Butler, A. J., and 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. doi: 10.1016/j.apmr.2004.12.039

Delvaux, V., Alagona, G., Gérard, P., De Pasqua, V., Pennisi, G., and de Noordhout, A. M. (2003). Post-stroke reorganization of hand motor area: a 1-year prospective follow-up with focal transcranial magnetic stimulation. *Clin. Neurophysiol.* 114, 1217–1225. doi: 10.1016/S1388-2457(03)00070-1

de Vet, H. C., Terwee, C. B., and Bouter, L. M. (2003). Current challenges in clinimetrics. *J. Clin. Epidemiol.* 56, 1137–1141. doi: 10.1016/j.jclinepi.2003.08.012

de Vet, H. C., Terwee, C. B., Knol, D. L., and Bouter, L. M. (2006). When to use agreement versus reliability measures. *J. Clin. Epidemiol.* 59, 1033–1039. doi: 10.1016/j.jclinepi.2005.10.015

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., and Knol, D. L. (2011). *Measurement in Medicine : A Practical Guide*. Cambridge; New York, NY: Cambridge University Press.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Eisen, S. V., Ranganathan, G., Seal, P., and Spiro, A. III. (2007). Measuring clinically meaningful change following mental health treatment. *J. Behav. Health Serv. Res.* 34, 272–289. doi: 10.1007/s11414-007-9066-2

Farzan, F., Barr, M. S., Levinson, A. J., Chen, R., Wong, W., Fitzgerald, P. B., et al. (2010). Reliability of long-interval cortical inhibition in healthy human subjects: a TMS-EEG study. *J. Neurophysiol.* 104, 1339–1346. doi: 10.1152/jn.00279.2010

Flansbjer, U. B., Holmbäck, A. M., Downham, D., Patten, C., and Lexell, J. (2005). Reliability of gait performance tests in men and women with hemiparesis after stroke. *J. Rehabil. Med.* 37, 75–82. doi: 10.1080/16501970410017215

Fleiss, J. L. (1999). *The Design and Analysis of Clinical Experiments*. New York, NY: Wiley.

Fridman, E. A., Hanakawa, T., Chung, M., Hummel, F., Leiguarda, R. C., and Cohen, L. G. (2004). Reorganization of the human ipsilesional premotor cortex after stroke. *Brain* 127(Pt 4), 747–758. doi: 10.1093/brain/awh082

Gray, S. (2015). *PktoPk.zip*. Available online at: http://ced.co.uk/downloads/signalscripts#analysis

Guyatt, G., Walter, S., and Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *J. Chronic Dis.* 40, 171–178. doi: 10.1016/0021-9681(87)90069-5

Harvill, L. (1991). Standard error of measurement. *Instruct. Topics Educat. Measurem. Summ.* 10, 33–41.

Heald, A., Bates, D., Cartlidge, N. E., French, J. M., and Miller, S. (1993). Longitudinal study of central motor conduction time following stroke. 1. Natural history of central motor conduction. *Brain* 116(Pt 6), 1355–1370. doi: 10.1093/brain/116.6.1355

Hébert, R., Spiegelhalter, D. J., and Brayne, C. (1997). Setting the minimal metrically detectable change on disability rating scales. *Arch. Phys. Med. Rehabil.* 78, 1305–1308. doi: 10.1016/S0003-9993(97)90301-4

Hoonhorst, M. H., Kollen, B. J., van den Berg, P. S., Emmelot, C. H., and Kwakkel, G. (2014). How reproducible are transcranial magnetic stimulation-induced MEPs in subacute stroke? *J. Clin. Neurophysiol.* 31, 556–562. doi: 10.1097/WNP.0000000000000114

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

Hummel, F., Celnik, P., Giraux, P., Floel, A., Wu, W. H., Gerloff, C., et al. (2005). Effects of non-invasive cortical stimulation on skilled motor function in chronic stroke. *Brain* 128(Pt 3), 490–499. doi: 10.1093/brain/awh369

Hummel, F., and Cohen, L. G. (2005). Improvement of motor function with noninvasive cortical stimulation in a patient with chronic stroke. *Neurorehabil. Neural Repair* 19, 14–19. doi: 10.1177/1545968304272698

Julkunen, P., Säisänen, L., Danner, N., Niskanen, E., Hukkanen, T., Mervaala, E., et al. (2009). Comparison of navigated and non-navigated transcranial magnetic stimulation for motor cortex mapping, motor threshold and motor evoked potentials. *Neuroimage* 44, 790–795. doi: 10.1016/j.neuroimage.2008.09.040

Kamen, G. (2004). Reliability of motor-evoked potentials during resting and active contraction conditions. *Med. Sci. Sports Exerc.* 36, 1574–1579. doi: 10.1249/01.MSS.0000139804.02576.6A

Khedr, E. M., Abdel-Fadeil, M. R., Farghali, A., and Qaid, M. (2009). Role of 1 and 3 Hz repetitive transcranial magnetic stimulation on motor function recovery after acute ischaemic stroke. *Eur. J. Neurol.* 16, 1323–1330. doi: 10.1111/j.1468-1331.2009.02746.x

Khedr, E. M., Shawky, O. A., El-Hammady, D. H., Rothwell, J. C., Darwish, E. S., Mostafa, O. M., et al. (2013). Effect of anodal versus cathodal transcranial direct current stimulation on stroke rehabilitation: a pilot randomized controlled trial. *Neurorehabil. Neural Repair* 27, 592–601. doi: 10.1177/1545968313484808

Koski, L., Lin, J. C., Wu, A. D., and Winstein, C. J. (2007). Reliability of intracortical and corticomotor excitability estimates obtained from the upper extremities in chronic stroke. *Neurosci. Res.* 58, 19–31. doi: 10.1016/j.neures.2007.01.007

Kujirai, T., Caramia, M. D., Rothwell, J. C., Day, B. L., Thompson, P. D., Ferbert, A., et al. (1993). Corticocortical inhibition in human motor cortex. *J. Physiol.* 471, 501–519. doi: 10.1113/jphysiol.1993.sp019912

Lexell, J. E., and Downham, D. Y. (2005). How to assess the reliability of measurements in rehabilitation. *Am. J. Phys. Med. Rehabil.* 84, 719–723. doi: 10.1097/01.phm.0000176452.17771.20

Liepert, J. (2006). Motor cortex excitability in stroke before and after constraint-induced movement therapy. *Cogn. Behav. Neurol.* 19, 41–47. doi: 10.1097/00146965-200603000-00005

Liepert, J., Bauder, H., Wolfgang, H. R., Miltner, W. H., Taub, E., and Weiller, C. (2000a). Treatment-induced cortical reorganization after stroke in humans. *Stroke* 31, 1210–1216. doi: 10.1161/01.STR.31.6.1210

Liepert, J., Restemeyer, C., Kucinski, T., Zittel, S., and Weiller, C. (2005). Motor strokes: the lesion location determines motor excitability changes. *Stroke* 36, 2648–2653. doi: 10.1161/01.STR.0000189629.10603.02

Liepert, J., Storch, P., Fritsch, A., and Weiller, C. (2000b). Motor cortex disinhibition in acute stroke. *Clin. Neurophysiol.* 111, 671–676. doi: 10.1016/S1388-2457(99)00312-0

Liu, H., and Au-Yeung, S. S. (2014). Reliability of transcranial magnetic stimulation induced corticomotor excitability measurements for a hand muscle in healthy and chronic stroke subjects. *J. Neurol. Sci.* 341, 105–109. doi: 10.1016/j.jns.2014.04.012

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Pub. Co.

Maeda, F., Gangitano, M., Thall, M., and Pascual-Leone, A. (2002). Inter- and intra-individual variability of paired-pulse curves with transcranial magnetic stimulation (TMS). *Clin. Neurophysiol.* 113, 376–382. doi: 10.1016/S1388-2457(02)00008-1

Malcolm, M. P., Triggs, W. J., Light, K. E., Shechtman, O., Khandekar, G., and Gonzalez Rothi, L. J. (2006). Reliability of motor cortex transcranial magnetic stimulation in four muscle representations. *Clin. Neurophysiol.* 117, 1037–1046. doi: 10.1016/j.clinph.2006.02.005

Manganotti, P., Acler, M., Zanette, G. P., Smania, N., and Fiaschi, A. (2008). Motor cortical disinhibition during early and late recovery after stroke. *Neurorehabil. Neural Repair* 22, 396–403. doi: 10.1177/1545968307313505

McGinley, M., Hoffman, R. L., Russ, D. W., Thomas, J. S., and Clark, B. C. (2010). Older adults exhibit more intracortical inhibition and less intracortical facilitation than young adults. *Exp. Gerontol.* 45, 671–678. doi: 10.1016/j.exger.2010.04.005

McGraw, K. O., and Wong, S. P. (1996). Forming Inferences about some intraclass correlation coefficients. *Psychol. Methods* 1, 30–46. doi: 10.1037/1082-989X.1.1.30

McMillan, A. S., Watson, C., Walshaw, D., and Taylor, J. P. (1998). Improved reproducibility of magnetic stimulation-evoked motor potentials in the human masseter by a new method for locating stimulation sites on the scalp. *Arch. Oral Biol.* 43, 665–668. doi: 10.1016/S0003-9969(98)00050-8

Medical Research Council of U.K. (1978). *Aids to Examination of the Peripheral Nervous System*. Palo Alto, CA; Pendragon House.

Menon, P., Kiernan, M. C., and Vucic, S. (2014). Cortical excitability differences in hand muscles follow a split-hand pattern in healthy controls. *Muscle Nerve* 49, 836–844. doi: 10.1002/mus.24072

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. doi: 10.1016/j.jclinepi.2010.02.006

Mortifee, P., Stewart, H., Schulzer, M., and Eisen, A. (1994). Reliability of transcranial magnetic stimulation for mapping the human motor cortex. *Electroencephalogr. Clin. Neurophysiol.* 93, 131–137. doi: 10.1016/0168-5597(94)90076-0

Nakamura, H., Kitagawa, H., Kawaguchi, Y., and Tsuji, H. (1997). Intracortical facilitation and inhibition after transcranial magnetic stimulation in conscious humans. *J. Physiol.* 498(Pt 3), 817–823. doi: 10.1113/jphysiol.1997.sp021905

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory*. New York, NY: McGraw-Hill.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Pascual-Leone, A., Cohen, L. G., Brasil-Neto, J. P., and Hallett, M. (1994). Non-invasive differentiation of motor cortical representation of hand muscles by mapping of optimal current directions. *Electroencephalogr. Clin. Neurophysiol.* 93, 42–48. doi: 10.1016/0168-5597(94)90090-6

Peinemann, A., Lehner, C., Conrad, B., and Siebner, H. R. (2001). Age-related decrease in paired-pulse intracortical inhibition in the human primary motor cortex. *Neurosci. Lett.* 313, 33–36. doi: 10.1016/S0304-3940(01)02239-X

Phillips, C. G., and Porter, R. (1977). Corticospinal neurones. Their role in movement. *Monogr. Physiol. Soc.* v–xii, 1–450.

Portney, L. G., and Watkins, M. P. (2009). *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Prashantha, D. K., Sriranjini, S. J., Sathyaprabha, T. N., Nagaraja, D., and Pal, P. K. (2013). Evaluation of the motor cortical excitability changes after ischemic stroke. *Ann. Indian Acad. Neurol.* 16, 394–397. doi: 10.4103/0972-2327.116955

Rossini, P. M., Desiato, M. T., and Caramia, M. D. (1992). Age-related changes of motor evoked potentials in healthy humans: non-invasive evaluation of central and peripheral motor tracts excitability and conductivity. *Brain Res.* 593, 14–19. doi: 10.1016/0006-8993(92)91256-E

Rothwell, J. C. (2010). Plasticity in the human motor system. *Folia Phoniatr. Logop.* 62, 153–157. doi: 10.1159/000314030

Ruohonen, J., and Ilmoniemi, R. J. (1999). Modeling of the stimulating field generation in TMS. *Electroencephalogr. Clin. Neurophysiol. Suppl.* 51, 30–40.

Sacco, R. L., Benjamin, E. J., Broderick, J. P., Dyken, M., Easton, J. D., Feinberg, W. M., et al. (1997). American Heart Association Prevention Conference. IPrevention, V., and Rehabilitation of Stroke. Risk factors. *Stroke* 28, 1507–1517. doi: 10.1161/01.STR.28.7.1507

Sagan, C. (1995). *The Demon-haunted World : Science as a Candle in the Dark*. New York, NY: Random House.

Sanger, T. D., Garg, R. R., and Chen, R. (2001). Interactions between two different inhibitory systems in the human motor cortex. *J Physiol* 530(Pt 2), 307–317. doi: 10.1111/j.1469-7793.2001.0307l.x

Schonfeldt-Lecuona, C., Thielscher, A., Freudenmann, R. W., Kron, M., Spitzer, M., and Herwig, U. (2005). Accuracy of stereotaxic positioning of transcranial magnetic stimulation. *Brain Topogr.* 17, 253–259. doi: 10.1007/s10548-005-6033-1

Schuck, P., and Zwingmann, C. (2003). The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. *Int. J. Rehabil. Res.* 26, 85–91. doi: 10.1097/00004356-200306000-00002

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

Smith, M. J., Keel, J. C., Greenberg, B. D., Adams, L. F., Schmidt, P. J., Rubinow, D. A., et al. (1999). Menstrual cycle effects on cortical excitability. *Neurology* 53, 2069–2072. doi: 10.1212/WNL.53.9.2069

Stinear, C. M., Byblow, W. D., and Ward, S. H. (2014). An update on predicting motor recovery after stroke. *Ann. Phys. Rehabil. Med.* 57, 489–498. doi: 10.1016/j.rehab.2014.08.006

Streiner, D. L., and Norman, G. R. (2008). *Health Measurement Scales : A Practical Guide to their Development and Use*. Oxford; New York, NY: Oxford University Press.

Swayne, O. B., Rothwell, J. C., Ward, N. S., and Greenwood, R. J. (2008). Stages of motor output reorganization after hemispheric stroke suggested by longitudinal studies of cortical physiology. *Cereb. Cortex* 18, 1909–1922. doi: 10.1093/cercor/bhm218

Takechi, U., Matsunaga, K., Nakanishi, R., Yamanaga, H., Murayama, N., Mafune, K., et al. (2014). Longitudinal changes of motor cortical excitability and transcallosal inhibition after subcortical stroke. *Clin. Neurophysiol.* 125, 2055–2069. doi: 10.1016/j.clinph.2014.01.034

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., 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. doi: 10.1016/j.jclinepi.2006.03.012

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., and Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Med. Educ.* 10:40. doi: 10.1186/1472-6920-10-40

Uy, J., Ridding, M. C., and Miles, T. S. (2002). Stability of maps of human motor cortex made with transcranial magnetic stimulation. *Brain Topogr.* 14, 293–297. doi: 10.1023/A:1015752711146

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

Werhahn, K. J., Conforto, A. B., Kadom, N., Hallett, M., and Cohen, L. G. (2003). Contribution of the ipsilateral motor cortex to recovery after chronic stroke. *Ann. Neurol.* 54, 464–472. doi: 10.1002/ana.10686

Wheaton, L. A., Villagra, F., Hanley, D. F., Macko, R. F., and Forrester, L. W. (2009). Reliability of TMS motor evoked potentials in quadriceps of subjects with chronic hemiparesis after stroke. *J. Neurol. Sci.* 276, 115–117. doi: 10.1016/j.jns.2008.09.012

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. doi: 10.1016/j.clinph.2004.02.027

Wyrwich, K. W., Tierney, W. M., and Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J. Clin. Epidemiol.* 52, 861–873. doi: 10.1016/S0895-4356(99)00071-2

Wyrwich, K. W., and Wolinsky, F. D. (2000). Identifying meaningful intra-individual change standards for health-related quality of life measures. *J. Eval. Clin. Pract.* 6, 39–49. doi: 10.1046/j.1365-2753.2000.00238.x

Yi, X., Fisher, K. M., Lai, M., Mansoor, K., Bicker, R., and Baker, S. N. (2014). Differences between Han Chinese and Caucasians in transcranial magnetic stimulation parameters. *Exp. Brain Res.* 232, 545–553. doi: 10.1007/s00221-013-3763-2