The reliability of commonly used electrophysiology measures

Brown KE[1], Lohse KR[2], Mayer IMS[3,4], Strigaro G[4], Desikan M[4], Casula EP[4], Meunier S[5], Popa T[5], Lamy J-C[5], Odish O[6], Leavitt BR[7], Durr A[5], Roos RAC[6], Tabrizi SJ[8], Rothwell JC[4], Boyd LA[1], Orth M[3]

Affiliations:

[1]Department of Physical Therapy, University of British Columbia, Vancouver, BC, Canada

[2]School of Kinesiology, Auburn University, Auburn, AL, USA

[3]Department of Neurology, Ulm University Hospital, Ulm, Germany

[4]Sobell Department of Motor Neuroscience and Movement Disorders, Institute of Neurology, University College London, London, UK

[5]APHP Department of Genetics, Groupe Hospitalier Pitié-Salpêtrière, and Institut du Cerveau et de la Moelle, INSERM U1127, CNRS UMR7225, Sorbonne Universités – UPMC Université Paris VI UMR_S1127, Paris, France

[6]Department of Neurology, Leiden University Medical Centre, 2300RC Leiden, The Netherlands

[7]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, Vancouver BC, V5Z 4H4 Canada

[8]Huntington's Disease Research Centre, UCL Institute of Neurology, London, UK

Contact for correspondence

Michael Orth, M.D., Ph.D.

Department of Neurology, Ulm University Hospital

Oberer Eselsberg 45/1; 89081 Ulm, Germany

e-mail: michael.orth@uni-ulm.de; Tel: +49-731 50063095; Fax: +49-731 50063082

Abstract

Background: Electrophysiological measures can help understand brain function both in healthy individuals and in the context of a disease. Given the amount of information that can be extracted from these measures and their frequent use, it is essential to know more about their inherent reliability.

Objective/Hypothesis: To understand the reliability of electrophysiology measures in healthy individuals. We hypothesized that measures of threshold and latency would be the most reliable and least susceptible to methodological differences between study sites.

Methods: Somatosensory evoked potentials from 112 control participants, long-latency reflexes, transcranial magnetic stimulation with resting and active motor thresholds, motor evoked potential latencies, input/output curves, and short-latency afferent inhibition and facilitation from 84 controls were collected at 3 visits over 24 months at 4 Track-On HD study sites. Reliability was assessed using intra-class correlation coefficients for absolute agreement and the effects of reliability on statistical power are demonstrated for different sample sizes and study designs.

Results: Measures quantifying latencies, thresholds, and evoked responses at high stimulator intensities had the highest reliability, and required the smallest sample sizes to adequately power a study. Very few between-site differences were detected.

Conclusions: Reliability and susceptibility to between-site differences should be evaluated for electrophysiological measures before including them in study designs. Levels of reliability vary substantially across electrophysiological measures, though there are few between-site differences. To address this, reliability should be used in conjunction with theoretical calculations to inform sample size and ensure studies are adequately powered to detect true change in measures of interest.

Keywords: transcranial magnetic stimulation; motor thresholds; SEP; study design; reliability; statistical power

Introduction

Electrophysiological measures can improve understanding of brain function both in healthy individuals and in the context of a disease. Numerous electrophysiological experimental paradigms probe the function of the cortex and white matter connections [1]. Some paradigms examine brain function at the time of stimulation while others aim to modulate brain function so that the effects outlast the time of stimulation and therefore probably reflect plasticity [2]. Commonly used techniques examining brain function include somatosensory-evoked potential latencies and amplitudes (SEP) to examine primary somatosensory cortical (S1) excitability and sensory afferent connections following peripheral nerve stimulation, and transcranial magnetic stimulation (TMS) in combination with electromyography (EMG) to explore excitability in the corticospinal tract (motor thresholds, motor evoked potential latencies and amplitudes) and local circuitry within the primary motor cortex (M1), e.g. cortical silent period duration [3]. Additionally, the combination of peripheral nerve stimulation and TMS, e.g. in sensory afferent inhibition and facilitation [4,5] or long latency reflexes [6] can be used to test cortical circuitry involved in sensorimotor integration [7].

Given the amount of information that can be extracted from these measures and their frequent use, it is essential to know more about their inherent reliability. In cross-sectional studies, for instance, poor reliability increases within-group variation, making it more difficult to measure between-group differences [8,9]. In clinical trials that are more dependent on within-individual changes, as opposed to group differences, measurement error can mask change over time [8,10,11]. Both of these issues can be addressed by understanding the reliability of the dependent measures. Finally, when designing multi-centre studies the susceptibility to differences between sites in the interpretation of the protocol for electrophysiological measures can introduce further variability in the data. It is therefore critical to identify these sources of added variability as well.

To date, there is discrepancy in the literature, with some investigations reporting low to moderate reliability across measures [12-15], and separate studies reporting high reliability [16-19]. These values are largely dependent on the measures investigated, the muscle of interest, and the methods by which reliability is quantified. As such, investigations to date largely suffer from small sample sizes, inclusion of few measures, and inconsistent methodological approaches [20]. Recent work outlined these shortcomings, and provided a more comprehensive evaluation of the reliability of TMS measures in healthy older individuals as well as individuals with stroke [21].

The current study utilised a large sample of electrophysiological data in healthy individuals across 24-months to address three main aims. First, we examined feasibility of each electrophysiological measure by calculating attrition rates across time. Second, we determined the reliability of different electrophysiological measures. Thirdly, we determined which electrophysiological measures may be susceptible to methodological differences by assessing between-site differences. Finally, we conducted simulation to quantify how reliability influences statistical power when using electrophysiological measures. Broadly, we hypothesized that measures of threshold and latency would be the most reliable and least susceptible to methodological differences between study sites.

Materials and Methods

Participants

This study used electrophysiology data from 112 control participants (67 females) enrolled in the Track-On study at the four study sites (London, Paris, Leiden, Vancouver) [22,23]. TrackOn is a longitudinal observational study in participants carrying a mutation in the *HTT* gene and matched controls [22,23].

Participants were assessed at baseline, as well as at 12 and 24-month follow-ups, at four study sites: London (29 participants at baseline; 26% of total 112), Paris (29 at baseline; 26%), Leiden (28 at baseline; 25%), and Vancouver (26 at baseline; 23%). Local ethics committees approved the study, and written informed consent was obtained from each participant. The Leiden site only collected SEPs; therefore, the long-latency reflexes (LLRs) and TMS-based measures have a total of 84 control participants' data.

Electrophysiology

For all data collection, participants were seated in a comfortable chair and asked to relax as much as possible, unless instructed otherwise. All measures were collected from the dominant hemisphere and hand, assessed with the Edinburgh Handedness Questionnaire [24].

Electroencephalography

Somatosensory evoked potentials (SEPs) were recorded following median nerve stimulation (pulse width 200 μs, square wave pulse, cathode distal, anode proximal) with surface electrodes using routine techniques [25]. SEPs were recorded with a silver/silver-chloride disk electrode over the somatosensory cortex (2 cm posterior of C3 in the international EEG 10-20 system) referenced against Fz. Briefly, stimulation at 3 Hz was delivered at three intensities: *sensory threshold*, defined as the minimum intensity at which participants could perceive the stimulation at the wrist; *motor threshold*, defined as the minimum intensity required to evoke a visible twitch in the target muscle; and at *150% of the motor threshold*. Recordings from 300 stimuli were collected. At three sites (London, Paris, Vancouver), surface EMG were recorded from the right abductor pollicis brevis (APB) muscle using silver/silver-chloride disc surface electrodes (1 cm diameter) in a belly tendon montage. The EMG signal was amplified and analogue filtered (30 Hz to 1 kHz) with a Digitimer D360 amplifier (Digitimer Ltd., Welwyn Garden City, UK) in London and Paris, or Powerlab 4/30 EMG System (AD Instruments, Colorado Springs, CO) in Vancouver. Data were digitised (sampling rate 4 kHz) for offline analysis using Signal software (Cambridge Electronic Devices, Cambridge, UK) in London and Paris, or LabChart (AD Instruments, Colorado Springs, CO) in Vancouver. In order to analyse SEP data, an average trace for each stimulation intensity was produced to extract the N20 latency and N20/P25 amplitude. The N20 component was identified as the first negative peak in a time window of 15-25 ms post-stimulus; N20 latency was defined as the time from stimulation to this peak. Latency was determined from the 150% of motor threshold trace; if no peak could be detected in this trace, motor threshold was used. N20/P25 amplitude was calculated as the peak-to-peak amplitude between the N20 and the following P25.

Long-latency Reflexes

Long-latency reflexes were collected using standard procedures [6]. Three hundred stimuli were delivered over the median nerve at the wrist as individuals maintained an APB contraction of 20-30% of their maximal voluntary contraction (MVC). To activate the APB, individuals were instructed to abduct their thumbs against a force transducer while monitoring visual feedback to ensure consistency. EMG was collected as described above. Average traces were used to determine LLR2 amplitudes, as well as the latencies of both LLR1 and LLR2. LLR1 was defined as the first visible deflection from baseline between 35-45 ms post-stimulus, while

LLR2 was identified in a time window of 45-55 ms. LLR2 amplitude was defined as the peak-to-peak amplitude between LLR2 and the following peak.

Transcranial Magnetic Stimulation

TMS was performed as previously described using established techniques [3,25]. At all sites single pulse TMS was delivered using a monophasic figure-of-eight shaped coil (Magstim 70 mm P/N 9790, Magstim Co., UK) connected to a Magstim 200^2 stimulator (Magstim Co., UK). Stimuli were given with a random inter-stimulus interval (ISI) of 4-5 seconds. The coil was held in such a way to induce a posterior-anterior flow with the coil handle positioned at an angle of 45° pointing backwards. The APB 'hot-spot' was located and both resting and active motor thresholds (RMT and AMT) were determined as described [26]. The optimal spot for APB activation was marked with a felt pen (London) or coordinates recorded using neuronavigation software (Vancouver; Brainsight™, Rogue Research Inc., Montreal, QC, Canada; Paris; N eXimia 2.2.0, Nextim Ltd, Helsinki, Finland). Following threshold determination, TMS was used to collect input/output curves at rest (110%, 130%, 150% RMT) or with pre-activation (125%, 150%, 175% AMT) as described including cortical silent period determination [27,28].

MEP latency was defined as the time between the stimulus and MEP onset, while the CSP was defined as the time from the beginning of the MEP to the return of voluntary EMG activity [29]. Both were determined subjectively through visual inspection from the MEP evoked from the highest intensity of stimulation for each individual. To quantify the size of the MEP, both peak-to-peak amplitude and curve area were calculated. Curve area was determined from the unrectified MEP using each waveform's absolute amplitude multiplied by the time between samples on the channel. To investigate sensorimotor integration, median nerve stimulation was paired with TMS at various ISIs using short-latency afferent inhibition (SAI) and afferent facilitation (SAF) as described [4,5]. These measures of sensorimotor integration (SAI, SAF) were not collected at the third visit as a result of an updated protocol for the Track-On HD study.

Statistical Analysis

In order to measure the reliability of different neurophysiological measures, we first calculated the average measures two-way random-effects intra-class correlation coefficient for absolute agreement (ICC), hereafter written as *ICC(2,k)* (for more details about the statistical rationale see supplementary material). In the random-effects ICC, both people and observations are treated as random effects (i.e., we assume that both people and time points are samples from a larger population [30]). Data were filtered prior to analysis to remove participants with missing data. The ICC(2,k) can be interpreted as the ratio of true variance to total variance for *k* measures (i.e., across all three time points, [31]). In our analyses, we selected ICC(2,k) ≥0.80 as cut-point indicating a relatively stable and reliable measure with relatively little variation within a person over time compared to the individual differences between people [32].

We also calculated a single measures two-way random-effects ICC for absolute agreement, referred to as *ICC(2,1)*. The ICC(2,1) reflects the average ratio of true variance to total variance captured by any single measurement (i.e., the ICC(2,1) is equivalent to the average of the off-diagonal of a correlation matrix between all time points). Presenting ICC(2,1) as a compliment to ICC(2,k) is important because ICC(2,k) is sensitive to the number of measurements, whereas ICC(2,1) is not. That is, if the ratio of r^2_{tx}=var(T)/var(X)=0.25 in the population, ICC(2,1) will approximate 0.25 (especially in large sample sizes) regardless of the number of observations taken, whereas ICC(2,k) will increase to very high levels as k increases.

Next, we were interested in electrophysiological measures that were not statistically different across study sites. We conducted a series of MANOVAs in which the three different time points were treated as dependent measures and study site was treated as a between-subjects factor. For TMS latency measures, we also included arm length as a covariate. The Pillai-Bartlett Trace (denoted V) was used in the determination of statistical significance. In the event of a statistically significant effect of study site in the MANOVA, we conducted follow-up univariate ANOVAs at each time point, followed by pairwise comparisons of the study-sites (in the event the univariate ANOVA was statistically significant).

Finally, we used the ICC(2,1) observed in the current data to explore the statistical power available using different electrophysiological measures. To this end we constructed statistical simulations for three different study designs: 1) an independent samples t-test, 2) a paired samples t-test, and 3) the within-between interaction in a mixed-factorial ANOVA (i.e., a Group by Test interaction in a classic randomized controlled trial). The details of the statistical simulations and the code for running them are provided online (https://github.com/keithlohse/power_reliability) and these results were corroborated with power-calculation software (G*Power 3.1.9.2; [33]). These simulations were run at nine different sample sizes n/group = [10, 20, 30, 40, 50, 60, 100, 200, 300] and two different effect-sizes d = [0.5, 0.8], which correspond to traditionally moderate and large effects [34]. The code for these simulations is adaptable, however, so researchers could always adapt the code for their own power analyses.

All analyses were conducted using SPSS v23.0 (IBM, Armonk, New York). All descriptive statistics are reported as mean (SD) unless otherwise indicated. A graphical depiction of our approach can be found in Figure 1.

Results

Cohort

At visits one to three, individuals had an average age of 48.1 years (SD: 10.7), 49.4 years (10.5), and 50.6 years (10.4), respectively; 67 participants were female (60%). 103 individuals were right-handed, 6 were left-landed, and 3 were ambidextrous.

There was a significant difference in arm length across study sites, and thus arm length was controlled for in our MANOVAs for latency-based measures. Arm length in Leiden (75.98cm (5.08)), was significantly larger than in London (72.10cm (5.7)), and in Paris (71.35cm (8.05)), but not in Vancouver (74.22cm (5.23)) (*F* (3,100)=3.16, *p*=0.03).

All electrophysiological measures were well tolerated and attrition rates were low. Attrition values for each study site as well as total participant numbers can be found in Table 1. As the Leiden site exclusively collected EEG measures, there are different values for those outcomes as compared to all the other electrophysiological measures. At visit two, retention of participants was 94% for EEG measures, and 92% for all other measures. At the third visit, EEG measures were assessed on 90% of the original sample, and TMS measures were assessed on 89% of the original sample. Between the second and third visits, attrition was low at 4% and 3% for EEG and TMS measures, respectively.

Assessing Reliability Over Time: Two-Way Random-Effect ICCs for Absolute Agreement

We first examined reliability using data from the same participants across several visits. Several measures met the ICC(2,k) cut-off ≥0.80 indicating high reliability (Figure 2; Table 2). These were SEPs at motor threshold (0.91) and 150% of motor threshold (0.91), N20 latency (0.90), the latency of both LLR1 and LLR2 (0.97, 0.98), MEP amplitude at 150% of RMT (0.81), RMT and AMT (0.89 and 0.81, respectively), and both resting and active MEP latency (0.92 and 0.89 respectively). Other measures met the ICC(2,k) cut-off ≥0.50 of moderate reliability while a number of measures had low reliability (ICC(2,k) cut-off <0.50) (Table 2, see Supplementary Table 1 for low reliability measures).

To compliment the ICC for absolute agreement, we also calculated a number of descriptive statistics showing the reliability of each measure. These measures, shown in Supplementary Table 2, are the standard deviation between participants (i.e., the standard deviation of the observed values, averaged across time-points), the average within-participant standard deviation (i.e., the average standard deviation within a person over time), and the average within-participant coefficient of variation (i.e., the within-participant standard deviation divided by the mean for each participant). Across measures, the MEP active latency, MEP resting latency, N20 latency, and long-loop reflex measures showed the smallest coefficient of variation within participants.

Assessing Agreement between Study Sites

We examined whether the measures differed statistically between study sites. All measures were analysed with a MANOVA to determine whether there were between-site differences. Significant effects shown for high and moderate reliability measures can be seen in Table 3, while all other results can be found in Supplementary Tables 3 and 4. Within the measures that met the ICC(2,k) criterion of 0.8, SEPs at both motor threshold and 150% of motor threshold differed between sites (ps<0.001). For interpreting SEP measures, it also important to establish that stimulation intensities between study sites were similar. Absolute stimulation intensities across visits were similar in Leiden and London while at the Paris site intensities at Visit 2 were substantially larger than at Visits 1 or 3 (Supplementary Table 5).

Resting MEP latency also had a main effect of study site that remained despite controlling for arm length (p=0.003). Post-hoc analysis revealed that there was a significant difference between Paris and Vancouver (p=0.02) at the first time-point, and between London and Paris at the second and third time-points (p=0.005, p=0.001). In general, three main trends emerged. First, lower stimulation intensities were more likely to have between-site differences. Second, active measures were more likely to have between-site differences than their resting counterparts. Third, measures analyzed by area under the curve had more site differences than when analyzed by quantifying peak-to-peak amplitude.

Finally, there were also methodological differences between study sites that might have affected the reliability of the results. For instance, the Vancouver and Paris sites used a neuronavigation system to mark the site for TMS measures, whereas the London site did not. (Recall that the Leiden site did not collect TMS data and was excluded from these calculations.) To address this question, we calculated ICCs for each measure at each study site separately. As shown in Supplementary Table 6, there was considerable heterogeneity in the reliability between study sites, but no consistent pattern emerged suggesting that the neuronavigation system affected the reliability for TMS measures.

Simulating the Effects of Reliability on Statistical Power

We next explored the hypothetical statistical power available using different electrophysiological measures (Tables 4-6). For a paired-samples t-test, for instance, any non-zero change is always going to be statistically significant when there is no measurement error. As measurement error is added, r^2_{TX} decreases, and statistical power decreases. Consider then a study design using a paired-samples t-test to detect a medium-sized effect where the change from pre-test to post-test is one-half of the pooled standard deviation, d = 0.5. If the primary outcome were RMT then the ICC(2,1)=0.73. As ICC(2,1) is an estimate of the true r^2_{TX}, we can make a conservative estimate that r^2_{TX}=0.64 and having 40 total participants would yield ~81%

power (Table 5). Conversely, if the primary outcome were SEP at sensory threshold, the ICC(2,1)=0.32. Looking at Table 5, the closest r^2_{TX} is 0.36 and now having 40 participants would only yield ~36% power. Indeed, we would need to increase our sample size to 100 to achieve even 74% statistical power due to the lower reliability of the SEP at sensory threshold. Consider now the same measures when examining an interaction effect in a 2x2 mixed-factorial ANOVA, with a medium-sized effect (d=0.5). For RMT, with an estimated r^2_{TX} of 0.64, having a sample size of 60 would yield 69% power, while increasing the sample size to 100 would yield 89% power (Table 6). In contrast, SEP at sensory threshold with an estimated r^2_{TX} of 0.36 would require a sample size of 200 in order to achieve 78% power, with a sample size of 40 only providing 23% power (Table 6).

Discussion

Electrophysiological methods can help assess brain function. It is important to know how reliable data generated with different electrophysiological methods are in order to sufficiently power a study using such measures. Here, using healthy control data from a longitudinal study conducted at four study sites, we show that the study had low attrition rates indicating it was well tolerated. Generally, measures quantifying latencies, thresholds, and evoked responses at high stimulator intensities had the highest reliability, and required the smallest sample sizes to adequately power a study. Very few between-site differences were detected. Our data can assist in adequately powering research studies or clinical trials using electrophysiological measures.

The electrophysiological data were collected from healthy controls at four Track-On study sites in three study visits a year apart. Attrition rates were low suggesting that the electrophysiological measurements were well tolerated. Amongst the individuals who did drop out, the main reasons related to other aspects of the Track-On study including the duration of the study day, which included many other assessments. The electrophysiological data were collected at the end of the study day, so that time constraints or participant burden often affected the electrophysiological part of the protocol.

We then employed two complimentary approaches to understanding the reliability of each measure. Measures with high ICC(2,k) values have a favourable signal to noise ratio and will be better suited for detecting even small differences between groups or over time, for example as introduced by an intervention, e.g. a treatment. In accordance with previous work, the most reliable measures included TMS thresholds, measures of latency, and sensory and motor evoked-potential amplitudes at higher stimulation intensities [15,35-42]. Most amplitude measures for recruitment curves at rest or when active, and sensory short-afferent inhibition or facilitation were of moderate reliability. Conversely, most area measures, or ratios (SPD/MEP size or conditioned/unconditioned MEP size) had low reliability. Measures with high reliability may relate most closely to brain structure; for example, latencies may reflect the integrity of a particular white matter tract. Since brain structure probably did not significantly change in our healthy controls over the 2 years of the Track-On study, the within-participant variability of these measures was also low. Such reasoning may also apply to evoked-response amplitudes, in particular SEPs that were equally reliable at motor threshold and 150% motor threshold intensities; however, it is also possible that in case of the MEPs they were most reliable when they were likely near a physiological maximum response and thus high reliability may also reflect a ceiling effect. Measures with lower reliability may record brain function that is less tightly linked to brain structure. This does not mean that the methods used for these measurements are inherently less reliable, as it is also possible that the greater within-participant variability results from physiological differences in brain function. For instance, the activity and excitability of the neuronal populations being stimulated may fluctuate so that the response to stimulation depends on the state of that population at the time of stimulation [43].

Understanding the reliability of dependent measures is critical for experimental design, specifically *a priori* power calculations. Given the *a priori* effect-size of interest, the necessary sample-size can be adapted to the reliability of the outcome measure in question. We next used a simulation approach to calculate statistical power for three statistical analyses, including a simple clinical trial design. As expected, studies using measures with high reliability require the smallest sample size. However, our results also indicate that measures with less than ideal reliability may still be usable if the sample size provides sufficient statistical power. We should note, however, that there are other methods for increasing power beyond increasing sample size. For instance, study design can improve statistical power through the proper use of covariates or the use of repeated measures to improve reliability (e.g., multiple baseline and post-tests; [44]). Furthermore, variance/standard deviation measures from previous *empirical* research will capture both variation in the true scores and measurement error, whereas simply assuming a *theoretical* effect-size (e.g., adequate power to detect a hypothetical d=0.5) does not take measurement error into account. That said, effect-sizes can vary drastically from sample to sample (especially in under-powered studies). Therefore, we would recommend that research be conservative about effect-size estimates in the design of future experiments and use the reliability data from the present study to inform effect size calculations, helping to avoid underpowered study designs.

Finally, we assessed between-site differences. This is important when determining which metrics to include in large, multi-site studies. Altogether, there were few study site effects. Generally, measures that had higher stimulation intensities, and thus were probably close to a physiological ceiling had fewer between-site differences. Further, measures obtained at rest appeared to be less influenced by study-site than measures where participants had to pre-activate the muscle that was recorded from. Consistent with previous work, participants per protocol had to maintain a contraction level of about 20-30% of the maximum contraction strength measured either with a force transducer or an oscilloscope. The TrackOn electrophysiology protocol involved training of study site personnel and data monitoring. However, the data suggest that these measures may not have been sufficient and thus may not have prevented variability in participants' levels of pre-activation. Finally, peak-to-peak amplitude measures were more consistent across sites than evoked-responses quantified as area under the MEP curve. We had expected that area measures would have been more robust since area measurements account for the possibility that the recruitment of additional motor units in response to higher stimulation intensities may not necessarily be synchronous. Temporal dispersion of responses would then be captured in an increase in the area under the MEP. However, our data indicate that this may not necessarily be the case so that amplitude measures may be better than area measurements. Another explanation, however, may be that at least with higher intensities amplitudes may be close to a physiological ceiling and hence the data are more stable than when measuring area.

When comparing across sites and individuals, it is essential to control for anthropometric factors that may influence the data. For example, in the current work, we controlled for arm length to eliminate any impact that the length of neural connections in the periphery could have on our latency-based variables. Controlling for arm length, there were no consistent differences between the sites for any latency measure, except rest MEP latencies. The identification of thresholds for peripheral nerve stimulation could be another source of between-site variation for which to control. Our protocol specified that median nerve motor threshold was based on a barely visible twitch in the target muscle, though M-wave amplitudes were monitored throughout. However, this may not account for differences between participants in the anatomical make-up of the mixed sensory and motor median nerve so that in some individuals motor fibres may run closer to the skin and therefore stimulation electrode and in others deeper within the forearm. There were still some notable differences between visits in the same participants at the same site. In a study running 24 months it cannot always be guaranteed that the same investigator performs all the experiments at a given site. This may introduce variability for instance in the amount of pressure applied to the peripheral nerve stimulation electrode which then can translate into differences in threshold measures. There may be other factors that we are not aware of that also introduce variability. Our data suggest that collectively controlling for as many of these variables as possible is essential to reduce site influences in multi-centre studies.

Limitations

Despite being the largest investigation of the reliability of these electrophysiological measures to date, these data are not without limitations. As previously noted, some of these centre around methodology. For example, we cannot exclude the possibility that different experimenters may have had slightly different approaches to applying the median nerve stimulation or surface electrodes, despite training given at the different study sites. In addition, many of the electrophysiological measures included were dependent on participants maintaining a contraction. It's possible that differences between time-points or across sites may result from slight variations in how the squeeze was performed or background level of EMG generated. Both of the aforementioned limitations should be minimized given that measures are based on threshold values or percentages of maximums. While we cannot directly quantify the influence of each of these factors on reliability, we can make qualitative conclusions about the direction of these effects of reliability; steps toward standardization will have a zero-to-beneficial effect on reliability whereas a lack of standardization will have a zero-to-harmful effect on reliability.

Furthermore, it is important to remember that the reliabilities we recorded in this study (whether ICCs or CoVs) are not inherent properties of the electrophysiological measures, but instead emerge from the interaction of the measure, the method, and the sample [9]. These reliabilities are representative to the extent that our methods and sample are representative. These values should be broadly representative because the sample was obtained from four international centres and the methods used in the TrackON study are inline with current standards for electrophysiological procedures [2]. However, the more a sample population deviates from these healthy adults (e.g., a clinical population) or the more methods deviate from our own (e.g., significantly greater or fewer stimulations to obtain an MEP), the more likely the reliability of the measure is to change. That said, the reliabilities observed in this study can still serve as a useful “anchor” for other experimenters in designing a study. For instance, the moderate reliability of SEP amplitude observed here could be improved by including multiple SEP pre-tests and post-tests in the study design.

Conclusion

In conclusion, we have systematically examined the variability of electrophysiology measures that are commonly used in clinical research studies. Across a large population of healthy individuals, attrition rates were low, suggesting that electrophysiological measures were well-tolerated. Levels of reliability varied substantially. This indicates that the choice of a dependent measure should be informed not only by its theoretical interest, but also its reliability, as poor reliability has profound, negative effects on statistical power. Similarly, methods susceptible to between-site differences should be used cautiously to minimize site related differences in large multi-centre studies. Taken together, these steps can help ensure that any differences between measurements truly reflect underlying biological differences rather than methodological variability. Further work could examine other electrophysiological measures and assess intra-rater and inter-rater reliability.

Acknowledgements: The authors thank the TrackOn study participants, and the CHDI/ High Q Foundation, a not-for-profit organization dedicated to finding treatments for HD. Track-On HD Investigators: A Coleman, J Decolongon, M Fan, T Koren (University of British Columbia, Vancouver); C Jauffret, (ICM and APHP, Pitié-Salpêtrière University Hospital, Paris); A Schoonderbeek, E P ‘t Hart (Leiden University Medical Centre, Leiden); C Berna, G Owen, (University College London, London).

Funding

This work was funded by the CHDI Foundation, Inc. The funding source had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

References

[1] Boniface S, Ziemann U. Plasticity in the Human Nervous System: Investigations with Transcranial Magnetic Stimulation. 2003.

[2] Rossini PM, Burke D, Chen R, Cohen LG, Daskalakis Z, Di Iorio R, et al. Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee. Clin Neurophysiol 2015;126:1071-107.

[3] Orth M, Rothwell JC. The cortical silent period: intrinsic variability and relation to the waveform of the transcranial magnetic stimulation pulse. Clin Neurophysiol 2004;115:1076-82.

[4] Tokimura H, Di Lazzaro V, Tokimura Y, Oliviero A, Profice P, Insola A, et al. Short latency inhibition of human hand motor cortex by somatosensory input from the hand. J Physiol 2000;523 Pt 2:503-13.

[5] Kessler KR, Ruge D, Ilic TV, Ziemann U. Short latency afferent inhibition and facilitation in patients with writer's cramp. Mov Disord 2005;20:238-42.

[6] Deuschl G, Eisen A. Long-latency reflexes following electrical nerve stimulation. The Inter national Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 1999;52:263-8.

[7] Manganotti P, Zanette G, Bonato C, Tinazzi M, Polo A, Fiaschi A. Crossed and direct effects of digital nerves stimulation on motor evoked potential: a study with magnetic brain stimulation. Electroencephalogr Clin Neurophysiol 1997;105:280-9.

[8] de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.

[9] de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. New York, NY: Cambridge University Press, 2011.

[10] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.

[11] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.

[12] Jung NH, Delvendahl I, Kuhnke NG, Hauschke D, Stolle S, Mall V. Navigated transcranial magnetic stimulation does not decrease the variability of motor-evoked potentials. Brain Stimul 2010;3:87-94.

[13] Kiers L, Cros D, Chiappa KH, Fang J. Variability of motor potentials evoked by transcranial magnetic stimulation. Electroencephalogr Clin Neurophysiol 1993;89:415-23.

[14] Orth M, Snijders AH, Rothwell JC. The variability of intracortical inhibition and facilitation. Clin Neurophysiol 2003;114:2362-9.

[15] Wassermann EM. Variation in the response to transcranial magnetic brain stimulation in the general population. Clin Neurophysiol 2002;113:1165-71.

[16] Kukke SN, Paine RW, Chao CC, de Campos AC, Hallett M. Efficient and reliable characterization of the corticospinal system using transcranial magnetic stimulation. J Clin Neurophysiol 2014;31:246-52.

[17] Bastani A, Jaberzadeh S. A higher number of TMS-elicited MEP from a combined hotspot improves intra- and inter-session reliability of the upper limb muscles in healthy individuals. PLoS One 2012;7:e47582.

[18] Du X, Summerfelt A, Chiappelli J, Holcomb HH, Hong LE. Individualized brain inhibition and excitation profile in response to paired-pulse TMS. J Mot Behav 2014;46:39-48.

[19] Liu H, Au-Yeung SS. Reliability of transcranial magnetic stimulation induced corticomotor excitability measurements for a hand muscle in healthy and chronic stroke subjects. J Neurol Sci 2014;341:105-9.

[20] Beaulieu LD, Flamand VH, Masse-Alarie H, Schneider C. Reliability and minimal detectable change of transcranial magnetic stimulation outcomes in healthy adults: A systematic review. Brain Stimul 2017;10(2):196-213.

[21] Schambra HM, Ogden RT, Martinez-Hernandez IE, Lin X, Chang YB, Rahman A, et al. The reliability of repeated TMS measures in older adults and in patients with subacute and chronic stroke. Front Cell Neurosci 2015;9:335.

[22] Kloppel S, Gregory S, Scheller E, Minkova L, Razi A, Durr A, et al. Compensation in Preclinical Huntington's Disease: Evidence From the Track-On HD Study. EBioMedicine 2015;2:1420-9.

[23] Orth M, Gregory S, Scahill RI, Mayer IS, Minkova L, Kloppel S, et al. Natural variation in sensory-motor white matter organization influences manifestations of Huntington's disease. Hum Brain Mapp 2016;37(12):4615-4628.

[24] Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 1971;9:97-113.

[25] Fischer M, Orth M. Short-latency sensory afferent inhibition: conditioning stimulus intensity, recording site, and effects of 1 Hz repetitive TMS. Brain Stimul 2011;4:202-9.

[26] Rossini PM, Barker AT, Berardelli A, Caramia MD, Caruso G, Cracco RQ, et al. Non-invasive electrical and magnetic stimulation of the brain, spinal cord and roots: basic principles and procedures for routine clinical application. Report of an IFCN committee. Electroencephalogr Clin Neurophysiol 1994;91:79-92.

[27] Terao Y, Ugawa Y. Basic mechanisms of TMS. J Clin Neurophysiol 2002;19:322-43.

[28] Chen R, Lozano AM, Ashby P. Mechanism of the silent period following transranial magnetic stimulation. Evidence from epidural recordings. Exp Brain Res 1999;128:539-42.

[29] Schippling S, Schneider SA, Bhatia KP, Munchau A, Rothwell JC, Tabrizi SJ, et al. Abnormal motor cortex excitability in preclinical and very early Huntington's disease. Biol Psychiatry 2009;65:959-65.

[30] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.

[31] Trevethan R. Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. Health Serv Outcomes Res Methodol 2016:doi:10.1007/s10742-016-0156-6.

[32] Kline P. The handbook of psychological testing. London: Routledge, 2000.

[33] Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175-91.

[34] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum, 1988.

[35] Koski L, Schrader LM, Wu AD, Stern JM. Normative data on changes in transcranial magnetic stimulation measures over a ten hour period. Clin Neurophysiol 2005;116:2099-109.

[36] Ellaway PH, Davey NJ, Maskill DW, Rawlinson SR, Lewis HS, Anissimova NP. Variability in the amplitude of skeletal muscle responses to magnetic stimulation of the motor cortex in man. Electroencephalogr Clin Neurophysiol 1998;109:104-13.

[37] Truccolo WA, Ding M, Knuth KH, Nakamura R, Bressler SL. Trial-to-trial variability of cortical evoked responses: implications for the analysis of functional connectivity. Clin Neurophysiol 2002;113:206-26.

[38] Kamen G. Reliability of motor-evoked potentials during resting and active contraction conditions. Med Sci Sports Exerc 2004;36:1574-9.

[39] Malcolm MP, Triggs WJ, Light KE, Shechtman O, Khandekar G, Gonzalez Rothi LJ. Reliability of motor cortex transcranial magnetic stimulation in four muscle representations. Clin Neurophysiol 2006;117:1037-46.

[40] Ngomo S, Leonard G, Moffet H, Mercier C. Comparison of transcranial magnetic stimulation measures obtained at rest and under active conditions and their reliability. J Neurosci Methods 2012;205:65-71.

[41] Hermsen AM, Haag A, Duddek C, Balkenhol K, Bugiel H, Bauer S, et al. Test-retest reliability of single and paired pulse transcranial magnetic stimulation parameters in healthy subjects. J Neurol Sci 2016;362:209-16.

[42] McDonnell MN, Ridding MC, Miles TS. Do alternate methods of analysing motor evoked potentials give comparable results? J Neurosci Methods 2004;136:63-7.

[43] Siebner HR, Hartwigsen G, Kassuba T, Rothwell JC. How does transcranial magnetic stimulation modify neuronal activity in the brain? Implications for studies of cognition. Cortex 2009;45:1035-42.

[44] McClelland GH. Increasing statistical power without increasing sample size. Am Psychol 2000;55:963-4.

Figure legends

Figure 1: Conceptual outline for including measures in a study design. The reliability of each potential electrophysiological measure should be determined unless it is known. A power analysis can then be used to assess the sample size required to accurately detect change based on the individualized reliability of the measure of interest. For multi-site studies, a further step should be undertaken to assess whether the measure is susceptible to site-differences before including it in the study design.

Figure 2: ICC(2,k) values for each electrophysiological measure. **A.** Measures with high reliability, ICC(2,k) ≥ 0.8. **B.** Measures with moderate reliability, ICC(2,k) ≥ 0.5. **C.** Measures with low reliability, ICC(2,k) < 0.5. RMT: resting motor threshold; AMT: active motor threshold; MEP: motor-evoked potential; SEP: somatosensory evoked potential; MT: motor threshold; LLR: long-latency reflexes; ST: sensory threshold; CRT: cortical relay time; SPD: silent period duration; SAI: short-latency afferent inhibition; SAF: short-latency afferent facilitation. 22, 24, 32, 34 denote inter-stimulus intervals.

Site	Time 1	Time 2	Time 3
London	29	28 (97)	25 (86, 89)
Paris	29	29 (100)	29 (100, 100)
Vancouver	26	20 (77)	21 (81, 105)
Leiden	28	28 (100)	26 (93, 93)
Total	**112**	**105 (94)**	**101 (90, 96)**

Table 1: The number of participants at each site and totalled across sites for each time point for the electrophysiological measures. The number in brackets is the percentage of participants returning from previous time points, such that the first number in brackets is in comparison to Time 1 and the second corresponds to Time 2. Note that Leiden only collected somatosensory evoked potentials.

Measure	**Time1** **Mean (SD)**	**Time2** **Mean (SD)**	**Time3** **Mean (SD)**	**Exclusions**	**ICC** **(2,k)**	**ICC** **(2,1)**
High Reliability						
RMT (% MSO)	45.17 (8.74)	45.96 (10.30)	46.25 (9.22)	26 (31)	0.89	0.73
AMT (% MSO)	35.73 (9.27)	35.35 (7.17)	36.13 (6.82)	25 (30)	0.81	0.59
MEP Latency Rest (ms)	22.40 (1.73)	21.77 (1.75)	22.18 (1.75)	28 (33)	0.92	0.79
MEP Latency Active (ms)	20.51 (1.85)	20.98 (2.05)	20.98 (1.73)	31 (37)	0.89	0.73
Amp 150% RMT (mV)	2.02 (1.84)	2.38 (2.30)	2.51 (2.13)	34 (40)	0.81	0.59
SEP Amp MT (µV)	1.81 (1.37)	2.01 (1.53)	2.12 (1.74)	48 (43)	0.91	0.77
SEP Amp 150% MT (µV)	2.27 (1.56)	2.59 (1.81)	2.72 (2.21)	54 (48)	0.90	0.74
N20 Latency (ms)	19.78 (1.20)	19.6 (1.27)	20.06 (1.39)	44 (39)	0.90	0.75
LLR1 Latency (ms)	38.51 (3.31)	38.56 (3.03)	38.21 (2.95)	45 (54)	0.97	0.92
LLR2 Latency (ms)	49.61 (3.44)	49.29 (3.24)	49.52 (3.00)	45 (54)	0.98	0.93
Mod. Reliability						
SEP Amp ST (µV)	0.86 (0.92)	0.69 (0.60)	0.81 (0.81)	53 (47)	0.57	0.31
LLR2 Amp (mV)	0.10 (0.07)	0.10 (0.07)	0.09 (0.08)	45 (54)	0.62	0.35
CRT (ms)	7.99 (2.87)	8.11 (2.51)	8.07 (2.62)	51 (61)	0.61	0.34
Amp 130% RMT (mV)	1.23 (1.23)	1.46 (1.47)	1.44 (1.33)	31 (37)	0.70	0.44
Amp 125% AMT (mV)	1.78 (1.36)	1.61 (1.12)	1.62 (1.39)	32 (38)	0.54	0.28
Amp 150% AMT (mV)	3.36 (1.95)	3.54 (2.27)	3.74 (2.34)	34 (40)	0.68	0.41
Amp 175% AMT (mV)	3.96 (2.03)	4.51 (2.31)	4.92 (3.42)	39 (46)	0.64	0.37
Area 130% RMT (mV ms)	8.29 (8.53)	5.97 (5.42)	5.09 (4.33)	34 (40)	0.60	0.33
Area 150% RMT (mV ms)	13.39 (12.46)	9.85 (9.09)	9.33 (7.40)	36 (43)	0.52	0.26
SPD 175% AMT (ms)	166.56 (45.26)	159.31 (36.9)	151.75 (37.4)	40 (48)	0.52	0.27
SAI N22 Amp (mV)	0.78 (0.65)	0.81 (0.72)		39 (46)	0.65	0.48
SAI N24 Amp (mV)	0.85 (0.82)	0.84 (0.69)		38 (45)	0.67	0.50
SAF N32 Amp (mV)	1.39 (1.26)	1.28 (1.17)		38 (45)	0.62	0.45
SAF N32 Area (mV ms)	130.83 (70.50)	135.07 (69.48)		33 (39)	0.55	0.38

Table 2: Reliability**.** Denotes mean (standard deviation) values for each dependent measure at each time-point, exclusions, ICC(2,k), and ICC(2,1). 'Exclusions' refers to participants who had to be excluded due to missing ≥1 time-point (shown as count (% of total n)). Exclusions for SEP measures are out of 112, whereas exclusions for TMS measures are out of 84 (as no TMS measures were recorded at the Leiden sites). High reliability denotes an ICC(2,k) ≥ 0.8. Moderate reliability denotes an ICC(2,k) between 0.5 and 0.8. Abbreviations: ST: sensory threshold, MT: motor threshold, LLR: long-latency reflex, CRT: cortical relay time, RMT: resting motor threshold, AMT: active motor threshold, SPD: silent period duration, SAI: short-latency afferent inhibition, SAF: short afferent facilitation. 22, 24, 32, 34 represent inter-stimulus intervals between the nerve stimulation and TMS. Values for dependent measures with low reliability can be found in the supplementary materials.

Measure	**V**	**df**	**p**	**London**	**Paris**	**Vancouver**	**Leiden**
High Reliability							
Rest Latency (ms)	0.35	6, 100	0.003	21.43 (1.55)	22.66 (1.75)	22.16 (1.41)	N/A
SEP Amp MT (μV)	0.32	9, 180	<0.001	1.63 (0.87)	0.97 (1.14)	1.87 (1.18)	2.79 (1.60)
SEP Amp 150% MT (μV)	0.42	9, 162	<0.001	2.16 (0.99)	1.22 (1.14)	2.74 (1.81)	3.60 (1.96)
Moderate Reliability							
Amp 125% AMT (mV)	0.44	6, 96	<0.001	2.10 (0.81)	1.41 (0.90)	1.26 (0.91)	N/A
Amp 150% AMT (mV)	0.31	6, 92	0.02	3.99 (1.40)	3.40 (2.03)	3.09 (1.87)	N/A
Area 130% RMT (mV)	0.29	6, 92	0.02	5.51 (3.76)	7.28 (6.17)	4.63 (2.84)	N/A
SPD 175% AMT (mV)	0.51	6, 80	0.001	159.08 (32.09)	165.72 (19.96)	138.95 (31.29)	N/A

Table 3: Effect of study site. MANOVA results are shown for significant results for measures of high and moderate reliability. Similar results that did not reach statistical significance or for measures with low reliability can be found in the appendix. For latency variables, the MANOVA controlled for the arm length of each participant. Mean (SD) are shown for each study site averaging across time points. Only participants with data from all three time-points were included in this analysis. Leiden data was only available for EEG measures. Abbreviations: SEP: somatosensory evoked potential; MT: motor threshold; RMT: resting motor threshold; AMT: active motor threshold; SPD: silent period duration.

Simulated independent t-test results, Cohen's d = 0.5.

	Measurement error in dependent variable					
	T	X	X	X	X	X
n/group	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	0.18	0.16	0.13	0.11	0.09	0.08
20	0.32	0.27	0.22	0.19	0.15	0.12
30	0.47	0.39	0.31	0.27	0.20	0.16
40	0.58	0.49	0.39	0.34	0.25	0.19
50	0.69	0.58	0.48	0.41	0.31	0.23
60	0.77	0.66	0.56	0.47	0.37	0.27
100	**0.94**	**0.87**	0.76	0.69	0.54	0.41
200	**0.99**	**0.99**	**0.97**	**0.94**	**0.83**	0.71
300	**1.00**	**0.99**	**0.99**	**0.99**	**0.95**	**0.87**

Simulated independent t-test results, Cohen's d = 0.8.

	Measurement error in dependent variable					
	T	X	X	X	X	X
n/group	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	0.40	0.33	0.27	0.22	0.19	0.13
20	0.68	0.61	0.50	0.42	0.34	0.23
30	**0.86**	**0.79**	0.68	0.60	0.47	0.33
40	**0.94**	**0.89**	**0.80**	0.71	0.59	0.42
50	**0.98**	**0.94**	**0.88**	**0.81**	0.69	0.52
60	**0.99**	**0.98**	**0.93**	**0.87**	0.77	0.58
100	**1.00**	**1.00**	**0.99**	**0.98**	**0.94**	**0.79**
200	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
300	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 4: Statistical power obtained as a function of sample size and reliability for independent-samples t-tests. Note that cells contain the statistical power (power of ≥0.8 is highlighted) observed across 10,000 simulated t-tests with a 1:1 group allocation ratio and α = 0.05 (i.e., % of significant results out of 10,000). Columns reflect decreasing reliability, X at different levels of $r^2_{TX'}$, from the original "true" outcome, T. These data were simulated based on a classical test theory model of $X_{ij} = T_i + \varepsilon_{ij}$ where T is a standard normal variable representing the true score and ε is a random normal variable representing measurement error. The variance of ε is adjusted to produce the desired correlation between T and X in the population.

Simulated paired t-test results, Cohen's d = 0.5.

	Measurement error in dependent variable					
	$T_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$
Total N =	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	**1.00**	0.53	0.26	0.16	0.11	0.09
20	**1.00**	**0.86**	0.49	0.32	0.20	0.14
30	**1.00**	**0.97**	0.68	0.46	0.28	0.19
40	**1.00**	**0.99**	**0.81**	0.58	0.36	0.24
50	**1.00**	**1.00**	**0.88**	0.69	0.44	0.29
60	**1.00**	**1.00**	**0.94**	0.77	0.51	0.33
100	**1.00**	**1.00**	**1.00**	**0.93**	0.74	0.51
200	**1.00**	**1.00**	**1.00**	**1.00**	**0.96**	**0.82**
300	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.94**

Simulated paired t-test results, Cohen's d = 0.8.

	Measurement error in dependent variable					
	$T_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$
Total N	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	**1.00**	**0.90**	0.55	0.35	0.22	0.15
20	**1.00**	**0.99**	**0.89**	0.66	0.43	0.28
30	**1.00**	**0.99**	**0.98**	**0.84**	0.60	0.41
40	**1.00**	**1.00**	**1.00**	**0.93**	0.74	0.51
50	**1.00**	**1.00**	**1.00**	**0.97**	**0.83**	0.62
60	**1.00**	**1.00**	**1.00**	**0.99**	**0.90**	0.69
100	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.90**
200	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**
300	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 5: Statistical power obtained as a function of sample size and reliability for paired-samples t-tests. Note that cells contain the statistical power (power of ≥0.8 is highlighted) observed across 10,000 simulated paired-samples t-tests with α = 0.05 (i.e., % of significant results out of 10,000). Columns reflect decreasing reliability, X at different levels of $r^2_{TX'}$, from the original "true" outcome, T. These data were simulated based on a classical test theory model of $X_{ij} = T_i + \varepsilon_{ij}$ where T is a standard normal variable representing the true score and ε is a random normal variable representing measurement error. The variance of ε is adjusted to produce the desired correlation between T and X in the population. Post-test T was a linear transformation

of pre-test, T_{post} = T_{pre}+d, and independent ε were added to the pre- and post-test to create the $X_{pre/post}$ variables.

Simulated interaction results, post-test Cohen's d = 0.5 (no pre-test difference).

	Measurement error in dependent variable					
	$T_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$
n/group =	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	**1.00**	0.34	0.16	0.12	0.09	0.08
20	**1.00**	0.62	0.29	0.19	0.13	0.10
30	**1.00**	**0.79**	0.40	0.28	0.17	0.12
40	**1.00**	**0.90**	0.51	0.34	0.23	0.15
50	**1.00**	**0.95**	0.61	0.42	0.26	0.18
60	**1.00**	**0.98**	0.69	0.48	0.32	0.21
100	**1.00**	**1.00**	**0.89**	0.71	0.48	0.32
200	**1.00**	**1.00**	**0.99**	**0.94**	0.78	0.55
300	**1.00**	**1.00**	**1.00**	**0.99**	**0.91**	0.73

Simulated interaction results, post-test Cohen's d = 0.8 (no pre-test difference).

	Measurement error in dependent variable					
	$T_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$	$X_{pre/post}$
n/group =	(no error)	r^2_{TX}=0.81	r^2_{TX}=0.64	r^2_{TX}=0.49	r^2_{TX}=0.36	r^2_{TX}=0.25
10	**1.00**	0.70	0.35	0.22	0.14	0.11
20	**1.00**	**0.95**	0.62	0.40	0.27	0.18
30	**1.00**	**0.99**	**0.80**	0.57	0.37	0.24
40	**1.00**	**0.99**	**0.90**	0.69	0.47	0.31
50	**1.00**	**1.00**	**0.95**	**0.79**	0.57	0.38
60	**1.00**	**1.00**	**0.98**	**0.86**	0.64	0.44
100	**1.00**	**1.00**	**0.99**	**0.98**	**0.85**	0.65
200	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**	**0.92**
300	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**

Table 6: Statistical power obtained as a function of sample size and reliability for the interaction term in a 2 (Group) by 2 (Time) mixed-factorial ANOVA. Note that cells contain the statistical power (power of ≥0.8 is highlighted) observed across 10,000 simulated interactions with α = 0.05 (i.e., % of significant results out of 10,000). Columns reflect decreasing reliability, X at different levels of r^2_{TX}, from the original "true" outcome, T. These data were simulated based on a classical test theory model of $X_{ij} = T_i + \varepsilon_{ij}$ where T is a standard normal variable representing the true score and ε is a random normal variable representing measurement error. The variance of ε is adjusted to produce the desired correlation between T and X in the population. Post-test T was a linear transformation of pre-test, $T_{post} = T_{pre}$+d for each group, and

independent ε were added to the pre- and post-test to create separate $X_{pre/post}$ variables in each group.