scientific reports

Check for updates

OPEN

A framework to assess the impact of number of trials on the amplitude of motor evoked potentials

Claudia Ammann[1,2], Pasqualina Guida[1], Jaime Caballero-Insaurriaga[1], José A. Pineda-Pardo[1,2], Antonio Oliviero[3] & Guglielmo Foffani[1,2,3]✉

The amplitude of motor evoked potentials (MEPs) elicited by transcranial magnetic stimulation (TMS) is a common yet highly variable measure of corticospinal excitability. The tradeoff between maximizing the number of trials and minimizing experimental time remains a hurdle. It is therefore important to establish how many trials should be used. The aim of this study is not to provide rule-of-thumb answers that may be valid only in specific experimental conditions, but to offer a more general framework to inform the decision about how many trials to use under different experimental conditions. Specifically, we present a set of equations that show how the number of trials affects single-subject MEP amplitude, population MEP amplitude, hypothesis testing and test–retest reliability, depending on the variability within and between subjects. The equations are derived analytically, validated with Monte Carlo simulations, and representatively applied to experimental data. Our findings show that the minimum number of trials for estimating single-subject MEP amplitude largely depends on the experimental conditions and on the error considered acceptable by the experimenter. Conversely, estimating population MEP amplitude and hypothesis testing are markedly more dependent on the number of subjects than on the number of trials. These tools and results help to clarify the impact of the number of trials in the design and reproducibility of past and future experiments.

Transcranial magnetic stimulation (TMS) is a safe, non-invasive technique based on delivering electromagnetic pulses to the cerebral cortex through a magnetic coil inducing a focused electric field in the underlying brain tissue[1]. When a single pulse of TMS is applied to the primary motor cortex with sufficient intensity, it depolarizes corticospinal neurons, eliciting a muscle contraction in the contralateral peripheral muscles, known as motor evoked potential (MEP)[1–3]. The peak-to-peak amplitude of MEPs recorded by surface electromyography (EMG) is commonly used to quantify the level of corticospinal excitability[4,5]. In the last few decades, TMS-induced MEPs have been increasingly used to obtain neurophysiological information about human motor function and mechanistic insights into neurological disorders[6–8].

Unfortunately, MEP amplitude displays high trial-to-trial variability, owing to both experimental and biological factors[9]. The intrinsic fluctuations in MEP amplitude depend on the state of ongoing oscillatory activity of cortical neurons beneath the TMS coil[10–12] and on the changing synchronization of motor neuron discharges at the spinal level[13]. Experimental factors like location on the scalp[14], coil orientation[15], stimulus intensity[9,16], and probably small changes in coil positioning (tilt, roll and twist) are also linked to MEP amplitude variability. The level of attention[17], and muscle activation of the subject[16,18] additionally affect MEP amplitude and variability. In population terms, MEP variability also depends on gender and age[19]. Even though some of these factors may be partly controlled by careful experimental designs and, to some extent, by the use of neuronavigation[20,21], MEP amplitude remains a substantially variable measure.

A common strategy to deal with trial-to-trial variability is to model MEP amplitude as a stochastic variable whose "true" probability distribution depends on all the possible sources of experimental and biological variability. Consequently, even though the "true" MEP amplitude does not exist in reality, its expected value can

[1]HM CINAC, Hospital Universitario HM Puerta del Sur, HM Hospitales, Universidad CEU-San Pablo, Madrid, Spain. [2]CIBERNED, Instituto de Salud Carlos III, Madrid, Spain. [3]Hospital Nacional de Parapléjicos, Toledo, Spain. ✉email: gfoffani.hmcinac@hmhospitales.com

be estimated by averaging over trials. When planning a TMS experiment, therefore, a basic methodological question always arises: how many trials should be used? One might simply say: the more, the better. However, experimental time is often limited and the biological and experimental conditions—and thus the "true" probability distribution—are likely to change over time. For example, an experiment might be designed to capture a biological phenomenon that is a priori delimited in time. Likewise, it may be difficult to guarantee stable attention and arousal during long experiments, and long protocols may even induce complex metaplasticity/anti-gating effects[22,23]. Estimating MEP amplitude thus becomes a tradeoff between maximizing the number of trials and minimizing the experimental time. Could 10 trials be enough? Or 20, or 30? What would be gained by using 100? In other words, what is the estimation error expected with a given number of trials? Should the same number of trials be used in a single-case study and in an experiment with 100 subjects? And does the number of trials have the same impact on statistical comparisons with independent-measures vs. repeated-measures designs? Recent studies have been designed to provide rule-of-thumb empirical answers to some of these questions, specifically to estimate single-subject MEP amplitude in certain experimental conditions[24–31]. However, since MEP variability depends on many experimental and biological sources and of the specific TMS technique employed, rule-of-thumb answers (e.g. at least 30 trials) are unlikely to fit all experimental situations. Here we aim to offer a more general theoretical framework to inform the decision about how many trials to use in TMS experiments.

The manuscript is organized as follows. First, we provide an analytical demonstration that some empirical approaches used in previous studies to define the minimum number of trials to estimate MEP amplitude[26–29,31] have limited implications. We then present a more principled general framework—derived from basic statistical reasoning—that clarifies the impact of number of trials on single-subject MEP amplitude, population MEP amplitude, hypothesis testing and test–retest reliability. We subsequently validate the equations with Monte Carlo simulations. Next, we apply the proposed framework in two experimental datasets. We first recorded 100 MEP trials in 20 subjects to provide a step-by-step application of the equations to estimate single-subject MEP amplitude and population MEP amplitude (Experiment 1). We then use the data from Experiment 1 to define the optimal number of trials and subjects to be used in a representative experiment designed to detect significant MEP amplitude differences between two stimulus intensities commonly employed in stimulus–response curves[9,16,26] (110% of the resting motor threshold [RMT] vs. 120%RMT) (Experiment 2). Beyond the specific examples, the equations and reasoning have general validity, so they can be used in a variety of experimental designs.

Results

Analytical results. *Number of trials for estimating single-subject MEP amplitude: previous studies.* In a hypothetical single-pulse TMS experiment in which MEP amplitude is collected for n_{max} trials in single subjects, the cumulative average $\widehat{\mu}_{trials}(n)$ is defined as the average MEP amplitude obtained with the first n trials, so that $\widehat{\mu}_{trials}(n_{max})$ is the sample average with all trials. For simplicity, we will refer to the sample average (with n_{max} trials) simply as $\widehat{\mu}_{trials}$. Previous studies empirically defined the optimal number of trials for estimating single-subject MEP amplitude as the minimum number of trials n_{opt} that allows the cumulative average to come within a certain level of 'acceptable similarity' to the sample average. Two main measures of 'acceptable similarity' were used: (i) a 95% confidence interval (n_{opt_ci})[26–29,31], and (ii) a ± 10% difference ($n_{opt\_\%diff}$)[28] around the sample average.

We can define the inclusion of the cumulative average within the desired level of acceptable similarity as a probability of inclusion p_{incl}, so that $\alpha = 1 - p_{incl}$. With central limit theorem assumptions, here we show that both n_{opt_ci} and $n_{opt\_\%diff}$ are analytical functions of n_{max}, namely:

$$n_{opt\_ci} = \frac{n_{max}}{1 + \left(\frac{z_{1-\alpha_{ci}/2}}{z_{1-\alpha/2}}\right)^2}, \tag{1}$$

$$n_{opt\_\%diff} = \frac{1}{\frac{1}{n_{max}} + \left(\frac{\eta\widehat{\mu}_{trials}}{z_{1-\alpha/2}\widehat{\sigma}_{trials}}\right)^2}, \tag{2}$$

where $z_{1-\alpha_{ci}/2}$ is the critical value of the standard normal distribution for a confidence interval of $1 - \alpha_{ci}$ (e.g. for a 95% c.i., $\alpha = 0.05$ and $z_{1-\alpha/2} = 1.96$), $z_{1-\alpha/2}$ is the critical value corresponding to the probability of inclusion p_{incl}, η is the relative error that defines the acceptable difference from the sample average (e.g. for ± 10%, $\eta = 0.1$), $\widehat{\mu}_{trials}$ and $\widehat{\sigma}_{trials}$ are the sample average and standard deviation across trials (computed with n_{max}). For derivations, see "Methods". Unfortunately, n_{opt_ci} does not depend on the variability $\widehat{\sigma}_{trials}$, and for both n_{opt_ci} and $n_{opt\_\%diff}$ the cumulative average is a priori bound to reach the required 'acceptable similarity' to the sample average with a number of trials that depends on and is upper bounded by the total number of trials available n_{max}.

In Eq. (1), the definition of the optimal number of trials n_{opt_ci} is solely a function of n_{max}, $z_{1-\alpha_{ci}/2}$ and $z_{1-\alpha/2}$. The above-cited studies were empirically trying to define the minimum number of trials n_{opt_ci} that allowed the cumulative average to come within a 95% confidence interval around the sample average. They thus assumed $\alpha_{ci} = 0.05$, which implies $z_{1-\alpha_{ci}/2} = 1.96$. They were also using $p_{incl} = 1$, which would correspond to a theoretical $z_{1-\alpha/2} = +\infty$, but in practice corresponded to an arbitrary $p_{incl} < 1$ due to the finite number of subjects. For example, if $z_{1-\alpha/2} = 2.493$, which corresponds to an arbitrary but mathematically elegant inclusion probability p_{incl} between 0.95 and 0.99, then $n_{opt\_ci} = n_{max}/\varphi$, where $\varphi = 1.618$ is the golden ratio. With this 'golden' inclusion probability, the 'optimal' number of trials estimated by the inclusion of the cumulative average within a 95% confidence interval around the true average would be $n_{opt\_ci} = 19$, 25 and 62 with $n_{max} = 30$, 40 and 100, respectively. With an empirical $p_{incl} = 1$, as used in previous studies[26–29,31], if the number of subjects increases, then the experimental estimate of n_{opt_ci} asymptotically tends to n_{max}.

Unlike Eq. (1), Eq. (2) does take into account the trial-to-trial variability of MEP amplitude $\widehat{\sigma}_{trials}$. Unfortunately, however, it still depends on (and is limited by) the total number of trials available n_{max}. For example, if $\widehat{\mu}_{trials} = 1$, $\widehat{\sigma}_{trials} = 0.5$, $\eta = 0.1$ and $z_{1-\alpha/2} = 2.493$, then $n_{opt\_\%diff} = 25$, 32 and 61 with $n_{max} = 30$, 40 and 100, respectively. With $p_{incl} = 1$ as previously used empirically[28], if the number of subjects increases, then the experimental estimate of $n_{opt\_\%diff}$ also asymptotically tends to n_{max}.

Number of trials for estimating single-subject MEP amplitude: a principled framework. In order to avoid the limitations of previous empirical studies attempting to define the number of trials for estimating single-subject MEP amplitude, we rescue a more principled measure of 'acceptable similarity' that had already been used in the early TMS literature[14]: the inclusion of the cumulative average within an acceptable difference (e.g. ± 10%) from the *true* average. The optimal number of trials n_{opt} for estimating single-subject MEP amplitude is thus simply obtained as the number of trials at which the confidence interval of the estimate of the true average equals the acceptable difference from the true average, i.e.

$$n_{opt} = \left[\frac{z_{1-\alpha/2}\sigma_{trials}}{\eta\mu_{trials}}\right]^2 = \left[\frac{z_{1-\alpha/2}}{\eta}CV_{trials}\right]^2, \tag{3}$$

where the critical value $z_{1-\alpha/2}$ is now defined by the desired probability of inclusion p_{incl} within the relative error η around the true average μ_{trials}, and CV_{trials} is the corresponding coefficient of variation (i.e. $CV_{trials} = \sigma_{trials}/\mu_{trials}$). For example, if $CV_{trials} = 0.5$, then 96 trials are necessary to ensure that the estimated single-subject MEP amplitude stays within 10% of the true value ($\eta = 0.1$) with 95% probability ($p_{incl} = 0.95$, $z_{1-\alpha/2} = 1.96$). Crucially, in Eq. (3) n_{opt} is not upper-bounded by the total number of trials available n_{max}. Therefore, n_{opt} can also be rigorously estimated from experimental data (without dependence on n_{max}), by substituting the true CV_{trials} with the sample estimate $\widehat{CV}_{trials}$.

Note that Eq. (3) can also be derived as the theoretical asymptotic limit of Eq. (2) for a very-large number of trials, when the sample average $\widehat{\mu}_{trials}$ converges to the true average μ_{trials}:

$$\lim_{n_{max}\to\infty}\left(\frac{1}{\frac{1}{n_{max}} + \left(\frac{\eta\widehat{\mu}_{trials}}{z_{1-\alpha/2}\widehat{\sigma}_{trials}}\right)^2}\right) = \left[\frac{z_{1-\alpha/2}}{\eta}CV_{trials}\right]^2. \tag{4}$$

Equation (3) can be solved for η to calculate the relative error (i.e. the acceptable difference from the true average) that is implicitly assumed when the MEP amplitude is estimated with a given number of trials n, i.e.

$$\eta(n) = \frac{z_{1-\alpha/2}}{\sqrt{n}}CV_{trials} = \frac{z_{1-\alpha/2}}{\mu_{trials}}SE_{trials}(n), \tag{5}$$

where $SE_{trials}(n)$ is simply the standard error of $\widehat{\mu}_{trials}$ estimating μ_{trials} with n trials. The statistical error thus decreases with the inverse of the square root of n. For example, if $CV_{trials} = 0.5$ and $z_{1-\alpha/2} = 1.96$, then reducing the number of trials n from 96 to 30 or 20 increases the relative error η from 10.0% to 17.9% and 21.9%, respectively.

Number of trials for estimating population MEP amplitude. In many studies the objective may be to estimate the average MEP amplitude of a population of N subjects, which we will refer to as the population MEP amplitude.

Substituting trials with subjects, Eq. (5) remains valid to calculate the relative error $\eta(N)$ that is assumed acceptable when the population MEP amplitude $\mu_{subjects}$ is estimated with N subjects, given the coefficient of variation across subjects $CV_{subjects}$ or the standard error of the population MEP amplitude $SE_{subjects}(N)$, i.e.

$$\eta(N) = \frac{z_{1-\alpha/2}}{\sqrt{N}}CV_{subjects} = \frac{z_{1-\alpha/2}}{\mu_{subjects}}SE_{subjects}(N). \tag{6}$$

In Eq. (6) the statistical error decreases with the inverse of the square root of the number of subjects N. In order to understand how the error depends on the number of trials n, we can decompose the variance between subjects with n trials, $\sigma^2_{subjects}(n)$, into the sum of the asymptotic variance between subjects with infinite trials $\sigma^2_{subjects}$ and the error variance of the sample average within subjects due to the finite number of trials n[32]:

$$\sigma^2_{subjects}(n) = \sigma^2_{subjects} + \frac{\sigma^2_{trials}}{n}, \tag{7}$$

where σ^2_{trials} is the MEP variance across trials, either assumed to be equal across subjects or pooled across subjects. The standard error of the population MEP amplitude then becomes

$$SE_{subjects}(N, n) = \sqrt{\frac{\sigma^2_{subjects} + \frac{\sigma^2_{trials}}{n}}{N}}. \tag{8}$$

The relative error η of the population MEP amplitude thus depends on the number of trials n as follows:

$$\eta(N,n) = \frac{z_{1-\alpha/2}\sqrt{\sigma^2_{subjects} + \frac{\sigma^2_{trials}}{n}}}{\mu_{subjects}\sqrt{N}}. \tag{9}$$

Equations (8) and (9) show that the statistical error can be reduced by increasing either the number of trials n or the number of subjects N. However, increasing the number of trials n provides only limited benefit. For example, consider a hypothetical population of $N = 20$ subjects with $\mu_{subjects} = 1.0$ mV, $\sigma_{subjects} = 0.5$ mV and $\sigma_{trials} = 0.5$ mV. The minimum relative error η achievable for estimating the population MEP amplitude with an infinite number of trials is 21.9%. If we reduce the number of trials from infinite to 10 or even 5, then the error only increases to 23.0% and 24.0%, respectively. With 10 trials, if we double σ_{trials} from 0.5 to 1.0 mV, the error only increases from 23.0 to 25.9%. Conversely, the error can always be decreased by increasing the number of subjects N.

Number of trials for hypothesis testing. In many experimental situations, one might be interested in knowing if a certain number of trials is sufficient to perform hypothesis testing, for example to test if MEP amplitude is significantly different in a population of patients compared to a population of controls (unpaired), or if it is significantly different before and after an intervention on the same population of subjects (paired). The same reasoning used to estimate the population MEP amplitude can be applied to express the t statistic for a Student's paired t-test as a function of the number of subjects N and trials n:

$$t(N,n) = \frac{\widehat{\mu}_{subjects1} - \widehat{\mu}_{subjects2}}{\sqrt{\frac{2\left[\sigma^2_{subjects}(1-r) + \frac{\sigma^2_{trials}}{n}\right]}{N}}}, \tag{10}$$

where $\widehat{\mu}_{subjects1}$ and $\widehat{\mu}_{subjects2}$ are the population MEP amplitudes of the two populations to be compared, assuming for simplicity equal variances, and r is the asymptotic correlation of MEPs between the two populations (i.e. the correlation that would be obtained within an infinite number of trials). Note that if we assume $r = 0$, then Eq. (10) represents an unpaired t-test with equal N and equal variances. A derivation of Eq. (10) is provided in the Methods.

The relationship between the number of trials and statistical power may be seen more directly in the corresponding formula for the calculation of the sample size N_{opt} in a power analysis for the t-test[32]:

$$N_{opt}(n) = 2\frac{\left[\sigma^2_{subjects}(1-r) + \frac{\sigma^2_{trials}}{n}\right]\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\mu_{subjects1} - \mu_{subjects2}\right)^2}, \tag{11}$$

where α is the probability of type 1 error and β is the probability of type 2 error ($1 - \beta$ is the power). With typical values of $\alpha = 0.05$ and $\beta = 0.20$ (i.e. $z_{1-\alpha/2} + z_{1-\beta} = 2.80$), Eq. (11) becomes:

$$N_{opt}(n) = 15.68\frac{\sigma^2_{subjects}(1-r) + \frac{\sigma^2_{trials}}{n}}{\left(\mu_{subjects1} - \mu_{subjects2}\right)^2}. \tag{12}$$

For example, if $\sigma_{subjects} = 0.5$ mV and $\sigma_{trials} = 0.5$ mV and we want to detect a difference $\mu_{subjects1} - \mu_{subjects2} = 0.2$ mV, Eq. (12) indicates the following. In a between-subjects design ($r = 0$), with only one trial ($n = 1$) we would need two groups of at least 196 subjects. By increasing the number of trials to $n = 5$ or 10, the number of subjects would conveniently decrease to 118 and 108, respectively. However, further increasing the number of trials would lead to negligible additional reduction of the number of subjects needed (e.g. 103 subjects with $n = 20$ trials, 101 subjects with $n = 40$ trials, 98 subjects with $n = \infty$ trials). In a within-subjects design with high correlation ($r = 0.9$), with only one trial ($n = 1$) we would need at least 108 subjects. By increasing the number of trials to $n = 5$, 10 or 20, the number of subjects would decrease considerably to 30, 20 and 15, respectively. Further increasing the number of trials would lead to a progressively smaller reduction of the number of subjects needed (e.g. 14 subjects with $n = 30$ trials, 13 subjects with $n = 40$, 10 subjects with $n = \infty$ trials).

Number of trials for test–retest reliability. Finally, the number of trials n clearly has an impact on the test–retest reliability of TMS measures[33], as reported in previous experimental studies[28,31]. In the case of MEP amplitude, we can show this impact analytically. For simplicity, we focus on Pearson's correlation, which is useful to assess test–retest reliability when only two time points are available, particularly if means and variances do not change across time points[34]. Substituting Eq. (7) in Eq. (40) (see "Methods"), the dependence of the Pearson's correlation coefficient $r(n)$ between repeated measures on the number of trials n within measures can be expressed as follows:

$$r(n) = r\frac{\sigma^2_{subjects}}{\sigma^2_{subjects} + \frac{\sigma^2_{trials}}{n}}, \tag{13}$$

where r and $\sigma^2_{subjects}$ are the asymptotic Pearson's correlation across repeated measures and the (pooled) variance across subjects with infinite trials, and σ^2_{trials} is the (pooled) variance across trials. Note that if mean and

variance do not change across time points (which should be the case in the context of test–retest reliability of TMS measures), then the Pearson's correlation coefficient is identical to the concordance correlation coefficient[34], which in turn is virtually identical to a group of intraclass correlation coefficients that estimate the degree of absolute agreement between non-interchangeable measurements[35–37]. Equation (13) clarifies that increasing the number of trials can only increase the test–retest reliability up to a limit (i.e. r), which is consistent with previous experimental observations[28,31].

Simulation results. *Single-subject MEP amplitude.* To validate Eq. (5), we simulated 10,000 single subjects with non-normally distributed MEPs at four levels of CV_{trials} (0.25, 0.50, 0.75 and 1.00). For each subject, we simulated 100 trials drawn from an independent lognormal distribution with mean μ_{trials} = 1.0 mV and standard deviation σ_{trials} = 0.25, 0.5, 0.75 or 1.00 mV, with a corresponding skewness = 0.77, 1.63, 2.67, 4.0. The lognormal distribution was obtained as the exponential of a normal distribution with mean

$$\mu = \log\left(\frac{\mu_{trials}^2}{\sqrt{\mu_{trials}^2 + \sigma_{trials}^2}}\right), \tag{14}$$

and variance

$$\sigma^2 = \log\left(1 + \frac{\sigma_{trials}^2}{\mu_{trials}^2}\right). \tag{15}$$

We then calculated the cumulative average MEP amplitude for each subject. We finally calculated the 95th percentile of the distribution across subjects of the absolute errors of the cumulative average estimating the true average (divided by 1 mV), as a function of the number of trials n. This 95th percentile was used as an estimate of the relative error η of the single-subject MEP amplitude. Note that this means that we considered a 95% probability of inclusion of the cumulative average within the relative error η from the true average [i.e. $z_{1-\alpha/2}$ = 1.96 in Eq. (5)]. The comparison between the simulated data and Eq. (5) is provided in Fig. 1A.

Population MEP amplitude. To validate Eq. (9), we simulated 10,000 populations of N = 10, 20, 30 and 40 subjects. For each population of subjects, the single-subject MEP amplitude $\mu_{trials}(s)$ of each subject s was drawn from a lognormal distribution with mean $\mu_{subjects}$ = 1.0 mV (i.e. the true population MEP amplitude) and standard deviation $\sigma_{subjects}$ = 0.5 mV (skewness = 1.63). For each subject s within each population, we simulated 100 trials drawn from an independent lognormal distribution with mean $\mu_{trials}(s)$ and standard deviation σ_{trials} = 0.5 mV. We then calculated the cumulative population MEP amplitude for each population of subjects. Finally, we calculated the 95th percentile of the distribution across subjects of the absolute errors of the cumulative population MEP amplitude, estimating the true population MEP amplitude (divided by 1 mV), as a function o the number of trials n. This 95th percentile (i.e. $z_{1-\alpha/2}$ = 1.96 in Eq. (9)) was used as an estimate of the relative error η of the population MEP amplitude. The comparison between the simulated data and Eq. (9) is provided in Fig. 1B.

T-statistic for hypothesis testing. To validate Eq. (10), we simulated 10,000 population pairs of N = 10, 20, 30 and 40 subjects each. For each population pair, the single-subject MEP amplitude $\mu_{trials}(s_i)$ of each subject s_i (with i = 1 or 2) was drawn from a bivariate lognormal distribution with either

(i) means $\mu_{subjects1}$ = 1.4 mV and $\mu_{subjects2}$ = 1.0 mV, standard deviation $\sigma_{subjects}$ = 0.5 mV and covariance 0 (unpaired t-test), or
(ii) means $\mu_{subjects1}$ = 1.2 mV and $\mu_{subjects2}$ = 1.0 mV, standard deviation $\sigma_{subjects}$ = 0.5 mV and covariance r* 0.25, with r = 0.9 (paired t-test).

The bivariate lognormal distribution was obtained as the exponential of a bivariate normal distribution with means

$$\mu_1 = \log\left(\frac{\mu_{subjects1}^2}{\sqrt{\mu_{subjects1}^2 + \sigma_{subjects}^2}}\right), \tag{16}$$

$$\mu_2 = \log\left(\frac{\mu_{subjects2}^2}{\sqrt{\mu_{subjects2}^2 + \sigma_{subjects}^2}}\right), \tag{17}$$

variances

$$\sigma_1^2 = \log\left(1 + \frac{\sigma_{subjects}^2}{\mu_{subjects1}^2}\right), \tag{18}$$

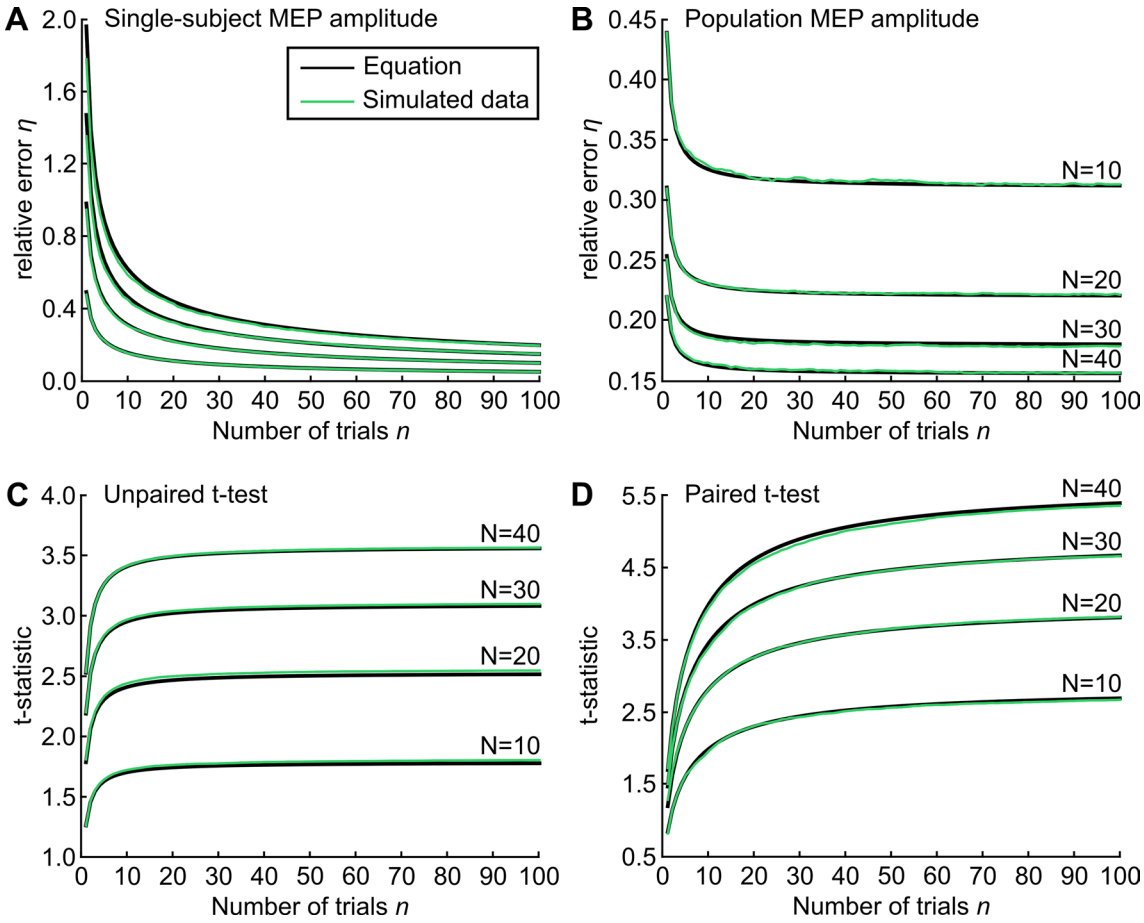

Figure 1. Number of trials for single-subject MEP amplitude, population MEP amplitude and hypothesis testing. (**A**) Single-subject MEP amplitude. With a given number of trials n (x-axis), the single-subject MEP amplitude is expected to be with 95% probability (i.e. $z_{1-\alpha/2} = 1.96$) within a relative error η (y-axis) around the true average, depending on the coefficient of variation (CV_{trials}= 0.25, 0.50, 0.75, 1.0). The lines represent Eq. (5) (black) and 10,000 single subjects simulated with lognormally distributed MEP amplitudes (green). (**B**) Population MEP amplitude. Representative example with $\mu_{subjects} = 1$ mV, $\sigma_{subjects} = 0.5$ mV, $\sigma_{trials} = 0.5$ mV. With a given number of trials n (x-axis), the population MEP amplitude is expected to be with 95% probability (i.e. $z_{1-\alpha/2} = 1.96$) within a relative error η (y-axis) around the true average, depending on the number of subjects (N = 10, 20, 30, 40). The lines represent Eq. (9) (black) and 10,000 populations of subjects simulated with lognormally distributed MEP amplitudes (green). (**C**) Unpaired t-test. Representative example with $\mu_{subjects1} = 1.4$ mV, $\mu_{subjects2} = 1.0$ mV, $\sigma_{subjects} = 0.5$ mV, $\sigma_{trials} = 0.5$ mV and $r = 0$. The t statistic is plotted as a function of the number of trials n, depending on the number of subjects (N= 10, 20, 30, 40). The lines represent Eq. (10) (black) and 10,000 populations of subjects simulated with lognormally distributed MEP amplitudes (green). (**D**) Paired t-test. Representative example with $\mu_{subjects1} = 1.2$ mV, $\mu_{subjects2} = 1.0$ mV, $\sigma_{subjects} = 0.5$ mV, $\sigma_{trials} = 0.5$ mV and $r = 0.9$. The t statistic is plotted as a function of the number of trials n, depending on the number of subjects (N = 10, 20, 30, 40). Lines as in (**C**). Note that equations (black lines) and simulated data (green lines) are highly overlapping.

$$\sigma_2^2 = \log\left(1 + \frac{\sigma_{subjects}^2}{\mu_{subjects2}^2}\right) \tag{19}$$

and covariance

$$\rho\sigma_1\sigma_2 = \log\left(1 + \frac{r\sigma_{subjects}^2}{\mu_{subjects1}\mu_{subjects2}}\right). \tag{20}$$

Note that we considered the unpaired t-test with equal sample sizes as a special case of the paired t-test (with null covariance). For each subject s_i within each population pair, we simulated 100 trials drawn from a lognormal distribution with mean $\mu_{trials}(s_i)$ and standard deviation $\sigma_{trials} = 0.5$ mV, and we calculated the cumulative average MEP amplitude across trials. For each population pair, we then computed the average and standard deviation across subjects of the cumulative MEP amplitude differences. To reduce bias, the estimate of the standard deviation was divided by the following correction factor:

$$c_4(N) = 1 - \frac{1}{4N} - \frac{7}{32N^2} - \frac{19}{128N^3}. \quad (21)$$

Population averages and standard deviations of the cumulative MEP amplitude differences were then averaged across population pairs. The t statistic was estimated with the standard formula, as a function of the number of simulated trials n and simulated subjects N:

$$t(n,N) = \frac{\bar{d}\ (n,N)}{s_d(n,N)/\sqrt{N}}, \quad (22)$$

where $\bar{d}\ (n,N)$ and s_d represent the mean and standard deviation of the cumulative MEP amplitude differences averaged across all population pairs. The comparisons between the simulated data and Eq. (10) are provided in Fig. 1C (unpaired) and in Fig. 1D (paired).

Experimental results. *Experiment 1.* In the first experiment we addressed the relatively simple problem of estimating MEP amplitude (Fig. 2A). In 20 subjects we set a stimulus intensity intended to evoke approximately 1–1.5 mV MEPs and we delivered 100 single pulses of TMS to the cortical location ('hot spot') representing the FDI. Note that 100 trials is an arbitrary number that is considerably higher than that which is typically used in TMS protocols. Importantly, we did not control for possible attentional drifts over the approximate 10 min required to complete the 100-trial protocol, but we are assuming stationarity for simplicity.

Were 100 trials sufficient to estimate single-subject MEP amplitude? The estimated MEP variability within subjects ($\widehat{CV}_{trials}$) was 0.61 (range 0.29 to 0.87). According to Eq. (3), if we wanted to guarantee that the estimated single-subject MEP amplitude was with 95% probability (i.e. $z_{1-\alpha/2} = 1.96$) within an arbitrary error of ± 10.0% (i.e. $\eta = 0.1$) from the true MEP amplitude, we should have increased the number of trials to 143 (range 33–291). Yet Eq. (5) indicates that with our 100 trials the actual difference from the true MEP amplitude was not much higher, just ± 12.0% (range 5.7–17.1%). Using only 30 or 20 trials, the error would increase to ± 21.8% and ± 26.7%, respectively (Fig. 2B,C).

Were 100 trials sufficient to estimate population MEP amplitude? The estimated MEP variability between subjects with 100 trials $\widehat{CV}_{subjects}$ was 0.39. Accordingly, Eq. (9) indicates that the estimated population MEP amplitude was with 95% probability within an error of ± 17.1% from the true population MEP amplitude. Importantly, this error would not increase much if the number of trials was decreased to 30 (± 17.7%), 20 (± 18.2%), 10 (± 19.3%.) or even 5 (± 21.4%), (Fig. 2D,E), and it virtually would not decrease further if we had an infinite number of trials (± 16.9%).

Experiment 2. As a representative example of hypothesis testing, we considered the problem of designing an experiment to test whether stimulus intensity affects MEP amplitude (although we actually know that it does). We thus decide to deliver stimuli at two intensities commonly used in stimulus–response curves: 110% and 120% of the RMT[9,16,26], and we use the results of Experiment 1 to make predictions for the following question: how many trials and subjects do we need to detect a difference in MEP amplitude between 110%RMT and 120%RMT?

In Experiment 1, the actual stimulus intensity employed was 122.5 ± 11.8% of the RMT, which elicited a population MEP amplitude $\widehat{\mu}_{subjects} = 1.48$ mV, with an estimation error of 17.1%, a pooled within-subjects MEP variability $\widehat{\sigma}_{trials(pooled)} = 1.01$ mV and an estimated asymptotic between-subjects MEP variability $\widehat{\sigma}_{subjects} = 0.57$. We thus make the following conservative estimations. (a) With 120%RMT intensity we will obtain a population MEP amplitude $\mu_{subjects1} = 1.48^*(1 - 0.171) = 1.23$ mV (i.e. the lower confidence limit from experiment 1). (b) With 110%RMT we will obtain a population MEP amplitude $\mu_{subjects2} = 1.23/2 = 0.62$ mV. (c) Both within-subject and between-subjects MEP variability will be the same at 110%RMT and at 120%RMT, i.e. $\sigma_{trials(pooled)} = 1.01$ mV and $\sigma_{subjects} = 0.57$ mV. (d) The asymptotic correlation between MEPs obtained at 110%RMT and at 120%RMT will be $r = 0.61$. The latter was estimated from the split-half correlation of the first 40 trials in Experiment 1 (i.e. the correlation of the mean MEPs estimated from the first 20 trials with the mean MEPs estimated from the next 20 trials), eliminating one outlier.

With the above numbers (Fig. 3A), Eq. (11) indicates that in order to detect a significant difference in MEP amplitude between 110%RMT and 120%RMT, with type-I error $\alpha < 0.05$ and type-II error $\beta < 0.20$ (i.e. power > 0.80), with infinite trials we would need only 6 subjects in a within-subjects design. This minimum number of subjects would increase to 7, 8, 10, and 14 with 30, 20, 10, and 5 trials, respectively. If instead we planned to perform the experiment in a between-subjects design ($r = 0$, i.e. one group tested at 110%RMT and the other group tested at 120%RMT), Eq. (11) tells us that with infinite trials we would need at least 14 subjects per group, which would increase to 16 and 18 subjects with 30 or 10 trials, respectively (Fig. 3B).

We decided to perform Experiment 2 in a within-subjects design with 10 trials per intensity and 16 subjects, in order to have more than enough power to detect a significant difference in a within-subject design (even with half of the trials), and almost sufficient power if assuming a between-subjects design. The two stimulus intensities (i.e. 110%RMT and 120%RMT) were delivered in the same experimental session, and the experiment was repeated twice to verify the consistency of the statistical results. As expected, MEP amplitude was greater at 120%RMT compared to 110%RMT both in the first session (1.57 ± 1.59 mV vs. 0.81 ± 0.85 mV) and in the second session (1.79 ± 1.64 mV vs. 0.76 ± 0.89 mV). Considering only the first 10 subjects (i.e. the minimum number of subjects to detect a significant difference as suggested by Eq. (11)), MEP amplitude was significantly higher with 120%RMT compared to 110%RMT, both in the first experimental session (paired t-test, $p = 0.010$) and in the second one ($p = 0.044$). The p-values decreased as expected considering the entire sample of 16 patients, both in

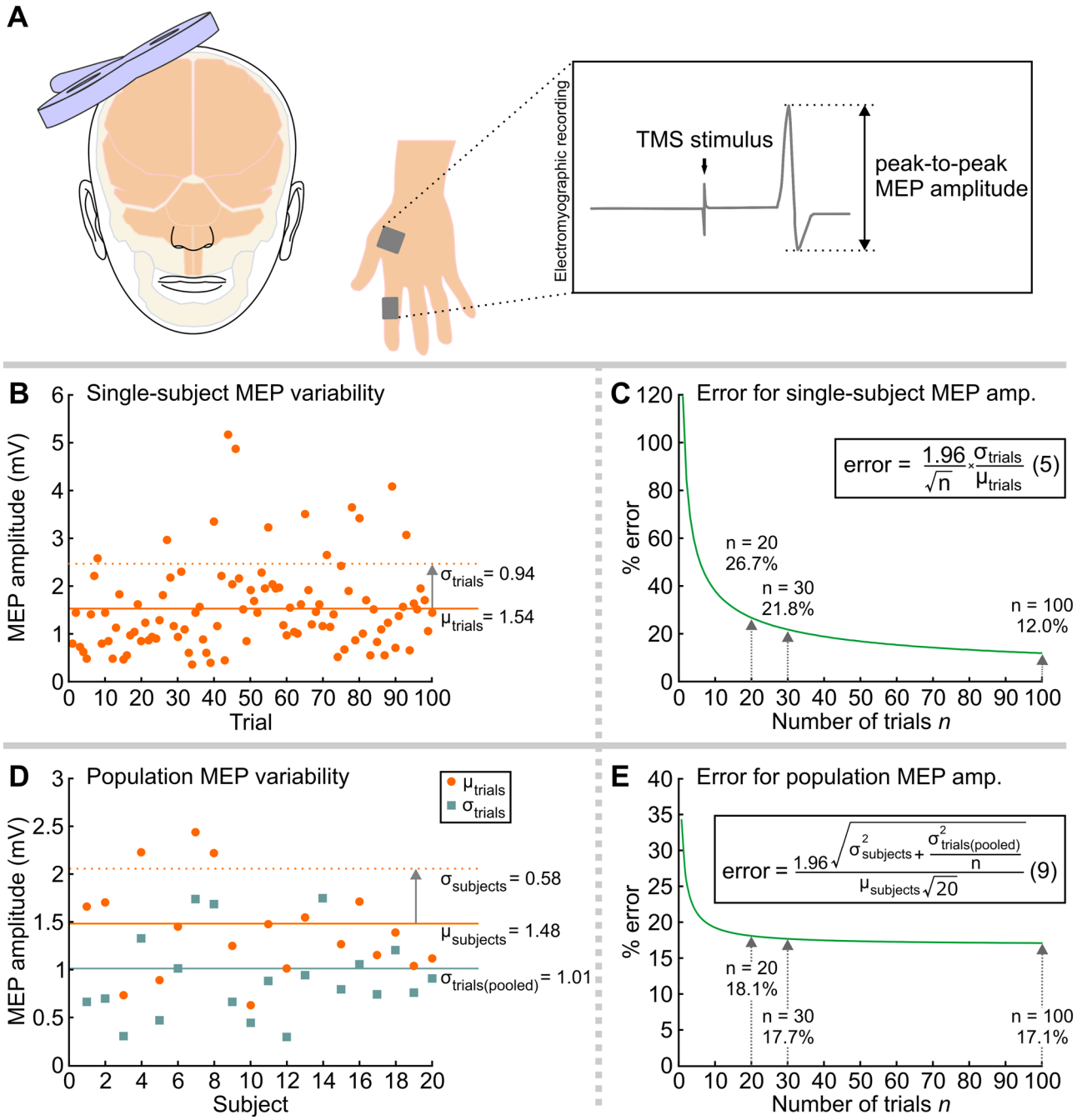

Figure 2. Validation with experimental data from Experiment 1. (**A**) Schematic experimental set up using TMS on the primary motor cortex inducing MEPs in the contralateral FDI muscle. A representative example of a recorded MEP is shown. (**B**) Peak-to-peak MEP amplitudes (mV) from one representative subject showing all 100 trials. (**C**) Experimental application of Eq. (5). (**D**) Average peak-to-peak MEP amplitude (μ_{trials}) and average standard deviation (σ_{trials}) from all subjects (N= 20) are represented. (**E**) Experimental application of Eq. (9). (**C,E**) For a 95% c.i., α= 0.05 and $z_{1-\alpha/2}$= 1.96.

the first experimental session ($p = 0.003$) and in the second one ($p < 0.001$) (Fig. 3C). As predicted, the difference remained significant even when only 5 trials were used, both in the first session ($p = 0.007$) and in the second one ($p = 0.001$). Conversely, if we assumed that the experiment was performed in a between-subjects design (i.e. two groups of 16 subjects), the p-values reached significance in the second session (unpaired t-test, $p = 0.034$), but not in the first one ($p = 0.10$), consistent with the lower statistical power that had been expected (Fig. 3D).

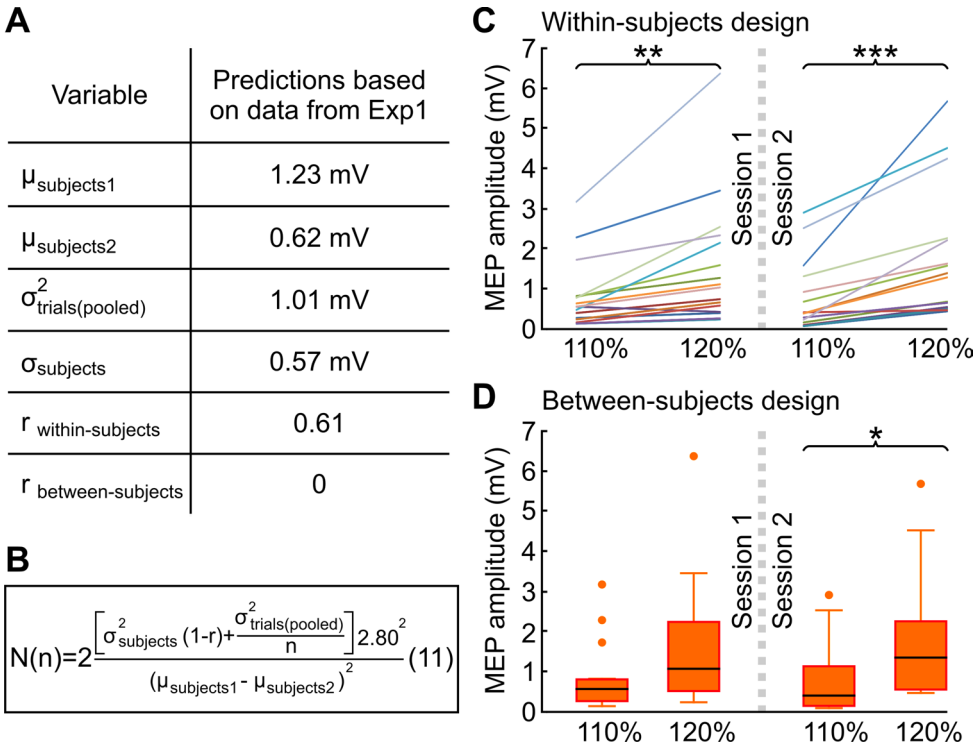

Figure 3. Experimental validation of Eq. (11). (**A**) Estimations based on data from Experiment 1 to calculate the number of trials needed to detect a difference in MEP amplitude between 110%RMT and 120%RMT. (**B**) With the numbers from (**A**), Eq. (11) determines that to detect a significant difference using 10 (or 30) trials in a within-subjects design it would require 10 (or 7) subjects, whereas in a between-subjects design it would require 18 (or 16) subjects. $\alpha = 0.05$, $\beta = 0.20$ (i.e. $z_{1-\alpha/2} + z_{1-\beta} = 2.80$). (**C**) Experimental validation of predictions made by Eq. (11) on MEP amplitude (mV) measured at two different intensities (110%RMT and 120%RMT) for a within-subjects design with 10 trials per TMS intensity and 16 subjects. The session was repeated twice (Experiment 2). Each colored line represents a single subject. Paired t-test; $^{**}p < 0.01$; $^{***}p < 0.001$. (**D**) Same as in (**C**) assuming a between-subjects design (i.e. two groups of 16 subjects), showing expected lower statistical power. Results are shown as box plots (horizontal lines: median (Q2), first quartile (Q1) and third quartile (Q3); whiskers: minimum and maximum value excluding outliers; outliers: points larger than Q3 + 1.5(Q3–Q1) or smaller than Q1–1.5(Q3–Q1). Unpaired t-test; $^{*}p < 0.05$.

Discussion

We presented a general framework of simple equations that show how the number of trials affects single-subject MEP amplitude, population MEP amplitude, hypothesis testing and test–retest reliability in TMS experiments. The equations were derived analytically, validated with Monte Carlo simulations, and applied to two sets of experimental data in a representative manner.

Analytical results. A number of recent experimental studies suggested that the minimum number of trials for estimating MEP amplitude would be around 30 trials[26,28,29]. However, we analytically showed that with the empirical approach used in these studies the estimated minimum number of trials essentially depends on total number of trials available n_{max}[Eqs. (1) and (2)] and does not depend on the trial-to-trial variability [Eq. (1)]. This probably explains why in these studies the estimated minimum number of trials n_{opt_ci} for MEPs collected at 120%RMT was higher when n_{max} was 40 (n_{opt_ci} = 29–31)[26,28,29], compared to when n_{max} was 30 (n_{opt_ci} = 21)[27]. Previous experimental estimates of the minimum number of trials to reliably estimate single-subject MEP amplitude thus do not lend themselves to generalization.

Equation (3) formalizes the intuition that the minimum number of trials to estimate single-subject MEP amplitude should depend on the trial-to-trial variability in the specific experimental conditions and on the acceptable statistical error defined by the experimenter[14]. Indeed, depending on stimulus intensity and on the stimulus–response curve of the individual subject in the specific experimental condition[18,38], MEP amplitude has a different trial-to-trial variability, as measured by the coefficient of variation (CV_{trials})[9,16,39,40]. This affects the minimum number of trials required to estimate single-subject MEP amplitude, which is proportional to the square of CV_{trials}. When the same equation is resolved in terms of the acceptable statistical error [Eq. (5)], it becomes explicit that increasing the number of trials dramatically reduces the error when only a few trials are available, but it offers a progressively smaller advantage as the number of trials increases (Fig. 1A). Nevertheless, the present study warns us that, if the acceptable error is low, in many experimental conditions estimating

single-subject MEP amplitude may require substantially more trials than previously suggested (but maintaining stationary conditions may become a challenge). However, increasing the number of trials can only improve the test–retest reliability of MEP amplitude up to a limit [Eq. (13)], in agreement with previous experimental results[28,31]. This is important, for example, for possible diagnostic applications[41,42], or for assessing the reproducibility of non-invasive brain stimulation techniques in individual subjects[43–45].

Equations (9), (10) and (11) define the impact of the number of trials for estimating population MEP amplitude and for hypothesis testing. Importantly, the non-linearity of the stimulus–response curve and its between-subjects variability contribute to both the between-subjects MEP amplitude variability $\sigma_{subjects}$ and the pooled within-subjects MEP amplitude variability σ_{trials}. This has a much higher impact on the minimum number of subjects than on the minimum number of trials required to estimate population MEP amplitude within a certain error or to detect a significant difference in hypothesis testing. In fact, the number of trials and trial-to-trial variability within subjects have a relatively minor impact on the estimation of population MEP amplitude, which mostly depends on the variability between subjects and on the number of subjects [Eq. (9); Fig. 1B]. Similarly, hypothesis testing is markedly more dependent on the number of subjects than on the number of trials [Eqs. (10) and (11)], particularly in unpaired experimental designs ($r = 0$; Fig. 1C). In paired designs ($0 < r < 1$), importantly, the number of trials becomes progressively more relevant if the asymptotic correlation r between repeated measures is higher (Fig. 1D). Nevertheless, even for highly reliable paired conditions (e.g. $r = 0.9$), a decrease in number of trials can always be compensated by an increase in number of subjects. In general, unless very few trials are used, increasing the number of trials will only induce a minor improvement in statistical power and reproducibility of comparisons between subjects (e.g. patients vs. controls) or within subjects (e.g. effect of an intervention). If more statistical power is needed, then the number of subjects rather than trials should be increased. Indeed, if sufficient subjects are available, theoretically the minimum number of trials per subject to detect any difference is always $n = 1$.

Simulation results. MEP amplitudes are typically not normally distributed. However, our analytical framework does not assume that MEP amplitudes are normally distributed: it assumes that the sample estimates of MEP amplitude means are normally distributed. Normal distribution of sample means is indeed guaranteed when the samples are normally distributed, but it also guaranteed by the central limit theory even when the samples are not normally distributed, if sufficient trials are available. To support this point, we validated Eqs. (5), (9) and (10) with Monte Carlo simulations that assumed lognormal distribution of single-trial MEP amplitudes within subjects and of single-subject MEP amplitude across subjects (Fig. 1). The results obtained with lognormal simulated data are highly consistent with the analytical equations. Note that very minor deviations from Eq. (5) are observed in the lognormal simulations, as expected, only with few trials and heavily skewed simulated data (skewness = 4 in Fig. 1; as a reference, the average skewness in Exp. 1 was 1.20, range [0.48–2.25]). Therefore, with low numbers of trials and/or in the presence of "outliers", the estimates obtained with the equations may be more accurate after normalizing the data, e.g. via an appropriate Box-Cox transformation[46,47]. Still, in most cases the equations can be readily applied to raw MEP data.

Experimental results. We provided a step-by-step application of the equations to estimate single-subject MEP amplitude and population MEP amplitude in a dataset of 100 MEP trials recorded in 20 subjects (Experiment 1). Our results show that 100 trials were sufficient to keep the estimation error of MEP amplitude below ± 20% in all our subjects, and they suggest that most experimental paradigms employing 20–30 trials (including ours) implicitly accept relatively large estimation errors for single-subject MEP amplitude. On the other hand, 100 trials were not only sufficient, but also unnecessarily high to estimate population MEP amplitude. The experimental results confirm that in the estimation of single-subject MEP amplitude the concept of "minimum number of trials" essentially depends on the error that is considered acceptable by the experimenter and the variability of MEPs in the individual subject. Conversely, the number of trials plays little role in the estimation of population MEP amplitude, which is more dependent on number of subjects.

We then used the data from Experiment 1 to define the optimal number of trials and subjects to be used in a representative experiment designed to detect significant MEP amplitude differences between two stimulus intensities (i.e. 110%RMT vs. 120%RMT; Experiment 2). Our results provide a practical example of how Eq. (11) can be used as a tool to assess the impact of the number of trials when designing new experiments. The same reasoning can be used to estimate the impact of the number of trials on experiments aiming to assess differences in MEP amplitude between groups of subjects (e.g. patients vs. controls) or changes in MEP amplitude before and after an intervention (e.g. non-invasive brain stimulation protocols).

Importantly, the equations have broad applicability and are generally valid for all experimental measures and conditions dealing with multiple trials per subject and populations of subjects. Within the TMS field, for example, the same equations can be directly applied to any measure of MEP amplitude (e.g. peak-to-peak, area, modulus, etc.) at any intensity on the stimulus–response curve, and to other single-pulse measures such as the silent period. Different experimental conditions (e.g. at rest, in activation, during a task, etc.) can be readily reflected in the equations by entering the corresponding values of within and between-subjects variability. The framework can also be extended, at least in principle, to more complex measures, such as the steepness of the stimulus–response curve and paired-pulse TMS measures. For these measures, however, some effort may be necessary to properly estimate within and between-subjects variability as a function of the number of trials. Indeed, the same equations and reasoning can also be applied to other fields (e.g. reaction times in behavioral tasks, etc.).

Practical recommendations. The aim of this study was not to provide rule-of-thumb answers that may be valid only in specific experimental conditions, but to offer a more general framework to inform the decision

about how many trials to use under different experimental conditions. Still, we can provide the following practical recommendations:

1. For estimating single-subject MEP amplitude, the minimum number of trials largely depends on the variability of the subject in the exact experimental conditions, and on the error considered acceptable by the experimenter. Equation (3) can be used to directly estimate the minimum number of trials, given the variability and the acceptable error. Equation (5) can be used to estimate the error, given the variability and the number of trials. An important caveat is that the estimate of single-subject MEP amplitude and its corresponding error refer only to the moment of the test. Their ability to represent the subject in general depends on test–retest reliability [Eq. (13)]. With this in mind, the general recommendation for estimating single-subject MEP amplitude is to use a relatively high number of trials.
2. For hypothesis testing, the number of trials plays a relatively minor role. Equation (11) can be used to explicitly estimate the impact of the number of trials in a power analysis. The general recommendation for hypothesis testing is to use at least few trials and to include a relatively high number of subjects.

Overall, we hope these simple equations will offer a useful tool to solve the issue of maximizing the number of trials and minimizing experimental time in many experimental situations, and to clarify the impact played by the number of trials on the design and reproducibility of past and future experiments.

Methods

Subjects. The study was performed according to the declaration of Helsinki and approved by the local Ethics Committee (Comité Ético de Investigación de HM Hospitales). We recruited 27 right-handed healthy participants (15 females; mean age ± standard deviation: 27.3 ± 5.7 years, 20–40 years old, 85% non-smokers) with a negative history of neurological or psychiatric conditions and medication-free at the time of the study. All subjects gave their informed consent.

Electromyographic recordings. We recorded EMG activity from the first dorsal interosseous (FDI) using disposable surface electrodes. EMG signals were band-pass filtered (2 Hz–2 kHz) and amplified (× 1000; D360, Digitimer Ltd, UK) and single trials were digitized (sample rate 5 kHz) using a CED 1401 A/D converter and Signal 5 software (Cambridge Electronic Design, Cambridge, UK). EMG signals were monitored online via visual feedback on a computer screen.

Transcranial magnetic stimulation. We used a 70-mm figure-eight-shaped magnetic coil connected to a Magstim 200^2 stimulator (Magstim Co. Ltd, UK) to perform monophasic single-pulse TMS. The coil was held tangential to the scalp with the handle oriented backwards and 45° from the midline. The induced current presented a posterior-anterior (PA) direction activating preferentially I1 waves[48,49]. Both experiments were performed using a frameless neuronavigation system (BrainSight, Rogue Research, Canada) to guide the coil position with the help of a magnetic resonance imaging template in standard space. For all experiments we measured the individual RMT, defined as the minimum TMS output intensity required to evoke a MEP peak-to-peak amplitude of ≥ 0.05 mV in five out of 10 consecutive trials in the resting FDI. We delivered TMS single pulses with 6 s ± 10% as inter-trial interval. This inter-trial interval was chosen to minimize the carryover effects in the initial transient state observed at intervals ≤ 5 s[24,50] and to be consistent with our recent studies[51–53].

Experimental procedures. We performed two independent experiments. Eighteen subjects participated in one experiment and 9 subjects participated in both. Subjects sat in a comfortable chair and were instructed to relax both arms and hands on a pillow keeping their eyes open for the duration of the experiment. Experiment 1 (n = 20; 11 females; mean age 27.7 ± 5.6 years): For each subject we determined the FDI 'hot spot' in the right motor cortex and measured the RMT. After establishing the TMS output intensity that evoked a peak-to-peak MEP amplitude of 1–1.5 mV, we recorded 100 MEPs at rest at that intensity. Experiment 2 (n = 16; 8 females; mean age 25.9 ± 4.8 years): Each subject performed two identical sessions, 7 days apart. In each session we determined the individual FDI 'hot spot' in the right motor cortex. We measured the RMT and recorded 40 MEPs at rest at different TMS output intensities (110%, 120%, 130%, and 140%RMT; randomized). Only the data from 110%RMT and 120%RMT were used in this study. In both experiments, single-trial MEP amplitude was estimated as peak-to-peak amplitude of recorded the EMG signal.

Derivation of Eq. (1). The optimal number of trials $n_{opt_{ci}}$ estimated by the inclusion of the cumulative average $\widehat{\mu}_{trials}(n)$ within a 95% confidence interval around the sample average $\widehat{\mu}_{trials}(n_{max})$, as used empirically in previous experimental studies[26–29,31], can be defined analytically. We will refer to the sample average $\widehat{\mu}_{trials}(n_{max})$ simply as $\widehat{\mu}_{trials}$, and to the true average as μ_{trials}.

First, the half width of the 95% confidence interval around the sample average $\widehat{\mu}_{trials}$ is simply $z_{1-\alpha_{ci}/2}SE(n_{max})$, where $z_{1-\alpha_{ci}/2}$ is the critical value (for a 95% c.i., α_{ci} = 0.05 and $z_{1-\alpha_{ci}/2}$ = 1.96) and $SE(n_{max})$ is the standard error of the estimate of the true average μ_{trials} with the maximum number of trials available n_{max}. Second, we can define the 'inclusion of the cumulative average' within the above confidence interval around the sample average in probabilistic terms, as the confidence interval of the estimate of the sample average made by cumulative average: $z_{1-\alpha/2}SE_{sample}(n)$, where $z_{1-\alpha/2}$ is the critical value defined by the probability of inclusion p_{incl} (i.e. $\alpha = 1 - \mathrm{p_{incl}}$) and $SE_{sample}(n)$ is the standard error of the cumulative average estimating the sample average with

n samples. Note that the above cited studies empirically used $p_{incl} = 1$, which would correspond to a theoretical $z_{1-\alpha/2} = +\infty$, but in practice corresponded to an arbitrary $p_{incl} < 1$ that depends on the number of subjects.

The optimal number of trials n_{opt_ci} is then defined as the number of trials at which the confidence interval of the estimate of the sample average made by the cumulative average equals the confidence interval of the estimate of the true average made by the sample average, i.e.

$$n_{opt\_ci} = n : z_{1-\alpha/2} SE_{sample}(n) = z_{1-\alpha_{ci}/2} SE(n_{max}). \tag{23}$$

In Eq. (23), $SE(n_{max})$ is given by the well-known formula:

$$SE(n_{max}) = \frac{\hat{\sigma}_{trials}}{\sqrt{n_{max}}}, \tag{24}$$

where $\hat{\sigma}_{trials}$ is the standard deviation of MEP amplitude across trials.

$SE_{sample}(n)$ is somewhat less straightforward. Let $\varepsilon(n)$ be the error in the estimate of the sample average made by the cumulative average with $n < n_{max}$, i.e.

$$\hat{\mu}_{trials}(n) = \hat{\mu}_{trials}(n_{max}) + \varepsilon(n). \tag{25}$$

From the decomposition of variances, it follows that:

$$\mathrm{Var}[\hat{\mu}_{trials}(n)] = \mathrm{Var}[\hat{\mu}_{trials}(n_{max})] + \mathrm{Var}[\varepsilon(n)], \tag{26}$$

where $Var[\varepsilon(n)]$ is the variance of the cumulative average estimating the sample average. Since the standard deviation of an estimator (in this case the cumulative average as an estimator of the sample average) is by definition the standard error of the estimator, we can write:

$$\mathrm{Var}[\varepsilon(n)] = SE_{sample}(n)^2. \tag{27}$$

$Var[\hat{\mu}_{trials}(n)]$ is the variance of the cumulative average estimating the true average, i.e.

$$\mathrm{Var}[\hat{\mu}_{trials}(n)] = SE(n)^2 = \frac{\hat{\sigma}^2_{trials}}{n} \tag{28}$$

and $Var[\hat{\mu}_{trials}(n_{max})]$ is the variance of the sample average estimating the true average, i.e.

$$\mathrm{Var}[\hat{\mu}_{trials}(n_{max})] = SE(n_{max})^2 = \frac{\hat{\sigma}^2_{trials}}{n_{max}}. \tag{29}$$

The variance of the cumulative average estimating the sample average $SE_{sample}(n)^2$ can thus be readily obtained by subtracting the variance of the sample average to the variance of the cumulative average estimating the true average, i.e.

$$SE_{sample}(n)^2 = SE(n)^2 - SE(n_{max})^2 = \frac{\hat{\sigma}^2_{trials}}{n} - \frac{\hat{\sigma}^2_{trials}}{n_{max}} = {\sigma_{trials}}^2 \left(\frac{1}{n} - \frac{1}{n_{max}} \right). \tag{30}$$

Substituting (24) and (30) in (23) we obtain:

$$n_{opt\_ci} = n : z_{1-\alpha/2} \hat{\sigma}_{trials} \sqrt{\frac{1}{n} - \frac{1}{n_{max}}} = \frac{z_{1-\alpha_{ci}/2} \hat{\sigma}_{trials}}{\sqrt{n_{max}}}, \tag{31}$$

which gives

$$n_{opt\_ci} = \frac{n_{max}}{1 + \left(\frac{z_{1-\alpha_{ci}/2}}{z_{1-\alpha/2}} \right)^2}, \tag{32}$$

corresponding to Eq. (1).

Derivation of Eq. (2). The optimal number of trials $n_{opt\_\%diff}$ estimated by the inclusion of the cumulative average within a ± 10% difference around the sample average, as used empirically in one previous study[28], can also be defined analytically, as follows:

$$n_{opt\_\%diff} = n : z_{1-\alpha/2} SE_{sample}(n) = \eta \hat{\mu}_{trials}, \tag{33}$$

where $z_{1-\alpha/2} SE_{sample}\left(n_{opt\_\%diff}\right)$ is the confidence interval of the estimate of the sample average made by the cumulative average, as in Eq. (23), $\eta = 0.1$ for ± 10% difference and $\hat{\mu}_{trials}$ is the sample average. Substituting (30) in (33), we obtain:

$$n_{opt\_\%diff} = n : z_{1-\alpha/2} \hat{\sigma}_{trials} \sqrt{\frac{1}{n} - \frac{1}{n_{max}}} = \eta \hat{\mu}_{trials}, \tag{34}$$

which gives:

$$n_{opt\_\%diff} = \frac{1}{\frac{1}{n_{max}} + \left(\frac{\eta\widehat{\mu}_{trials}}{z_{1-\alpha/2}\widehat{\sigma}_{trials}}\right)^2}, \tag{35}$$

corresponding to Eq. (2).

Derivation of Eq. (10). To derive Eq. (10), we start from the estimation of the t statistic in a paired Student's t-test, i.e.

$$t = \frac{\widehat{\mu}_{subjects1} - \widehat{\mu}_{subjects2}}{\sqrt{\frac{\widehat{\sigma}^2_{subjects1} + \widehat{\sigma}^2_{subjects2} - 2\mathrm{cov}(\widehat{\mu}_{trials1}, \widehat{\mu}_{trials2})}{N}}}, \tag{36}$$

where $\widehat{\boldsymbol{\mu}}_{trials1}$ and $\widehat{\boldsymbol{\mu}}_{trials2}$ are vectors of estimated single-subject MEP amplitudes for two repeated measures from the same population of subjects. Note that if we impose $\mathrm{cov}\left(\widehat{\boldsymbol{\mu}}_{\mathrm{trials1}}, \widehat{\boldsymbol{\mu}}_{\mathrm{trials2}}\right) = 0$, then Eq. (36) becomes the t statistic for an unpaired t-test between two populations with an equal number of subjects.

We assume equal variances (or pool them) so that $\widehat{\sigma}^2_{subjects1} + \widehat{\sigma}^2_{subjects2} = 2\widehat{\sigma}^2_{subjects}$, and we model the estimated single-subject MEP amplitudes as:

$$\widehat{\boldsymbol{\mu}}_{trials} = \boldsymbol{\mu}_{trials} + \boldsymbol{\varepsilon}, \tag{37}$$

where $\boldsymbol{\mu}_{trials}$ is the vector of true single-subject MEP amplitudes across subjects and $\boldsymbol{\varepsilon}$ is the corresponding error vector for estimating the single-subject MEP amplitude with a limited number of trials. Assuming that the errors are independent, the covariance term can be rewritten as follows:

$$cov(\widehat{\boldsymbol{\mu}}_{trials1}, \widehat{\boldsymbol{\mu}}_{trials2}) = r(n)\sigma^2_{subjects}(n) = cov(\boldsymbol{\mu}_{trials1} + \boldsymbol{\varepsilon}_1, \boldsymbol{\mu}_{trials2} + \boldsymbol{\varepsilon}_2) = cov(\boldsymbol{\mu}_{trials1}, \boldsymbol{\mu}_{trials2}) = r\sigma^2_{subjects}, \tag{38}$$

where $r(n)$ and $\sigma^2_{subjects}(n)$ are the Pearson's correlation across repeated measures and the (pooled) variance across subjects with n trials, whereas r and $\sigma^2_{subjects}$ are the asymptotic Pearson's correlation across repeated measures and (pooled) variance across subjects with infinite trials. Substituting Eqs. (7) and (38) in Eq. (36), we obtain:

$$t(N,n) = \frac{\widehat{\mu}_{subjects1} - \widehat{\mu}_{subjects2}}{\sqrt{\frac{2\left[\sigma^2_{subjects}(1-r) + \frac{\sigma^2_{trials}}{n}\right]}{N}}}, \tag{39}$$

which corresponds to Eq. (10). Note that

$$r = r(n)\frac{\sigma^2_{subjects}(n)}{\sigma^2_{subjects}} = r(n)\frac{\sigma^2_{subjects}(n)}{\sigma^2_{subjects}(n) - \frac{\sigma^2_{trials}}{n}} > r(n). \tag{40}$$

Therefore, $r(n)$ provides a lower bound for r, and r can be estimated from the data. Note that Eq. (40) corresponds to a classic correction for attenuation[54,55].

Code availability

The main code is given within the manuscript in form of equations (which are sufficiently simple to be readily implemented in any spreadsheet or programming language). The experimental data are available upon reasonable request.

Received: 27 April 2020; Accepted: 30 October 2020
Published online: 08 December 2020

References

1. Barker, A. T., Jalinous, R. & Freeston, I. L. Non-invasive magnetic stimulation of human motor cortex. *Lancet* **1**, 1106–1107 (1985).
2. Day, B. L. *et al.* Motor cortex stimulation in intact man. 2. Multiple descending volleys. *Brain* **110**(Pt 5), 1191–1209 (1987).
3. Rothwell, J. C. *et al.* Motor cortex stimulation in intact man. 1. General characteristics of EMG responses in different muscles. *Brain* **110**(Pt 5), 1173–1190 (1987).
4. Di Lazzaro, V. & Rothwell, J. C. Corticospinal activity evoked and modulated by non-invasive stimulation of the intact human motor cortex. *J. Physiol.* **592**, 4115–4128 (2014).
5. Bestmann, S. & Krakauer, J. W. The uses and interpretations of the motor-evoked potential for understanding behaviour. *Exp. brain Res.* **233**, 679–689 (2015).
6. Hallett, M. *et al.* Contribution of transcranial magnetic stimulation to assessment of brain connectivity and networks. *Clin. Neurophysiol.* **128**, 2125–2139 (2017).
7. Ziemann, U. Thirty years of transcranial magnetic stimulation: where do we stand?. *Exp. brain Res.* **235**, 973–984 (2017).
8. Derosiere, G., Vassiliadis, P. & Duque, J. Advanced TMS approaches to probe corticospinal excitability during action preparation. *Neuroimage* **213**, 116746 (2020).
9. Kiers, L., Cros, D., Chiappa, K. H. & Fang, J. Variability of motor potentials evoked by transcranial magnetic stimulation. *Electroencephalogr. Clin. Neurophysiol.* **89**, 415–423 (1993).

10. Bergmann, T. O. *et al.* EEG-guided transcranial magnetic stimulation reveals rapid shifts in motor cortical excitability during the human sleep slow oscillation. *J. Neurosci.* **32**, 243–253 (2012).
11. Keil, J. *et al.* Cortical brain states and corticospinal synchronization influence TMS-evoked motor potentials. *J. Neurophysiol.* **111**, 513–519 (2014).
12. de Goede, A. A. & van Putten, M. J. A. M. Infraslow activity as a potential modulator of corticomotor excitability. *J Neurophysiol.* **122**, 325–335 (2019).
13. Rösler, K. M., Roth, D. M. & Magistris, M. R. Trial-to-trial size variability of motor-evoked potentials: a study using the triple stimulation technique. *Exp. Brain Res.* **187**, 51–59 (2008).
14. Brasil-Neto, J. P., McShane, L. M., Fuhr, P., Hallett, M. & Cohen, L. G. Topographic mapping of the human motor cortex with magnetic stimulation: factors affecting accuracy and reproducibility. *Electroencephalogr. Clin. Neurophysiol.* **85**, 9–16 (1992).
15. Volz, L. J., Hamada, M., Rothwell, J. C. & Grefkes, C. What makes the muscle twitch: motor system connectivity and TMS-induced activity. *Cereb. Cortex* **25**, 2346–2353 (2015).
16. Darling, W. G., Wolf, S. L. & Butler, A. J. Variability of motor potentials evoked by transcranial magnetic stimulation depends on muscle activation. *Exp. Brain Res.* **174**, 376–385 (2006).
17. Mars, R. B., Bestmann, S., Rothwell, J. C. & Haggard, P. Effects of motor preparation and spatial attention on corticospinal excitability in a delayed-response paradigm. *Exp. Brain Res.* **182**, 125–129 (2007).
18. Devanne, H., Lavoie, B. A. & Capaday, C. Input-output properties and gain changes in the human corticospinal pathway. *Exp. Brain Res.* **114**, 329–338 (1997).
19. Pitcher, J. B., Ogston, K. M. & Miles, T. S. Age and sex differences in human motor cortex input-output characteristics. *J. Physiol.* **546**, 605–613 (2003).
20. Cincotta, M. *et al.* 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 (2010).
21. Julkunen, P. *et al.* Comparison of navigated and non-navigated transcranial magnetic stimulation for motor cortex mapping, motor threshold and motor evoked potentials. *Neuroimage* **44**, 790–795 (2009).
22. Delvendahl, I. *et al.* Occlusion of bidirectional plasticity by preceding low-frequency stimulation in the human motor cortex. *Clin. Neurophysiol.* **121**, 594–602 (2010).
23. Siebner, H. R. A primer on priming the human motor cortex. *Clin. Neurophysiol.* **121**, 461–463 (2010).
24. Schmidt, S. *et al.* An initial transient-state and reliable measures of corticospinal excitability in TMS studies. *Clin. Neurophysiol.* **120**, 987–993 (2009).
25. 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* **7**, e47582 (2012).
26. Cuypers, K., Thijs, H. & Meesen, R. L. J. Optimization of the transcranial magnetic stimulation protocol by defining a reliable estimate for corticospinal excitability. *PLoS ONE* **9**, e86380 (2014).
27. Chang, W. H. *et al.* Optimal number of pulses as outcome measures of neuronavigated transcranial magnetic stimulation. *Clin. Neurophysiol.* **127**, 2892–2897 (2016).
28. Goldsworthy, M. R., Hordacre, B. & Ridding, M. C. Minimum number of trials required for within- and between-session reliability of TMS measures of corticospinal excitability. *Neuroscience* **320**, 205–209 (2016).
29. Bashir, S. *et al.* The number of pulses needed to measure corticospinal excitability by navigated transcranial magnetic stimulation: eyes open vs close condition. *Front. Hum. Neurosci.* **11**, 121 (2017).
30. Hashemirad, F., Zoghi, M., Fitzgerald, P. B. & Jaberzadeh, S. Reliability of motor evoked potentials induced by transcranial magnetic stimulation: the effects of initial motor evoked potentials removal. *Basic Clin. Neurosci.* **8**, 43–50 (2017).
31. Biabani, M., Farrell, M., Zoghi, M., Egan, G. & Jaberzadeh, S. The minimal number of TMS trials required for the reliable assessment of corticospinal excitability, short interval intracortical inhibition, and intracortical facilitation. *Neurosci. Lett.* **674**, 94–100 (2018).
32. Bloch, D. A. Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Stat. Med.* **5**, 663–667 (1986).
33. Brown, K. E. *et al.* The reliability of commonly used electrophysiology measures. *Brain Stimul.* **10**, 1102–1111 (2017).
34. Lin, L. I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268 (1989).
35. Nickerson, C. A. A note on "A concordance correlation coefficient to evaluate reproducibility". *Biometrics* **53**, 1503–1507 (1997).
36. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46 (1996).
37. Liljequist, D., Elfving, B. & Skavberg Roaldsen, K. Intraclass correlation: a discussion and demonstration of basic features. *PLoS ONE* **14**, e0219854 (2019).
38. van Kuijk, A. A. *et al.* Stimulus-response characteristics of motor evoked potentials and silent periods in proximal and distal upper-extremity muscles. *J. Electromyogr. Kinesiol.* **19**, 574–583 (2009).
39. Capaday, C., Lavoie, B. A., Barbeau, H., Schneider, C. & Bonnard, M. Studies on the corticospinal control of human walking. I. Responses to focal transcranial magnetic stimulation of the motor cortex. *J. Neurophysiol.* **81**, 129–139 (1999).
40. Klein-Flügge, M. C., Nobbs, D., Pitcher, J. B. & Bestmann, S. Variability of human corticospinal excitability tracks the state of action preparation. *J. Neurosci.* **33**, 5564–5572 (2013).
41. Benussi, A. *et al.* Classification accuracy of transcranial magnetic stimulation for the diagnosis of neurodegenerative dementias. *Ann. Neurol.* **87**, 394–404 (2020).
42. Padovani, A. *et al.* Diagnosis of mild cognitive impairment due to Alzheimer's disease with transcranial magnetic stimulation. *J. Alzheimers. Dis.* **65**, 221–230 (2018).
43. Hinder, M. R. *et al.* Inter- and Intra-individual variability following intermittent theta burst stimulation: implications for rehabilitation and recovery. *Brain Stimul.* **7**, 365–371 (2014).
44. Dyke, K., Kim, S., Jackson, G. M. & Jackson, S. R. Intra-subject consistency and reliability of response following 2 ma transcranial direct current stimulation. *Brain Stimul.* **9**, 819–825 (2016).
45. Ammann, C., Lindquist, M. A. & Celnik, P. A. Response variability of different anodal transcranial direct current stimulation intensities across multiple sessions. *Brain Stimul.* **10**, 1–10 (2017).
46. Nielsen, J. F. Logarithmic distribution of amplitudes of compound muscle action potentials evoked by transcranial magnetic stimulation. *J. Clin. Neurophysiol.* **13**, 423–434 (1996).
47. Roy Choudhury, K. *et al.* Intra subject variation and correlation of motor potentials evoked by transcranial magnetic stimulation. *Ir. J. Med. Sci.* **180**, 873–880 (2011).
48. Sakai, K. *et al.* Preferential activation of different I waves by transcranial magnetic stimulation with a figure-of-eight-shaped coil. *Exp. brain Res.* **113**, 24–32 (1997).
49. Di Lazzaro, V. *et al.* Comparison of descending volleys evoked by transcranial magnetic and electric stimulation in conscious humans. *Electroencephalogr. Clin. Neurophysiol.* **109**, 397–401 (1998).
50 Julkunen, P., Säisänen, L., Hukkanen, T., Danner, N. & Könönen, M. Does second-scale intertrial interval affect motor evoked potentials induced by single-pulse transcranial magnetic stimulation? *Brain Stimul.* **5**(4), 526–532 (2012).
51. Dileone, M. *et al.* Dopamine-dependent changes of cortical excitability induced by transcranial static magnetic field stimulation in Parkinson's disease. *Sci. Rep.* **7**, 4329 (2017).

52. Dileone, M., Mordillo-Mateos, L., Oliviero, A. & Foffani, G. Long-lasting effects of transcranial static magnetic field stimulation on motor cortex excitability. *Brain Stimul.* **11**, 676–688 (2018).
53. Ammann, C. *et al*. Cortical disinhibition in Parkinson's disease. *Brain*. https://doi.org/10.1093/brain/awaa274 (2020).
54. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101 (1904).
55. Saccenti, E., Hendriks, M. H. & Smilde, A. K. Corruption of the Pearson correlation coefficient by measurement error and its estimation bias and correction under different error models. *Sci. Rep.* **10**, 438 (2020).

Acknowledgements

This research was funded by MINECO/AEI/FEDER, UE (SAF2017-86246-R and PRE2018-086735), and by Comunidad de Madrid (2017-T2/BMD-5231).

Author contributions

C.A. and G.F. designed the study. G.F. developed the mathematical formalism. C.A. and P.G. recruited the subjects and performed the experiments. C.A. and G.F. analyzed the data and drafted the first version of the manuscript. C.A. crafted the figures. C.A., P.G., J.C.I., J.A.P.P., A.O. and G.F. provided critical review and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.