Analysis and Comparison of Methods for Determining Transcranial Magnetic Stimulation Thresholds

Boshuo Wang[a], Angel V. Peterchev[a,b,c,d], and Stefan M. Goetz[a,b,d,e,*]

[a] Department of Psychiatry and Behavior Sciences, School of Medicine, Duke University, Durham, NC, USA

[b] Department of Electrical and Computer Engineering, School of Engineering, Duke University, Durham, NC, USA

[c] Department of Biomedical Engineering, School of Engineering, Duke University, Durham, NC, USA

[d] Department of Neurosurgery, Duke University, NC, USA

[e] Department of Engineering, School of Technology, University of Cambridge, Cambridge, UK

[*] Corresponding author: stefan.goetz@duke.edu, smg84@cam.ac.uk

Abstract

Background: Transcranial magnetic stimulation (TMS) is a noninvasive neuromodulation method with both clinical and research applications. However, cortical neuronal variability presents challenges for TMS thresholding methods.

Objective: To analyze existing methods and their variations for obtaining TMS threshold, introduce new methods from other fields, and compare their performances in terms of speed and accuracy.

Methods: We examine existing methods based on the relative-frequency approximation of the probability definition of TMS threshold and adaptive methods based on a probabilistic threshold model and maximum-likelihood as well as maximum a posteriori estimation. Variations in the estimation and stepping procedure are included to improve the performance of the latter. We adapt a Bayesian estimation method which iteratively incorporates information from the TMS response into the probability density function and introduce a family of non-parametric stochastic root-finding method with different convergence criteria and stepping rules. In addition, we demonstrate accelerated threshold detection utilizing population-level prior information. We apply all thresholding methods on a virtual test bed and characterize their estimation error.

Results. The relative-frequency methods require a large number of stimuli and, despite good accuracy on the population level, have wide error distributions for individual subjects. The parametric estimation methods obtain threshold much faster, and their accuracy depends on their estimation and stepping rules, with their performance significantly improved when population-level prior information is included. Stochastic root-finding methods are comparable to adaptive estimation methods but are much simpler to implement and do not rely on an underlying model.

Conclusion. TMS thresholding methods have a wide range of behavior and difference in performance and the optimal methods requires various considerations of trade-off.

Keywords: transcranial magnetic stimulation, threshold search, parametric estimation, Bayesian estimation, stochastic root-finding

Introduction

Motivation

Transcranial magnetic stimulation (TMS) is a noninvasive technique to activate neurons in the brain of conscious subjects (Goetz and Deng, 2017; Peterchev et al., 2015). Certain pulse rhythms can further modulate neural circuits, i.e., change how they process endogenous signals (Zhong et al., 2021). A key parameter of the stimulation pulses is their strength, which influences not only how many neurons but also which neuron types are stimulated (Aberra et al., 2020; Di Lazzaro et al., 2012). As the TMS pulse strength determines the activation of neuron types with different function, it understandably also influences efficacy of neuromodulation protocols and can even reverse the direction of the effect (Halawa et al., 2019; Lang et al., 2006; Sasaki et al., 2018; Sommer et al., 2006; Turi et al., 2021). Finally, the stimulation strength is a key parameter not only for choosing the best strength level for an effective procedure but also for safety consideration, and limitations for various modulatory protocols are recommended by a consensus of the field to avoid seizures (Rossi et al., 2009, 2020; Wassermann, 1998).

Typically, the pulse strength is individualized to a specific subject's or patient's anatomy and physiology through the so-called threshold—usually as motor threshold based on the muscular responses of finger muscles to stimulation of their representation in the primary motor cortex and rarely as phosphene threshold on visual artifacts elicited in the visual cortex (Antal et al., 2004; Deblieck et al., 2008; Goetz and Kammer, 2022; Stewart et al., 2001; Wassermann, 2002). For the motor threshold, the size of motor evoked potentials (MEPs) following stimulation in the primary motor cortex measured in electromyography (EMG) at the target muscle has widely replaced the visual observation of muscle twitches due to the large potential of overestimation for the latter (Westin et al., 2014). The threshold is considered as the lowest stimulation that can elicit MEPs of at least 50 μV peak to peak at in at least half of the trials (previously muscle twitches in half of the time).

The motor threshold is typically also used as a reference for other brain targets and its detection is recommended before any further neuromodulatory intervention, even for recurring subjects or patients (Cotovio et al., 2021; Rossini et al., 1994, 2015). Furthermore, the threshold and its changes can serve as a quantitative metric, e.g., for the diagnosis of lesions (Kobayashi and Pascual-Leone, 2003). This implies the need for a well-defined, rapid, accurate, and unbiased thresholding procedure.

However, the MEPs in response to stimulation in the primary motor cortex are highly stochastic due to the neuronal variability—a pulse administered twice with identical strength can lead to a maximum MEP response close to saturation once and no response at all in the other case (Goetz and Peterchev, 2012; Jung et al., 2010; Kiers et al., 1993). Moreover, the variability is far from Gaussian and appears to involve several variability sources (Capaday, 2021; Goetz et al., 2014, 2022; Kiers et al., 1993). Accordingly, the estimation procedure can become very time consuming and instable. In contrast to the associations the term threshold may elicit, the motor threshold is furthermore far from a binary transition between no response and response but a point on the dose–response curve, which was considered close to the lowest stimulation strength that still elicits any detectable excitatory response (Bawa and Lemon, 1993; Goetz et al., 2019; Mathias et al., 2014). Finally, the motor threshold is also likely not the lowest stimulation strength to evoke detectable responses, as recent studies demonstrate that the 50 μV threshold is but likely closer to the middle of the logarithmic dose–response curve of MEPs (Goetz et al., 2018; Li et al., 2021). In consequence, the large non-Gaussian variability challenges statistical methods and estimators. The smoother transition of the expected value of the responses in dependence of the stimulation strength, on the other hand, may allow the use of the MEP amplitude value to find the threshold faster.

Thresholding in Psychophysics

The problem of detecting a sensation threshold from data with strong stochastic influence has a longer tradition in psychophysics (Treutwein, 1995). Accordingly, the methods in that field are far more sophisticated than the commonly used and also more recent approaches in TMS. However, there are very important differences in psychophysics. Most importantly, the latter usually deals only with binary responses to a stimulus (perceived or not perceived), whereas EMG signals in neural stimulation in principle provide a continuous response strength and thus more subtle information. The fact that almost all existing methods in TMS seem to aggressively reduce the responses to such binary events could already be a hint of the role of psychophysics as an important inspirational source for the former. To consider this additional analog information of the response amplitude appropriately, stochastic root-finding methods that have been known in mathematics for a long time are adapted. We will present novel as well as established representatives of both classes and compare their performance in a realistic MEP model that describes their intricate statistical features.

Existing Methods in the Literature

The TMS threshold is usually defined as the lowest stimulation amplitude x—given as a percentage relative to the maximum stimulator output (MSO)—leading with a 50% probability to a peripheral EMG response V_{pp} of at least 50 μV peak-to-peak amplitude (Groppa et al., 2012; Rossini et al., 1994, 2015):

$$x_{\mathrm{th}} = \underset{x}{\operatorname{argmin}}\left(P\left(V_{\mathrm{pp}} \geq 50\ \mu\mathrm{V} \middle| x\right) = 50\%\right) \tag{1}$$

The value of 50 μV was chosen relatively arbitrarily as a low level that is not far from the noise limit of most EMG detection systems, and the procedures for detecting this point is still very rudimental in daily clinical application (Anderson and George, 2009; Mills and Nithi, 1997; Tranulis et al., 2006). The most common practice is estimating the relative-frequency of positive and negative responses (measured EMG voltage greater and lower than 50 μV, respectively) by appropriate stepping of the test amplitude to find the 50% probability level. At each test amplitude, a number of stimuli—usually up to ten—have to be administered accordingly, hence the name five-out-of-ten method. Such an approximation of the probability by a relative-frequency within just ten stimuli can be considered rather imprecise and a larger number of maximum twenty trials per stimulation amplitude is required to obtain more reproducible results for clinical and research purposes (Awiszus, 2012; Rossini et al., 2015). Another well-described method based on the relative response frequency was proposed by MILLS and NITHI (1997). Instead of detecting the threshold directly, it uses one point above threshold (*upper threshold*) and one below (*lower threshold*), and the threshold estimate itself is defined as the mean of these two points. Mathematically, this relates to a local linearization of the sigmoid curve of the neuronal response in dependence of the stimulation amplitude in the probability space. Similar to the relative-frequency method, both upper and lower thresholds are defined statistically, e.g., all ten responses suprathreshold and subthreshold. Both methods require a high number of stimuli and long duration to complete due to the necessary pause between two stimuli for preventing inter-pulse correlation (Nielsen, 1996). Furthermore, a high number of stimuli also bears a risk of unintended neuromodulation and side effects (Wassermann, 1998).

An alternative was introduced by AWISZUS (2003). The methodology is adapted from psychophysics (Pentland, 1980) and does no longer try to determine relative frequencies. Instead, a maximum-likelihood test is performed for a predefined model describing the parameter distribution of the input–output curve—most commonly of a cumulative Gaussian function with two parameters. The result of such parameter estimation by sequential testing (PEST) provides the most likely parameters, which include the threshold. An essential condition is a good selection

of the tested stimulus amplitudes (Strasburger and Treutwein, 1999). In contrast to frequency-based methods, each amplitude is usually just applied once.

These dominant procedures have been compared in several studies in the literature. A substantial test is reported by Mishory et al. (2004). The number of required stimuli was markedly lower for the model test (about a third) compared to a classical approach such as the MILLS–NITHI two-threshold method (Mills and Nithi, 1997). Tranulis et al. (2006) checked the MILLS–NITHI method against their own curve-fitting approach and reported huge numbers of required stimuli for both.

Despite the presented experimental performance of model estimation methods, the existing algorithms still seem to provide great potential for optimization. A key feature in this context might be the stepping rule, which chooses the next stimulation amplitude adaptively by taking into account the information from the previous observations. As later results will show and literature from psychophysics suggests (Strasburger and Treutwein, 1999), the stepping rule can notably influence performance. AWISZUS (2003) choses the maximum-likelihood estimate for the threshold to be the next stimulation amplitude. The threshold estimate is obtained from a reduced single-parameter model of the response and used a fixed nonstochastic relationship for the slope of the input-output curve at the threshold. This relationship is assigned based on some information from typical measurements (Awiszus, 2003), but does not allow any deviation from this slope–threshold regression line. The coefficient of 0.07 in this context is accordingly merely a fitting factor and only justified by an ensemble trend. If a subject does not meet this reduced model, the method may place all stimuli systematically too high or too low with respect to the intended stepping. Furthermore, the approach of using a threshold estimate of the last step as the amplitude for the next sample might be problematic in principle. If the likelihood distribution is not perfectly symmetric, states with extremely slow convergence are possible and the stepping approach is not able to increase the level of symmetry systematically. Instead of that, the peak of the likelihood distribution is shifted in tiny steps towards the correct parameter set and the spread is not reduced methodically. Convergence was observed to become very slow in these cases and can be completely overthrown by numerical issues.

A way for incorporating additional statistical information about typical results of the thresholding procedure itself in the same style as used here was proposed before (Goetz et al., 2011; Qi et al., 2011). Again, the method is adapted from another scientific field (Strasburger and Treutwein, 1999). Although the authors called it a BAYESIAN method, this name is not adopted in this text. It incorporates *a-posteriori* information into a classical FISHER maximum-likelihood estimation (MLE) using BAYES' theorem. This has been known in mathematics (Fisher, 1922, pp. 326) and applied fields, such as communication engineering (Proakis and Salehi, 2008), as maximum a-posteriori (MAP) for about a century. In this text, BAYESIAN is specifically used only for approaches that are consequently and completely premised on BAYESIAN statistics (Kontsevich and Tyler, 1999; Paninski, 2005; Wichmann and Hill, 2001). The BAYESIAN style particularly refers to the updating rule for the probability function after new observations, which should therefore follow the BAYESIAN learning paradigm. In contrast to MLE and MAP methods, which are point estimators, BAYESIAN statistics are furthermore intrinsically defined as grounded on a probabilistic distribution.

Methods

Classification of Thresholding Methods

The estimation of a neuronal property such as the excitation threshold can be performed in two fundamentally different ways. The model-based parameter estimators (MLE, MAP, and the BAYESIAN methods) require a probabilistic description of the input–output black box being analyzed in dependence of a certain parameter set. A

corrupt model will very likely deteriorate the quality of the estimate. The early approaches, in contrast, namely the aforementioned relative-frequency method based on recommendations of the International Federation of Clinical Neurophysiology (IFCN) (Rossini et al., 1994) or the MILLS–NITHI two-threshold method (Mills and Nithi, 1997), do not need such a predefined model description. They refer directly to the definition of the threshold. Several stochastic root-finding methods can also be counted among this class of non-parametric systems. In the following, the existing and several new or adapted algorithms are shortly described and compared for the application in TMS.

Nonparametric Threshold Detection

IFCN Relative-Frequency Method

The relative-frequency method was already roughly described. Despite aggressive criticism regarding its "mathematical validity" or "mathematical usefulness" (Awiszus, 2011, 2012, 2014b), this widely used procedure act as a benchmark for the more advanced methods. The procedure described by the IFCN still leaves freedom of interpretation. For the comparison, it is defined as follows to realistically represent the conditions in many laboratories:

(1) The starting amplitude is known from the hotspot search (see Simulation Routine for definition of hotspot), in which the corresponding neuron population being responsible for activating the observed muscle group (e.g., the *first dorsal interosseus*, FDI, or the *abductor pollicis brevis*, APB) is located with test stimuli on the primary motor cortex. The required amplitude adjustments in this step lead to an amplitude not too far above the desired threshold as focality would be counteracted at stronger stimuli.

(2) At a given stimulation amplitude, a maximum number (ten or twenty) of stimuli is applied until either positive responses of half the number (five or ten) or negative responses of more than half the number (six or eleven) have been observed.

(3) Being suprathreshold by definition in the situation of positive responses at least half the number of stimuli, the amplitude is stepped downward, typically by 1% or 2% MSO, and Step (2) is repeated.

(4) The latter outcome in Step (2) precludes 50% positive responses within the number of stimuli and the amplitude is subthreshold. The threshold itself is defined as the last level before this termination point.

Versions with the number of stimuli of ten and twenty and step size of 1% and 2% MSO are performed to show the effect of number of stimuli per step and step size on the accuracy and the number of stimuli required for the methods (Table 1).

MILLS–NITHI Two-Threshold Method

The original MILLS–NITHI method (Mills and Nithi, 1997) is carried out as follows:

(1) The subject is stimulated with pulses starting at an amplitude of 20% MSO and increasing in 10% MSO steps until the first positive response is identified.

(2) By decreasing the stimulation in 1% MSO steps, the lower threshold (LT) is designated as the maximum intensity at which 10 stimuli all produced no response.

(3) Starting from the lowest amplitude that had not resulted in a "no response" and increasing the stimulation in 1% MSO steps, the upper threshold (UT) is designated as the minimum intensity at which 10 stimuli all produced positive responses.

(4) The threshold is calculated as the mean of LT and UT.

In our study, a modified version was performed using the same starting amplitude as the IFCN method. The first negative response is identified by decreasing the amplitude in 10% steps and the LT and UT are then similarly obtained as in the original methods but in reversed order. Step sizes of 1% and 2% MSO were tested (Table 1).

Table 1 Summary of relative-frequency methods

Method	Variation	Step size (% MSO)
IFCN	Five-out-of-ten	1
		2
	Ten-out-of-twenty	1
		2
MILLS–NITHI	Original	1
		2
	Modified	1
		2

Stochastic Root-Finding Methods

The IFCN and MILLS–NITHI methods described above both require binary information reduction through quantization of the response (above versus below 50 µV). In contrast to such quantization, robust numeric root-finding algorithms that are still relatively unknown in TMS can directly act on analog data, e.g., the voltage amplitude of the EMG response. Preserving valuable *soft* information in the response, such root-finding approaches should enable an improvement of convergence. The here proposed sequential formulation is derived from the method of ROBBINS and MONRO (Lai, 2003; Robbins and Monro, 1951), with the neuronal variability playing the role of a noise source. The requirements for this approach are rather weak and comprise constant monotonicity on average (Lai, 2003).

The next stimulation amplitude x_{i+1} is defined relative to the last one using an iterative analog control sequence (ACS) that is intended to converge towards the threshold:

$$\begin{aligned} x_{i+1} &= x_i - a_i \cdot \Delta y_i \\ &= x_i - a_i \cdot (y_i - y_{\text{th}}) \quad . \end{aligned} \tag{2}$$

Here, $i \geq 1$ is the step number, a_i is the step size, x_i denotes the stimulation amplitude, and y_i and y_{th} are logarithms of the response ($y_i = \lg(V_i/[1\text{ V}])$) observed at step i and at the real threshold V_{th} (50 µV peak-to-peak by definition). The log-transformation takes into account the stochastic characteristics of EMG responses and the first approximation log-normal distribution of the response amplitude following an inductive stimulus in the motor cortex (Nielsen, 1996). The resulting distribution of y_i has reduced skewness and should be benign, optimally approximately GAUSSIAN, accordingly. Performing the whole procedure in the logarithmic space also grants the aforementioned requirement of constant monotonicity. The iteration for the EMG response thus becomes

$$x_{i+1} = x_i - a_i \lg(V_i/V_{\text{th}}) \quad . \tag{3}$$

Several versions were implemented for the sequence of step sizes a_i in order to compare the convergence properties (Table 2). Unlike classical mathematical problems, where almost sure (a.s.) convergence[1] is of importance, the TMS sequence has only a very limited number of steps. Thus, quick convergence seems more attractive here rather than any statement for the case of infinite duration, although a.s. convergence is covered by a certain class of control sequences (Williams, 1991). In the basic version, the control sequence followed a harmonic series

$$a_i = a_0/i \quad \text{for } i \geq 1. \tag{4}$$

[1] Defined as convergence with probability one: $x_i \xrightarrow{\text{a.s.}} \tilde{x}: \forall \epsilon > 0, \exists\, I, \delta: P(|x_i - \tilde{x}| > \delta) < \epsilon, \forall\, i > I$.

Following the style of Delyon and Juditsky (1993), adaptive methods are utilized in which the control sequence is adjusted at step i ($i > 1$) only when the response changes from subthreshold to suprathreshold or vice versa from the previous step $(i-1)$, i.e., following a sign change in Δy_i: $\mathrm{sgn}(\Delta y_i) \neq \mathrm{sgn}(\Delta y_{i-1})$. For a.s. harmonic convergence in ACS1, the adaptive control sequence is

$$a_i = a_0/(c_i + 1) \quad \text{for } i \geq 1, \tag{5}$$

where c_i is the number of sign changes of Δy_i up to step i,

$$c_i = \begin{cases} c_{i-1} & \mathrm{sgn}(\Delta y_i) = \mathrm{sgn}(\Delta y_{i-1}) \\ c_{i-1} + 1 & \mathrm{sgn}(\Delta y_i) \neq \mathrm{sgn}(\Delta y_{i-1}) \end{cases} \quad \text{for } i > 1, \tag{6}$$

with $c_1 \triangleq 0$. To test faster geometric convergence of $2^{-i/2}$ and 2^{-i}, ACS3 and ACS5 use control sequence of

$$a_i = a_0/b^{c_i} \quad , \tag{7}$$

with b as $\sqrt{2}$ and 2 to, respectively.

Further variations of these methods were designed to examine the convergence rate and performance. To take advantage of both fast convergence using larger b (e.g., ACS3) and the a.s. convergence properties of ACS1, a hybrid method ACS3–1 was designed to switch the stepping rule from ACS3 to ACS1 when the threshold estimate is close to the true threshold and the expected change in x_i ($|\Delta x_i| = |a_i \cdot \Delta y_i|$) is small, e.g., less than 1.5% MSO. Second-order versions control sequences were also implemented, in which the next stimulation amplitude is defined using information from previous two steps

$$x_{i+1} = x_i - a_i \cdot (\Delta y_i + w \cdot \Delta y_{i-1}) \quad \text{for } i > 1, \tag{8}$$

with w being the weight for the second-order term. Quantized versions, i.e., digital control sequences (DCS), were implemented for comparison as well. The analog value Δy_i was reduced to a modified sign function $\Delta y_i = \mathrm{sgn}(y_{\mathrm{th}} - y_i) \in \{-1, +1\}$ for the two different cases of below threshold or above (or equal) threshold.

The definition of appropriate initial a_0 is assisted by the iteration (2) itself, which bears the character of a slope (Nemirovski et al., 2009) after rearrangement per

$$a_i^{-1} = \frac{y_{\mathrm{th}} - y_i}{x_{i+1} - x_i} \quad . \tag{9}$$

For targeting the threshold quickly, a_0 should accordingly be not too far from the cotangent of the line connecting the present stimulation–response (x_i, y_i) towards the threshold $(x_{\mathrm{th}}, y_{\mathrm{th}})$. Larger values increase the speed of approach of x_i towards x_{th}, but may cause the stepping to overshoot and oscillate around the target, whereas smaller numbers promote lengthy one-sided convergence. Statistical information about typical and maximum slopes can fill this gap here. For the numerical comparison here, the default initial step size was chosen such that it approximates typical cotangents around the inflection point of the recruitment curve (see Response Model) accordingly:

$$a_0 = \frac{0.2}{\lg(10^{-2}) - \lg(10^{-5})} \approx 0.067 \quad . \tag{10}$$

Here $\lg(10^{-5})$ (10 μV) and $\lg(10^{-2})$ (10 mV) are approximate response amplitudes near the EMG noise floor and saturation level at the lower and higher ends of the input-out curve respectively, and 0.2 (20% MSO) is a sufficient change in stimulus strength to cover the transition of the sigmoidal response curve. Initial step size from one tenth to ten times the default values evenly sampled on a logarithmic scale were also tested. In second-order sequences, the weight of the second-order term also affects convergence. Larger/positive w speed up the approach to threshold, however, could generate oscillations depending on the value of a_0; smaller/negative w reduces the speed of convergence, but could help to mitigate the oscillation if a_0 is chosen too large. A default value of −0.1 is used and values between −1 and 1 were tested for ACS4.

Table 2 Summary of root-finding methods

Version	Control sequence	Order	Convergence of a_i
ACS1-N	Analog	1	Non-adaptive, i^{-1}
ACS2-N	Analog	2	
DCS1-N	Digital	1	
ACS1	Analog	1	Adaptive, i^{-1}
ASC2	Analog	2	
DCS1	Digital	1	
ACS3	Analog	1	Adaptive, $2^{-i/2}$
ASC4	Analog	2	
DCS3	Digital	1	
ACS3–1	Analog	1	Adaptive, $2^{-i/2}$ switching§ to i^{-1}
ACS5	Analog	1	Adaptive, 2^{-i}
ASC6	Analog	2	
DCS5	Digital	1	

§ Transition when the maximum change of x_i—$|\Delta x_i| = |a_i \cdot \Delta y_i|$–is expected to be less than 1.5% MSO.

Parametric Estimation

Basic Maximum-Likelihood Threshold Hunting

The implementation of the basic MLE method followed substantially the suggestions and definitions by AWISZUS (2003). The probability space is described by a GAUSSIAN model with two degrees of freedom—the threshold $t = x_{\mathrm{th}}$ and the slope s at threshold:

$$p(r_i = 1|x_i, t, s) = \Phi\left(\frac{x_i - t}{s}\right) = \frac{1}{s\sqrt{2\pi}} \int_{-\infty}^{x_i} \exp\left(-\frac{(\xi - t)^2}{2s^2}\right) \mathrm{d}\xi \quad . \tag{11}$$

Here, r_i denotes the binary response of a single stimulus x_i:

$$r_i = \begin{cases} 1 & V_i \geq 50\ \mu V \\ 0 & V_i < 50\ \mu V \end{cases} \quad .$$

The likelihood function is defined for independent events

$$p(\mathbf{r}|\mathbf{x}, t, s) = \prod_{\{i|r_i=1\}} p(r_i = 1|x_i, t, s) \prod_{\{i|r_i=0\}} \left(1 - p(r_i = 0|x_i, t, s)\right) \quad , \tag{12}$$

where $\mathbf{x}$ is the list of stimuli and $\mathbf{r}$ is the list of binary responses. According to the classical FISHER approach, the maximum of this expression for the arguments t and s provides the most likely parameter set as generally known ($\mathrm{argmax}_{(t,s)}\, p(\mathbf{r}|\mathbf{x}, t, s)$). Due to the strict monotonicity of the logarithm, the product can be split into an equivalent sum, which forms the log-likelihood. The MLE of t provides both the threshold estimate at the current step and stimulation amplitude at the next step. AWISZUS simplified the estimation (12) by using only the threshold parameter t, whereas the slope parameter was fixed to the value $s = 0.07 \cdot t$ (Awiszus, 2003; Julkunen, 2019). We replicated AWISZUS's method and also performed a full MLE with both parameters (Table 3). When the likelihood function has a plateau, i.e., a range of t or (t, s) combinations resulting in the same maximum, the mean of the argument t over the range or (t, s) combinations is chosen as the next stimulation amplitude (Awiszus, 2003). For computational purposes, the parameter space was discretized with $\Delta t = 0.001$ for $t \in [0, 1]$, and 400 logarithmical-distributed samples were chosen for $s \in [0.001, 0.5]$.

Advanced Estimation and Stepping

Both the stimulation and the stepping rule of the basic MLE method can be improved (Table 3). A modified estimation method was designed based on a MAP-like approach and inverses the likelihood ratio from (12). This was also suggested by Qi et al. (2011). From the perspective of the causality of the model, the likelihood function is directed backwards. For incorporating statistical information, this was modified using BAYES' theorem. This feature was implemented in order to improve the performance for common cases per

$$p(\boldsymbol{\theta}|\mathbf{x},\mathbf{r}) = \frac{p(\boldsymbol{\theta})}{p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x})} \cdot p(\mathbf{r}|\mathbf{x},\boldsymbol{\theta}) \quad , \tag{13}$$

in which the vector $\boldsymbol{\theta} = (t, s)$ denotes the parameter set. The additional information about the distribution of typical parameters $p(\boldsymbol{\theta})$ has to be provided in advance (see Response Model section). Despite that, this is not *a-priori* information from an epistemological point of view. It can be collected using the algorithm itself. The last factor $p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x})$ can be evaluated by normalization

$$p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}) = \int_{\mathbb{R}_+^n} p(\mathbf{r}|\mathbf{x}) p(\mathbf{r}|\mathbf{x},\boldsymbol{\theta}) \mathrm{d}^{\mathrm{n}}\theta \quad , \tag{14}$$

and provides a measure for the consistency or matching of *a-posteriori* statistics with the observations.

The rule for choosing the stimulus amplitude for the next step is modified to reduce the deviation of the estimation systematically. Taking the amplitude that maximizes the likelihood function may be disadvantageous for sampling and could result in strong biasing due to (randomly occurring) inappropriate sampling in the beginning. And in the context of prior information, such approach tends to cluster sampling even if the estimate is still far from the targeted value. Inspired by the stepping method by KONTSEVICH and TYLER (1999), the next stimulus is placed at the amplitude that is expected to lead to the smallest second-order moment of $p(\boldsymbol{\theta}|\mathbf{r})$ in the next step. In principle, this method also allows prioritizing a certain parameter. For the task of threshold detection, only $t = \theta^{(1)}$ is of special interest. The appropriate formula reads

$$x_{i+1} = \underset{x_{i+1}}{\operatorname{argmin}}\, \mathbf{E}\left(\int_{\mathbb{R}_+^n} \left(\theta^{(1)} - \tilde{\theta}^{(1)}\right)^2 p(\boldsymbol{\theta}|\mathbf{r},\mathbf{x},x_{i+1})\, \mathrm{d}^n\theta \right) \quad , \tag{15}$$

where $\tilde{\theta}^{(1)}$ can either be the estimated threshold t_{est} or the expected mean $\bar{\theta}^{(1)} = \mathbf{E}\left(\int_{\mathbb{R}_+^n} \theta^{(1)} p(\boldsymbol{\theta}|\mathbf{r},\mathbf{x},x_{i+1})\, \mathrm{d}^n\theta\right)$, and $n = 2$ is the number of parameters. The former expression resembles the inertia of the likelihood function, whereas the latter is the variance. Accordingly, these properties are used to reference these methods later (Table 3). Herein, the accuracy of the less important slope parameter $s = \theta^{(2)}$ is also improved from step to step, but only to such an extent that it does not limit the estimation accuracy of the threshold. The integration can be numerically challenging in some cases, e.g., for steep peaks. Therefore it is highly recommended to implement a divergence detection, for instance, the here taken difference to the maximum of the likelihood function, and either re-calculate the integral or use the maximum instead.

KONTSEVICH–TYLER BAYESIAN Estimation

Strictly BAYESIAN methods differ from point estimators and are unknown in the field of TMS. The here chosen embodiment was adapted from KONTSEVICH and TYLER (1999). All central elements were taken as described, and only the estimated model was modified for consistency and was the GAUSSIAN black box $p(r|\boldsymbol{\theta},x)$ as in the other two estimators, i.e., (11). Based on this, new information is embedded into the probability density function $p(r|x,\boldsymbol{\theta})$ at every step i using BAYES' theorem following

$$p_i(\boldsymbol{\theta}|x,r) = \frac{p_i(\boldsymbol{\theta})p(r|x,\boldsymbol{\theta})}{\int_{\mathbb{R}_+^n} p_i(\boldsymbol{\theta})p(r|x,\boldsymbol{\theta})\mathrm{d}^n\theta} \quad . \tag{16}$$

The next stimulation amplitude x_{i+1} is determined using an information-theoretical approach which minimizes the expected entropy at the next step (Kontsevich and Tyler, 1999)

$$x_{i+1} = \underset{x}{\operatorname{argmin}}\, \mathbf{E}(H_i(x,r)\,) \quad , \tag{17}$$

in which the entropy $H_i(x,r)$ is defined as

$$H_i(x,r) = -\int_{\mathbb{R}_+^n} p_i(\boldsymbol{\theta}|x,r)\log\big(p_i(\boldsymbol{\theta}|x,r)\big)\,\mathrm{d}^n\theta \quad , \tag{18}$$

and the expectation is as sum of the entropy $H_i(x,r)$ weighted by probability $p_i(r|x)$ of sub- and supra-threshold responses

$$p_i(r|x) = \int_{\mathbb{R}_+^n} p_i(r|x,\boldsymbol{\theta})p_i(\boldsymbol{\theta})\mathrm{d}^n\theta \quad r = 0{,}1. \tag{19}$$

Each stimulus–response pair narrows down the posterior probability density distribution to:

$$p_{i+1}(\boldsymbol{\theta}) = p_i(\boldsymbol{\theta}|x_{i+1},r_{i+1}) \quad . \tag{20}$$

The first moment of the probability density function provides the estimates for the parameters at each step per

$$\theta_i^{(l)} = \int_{\mathbb{R}_+^n} \theta^{(l)} p_i(\boldsymbol{\theta})\mathrm{d}^n\theta \quad , \tag{21}$$

where l is the parameter index (e.g., $l = 1{,}2$ for $\boldsymbol{\theta} = (t,s)$). In contrast to the point estimators methods, the probably density functions (11) and (16) are explicitly computed and stored for the whole parameter space: $r \in \{0{,}1\}$, $x \in [0,1]$, $t \in [0,1]$, and $s \in [0.001, 0.5]$. For computational purposes, the parameter space was discretized with $\Delta x = 0.01$ and $\Delta t = 0.0025$, and 250 logarithmical-distributed samples were chosen for s, corresponding to well over ten million points. The initial distribution of $p_0(\boldsymbol{\theta})$ was chosen as the population distribution $p(t,s)$ (see Response Model).

Table 3 Summary of parametric methods

Method	Threshold estimation method	Stepping method
MLE-1, MLE-1 (AWISZUS)	Single-parameter MLE	Single-parameter MLE
MAP-1, MLE-1	Single-parameter MAP	Single-parameter MLE
MLE-2, MLE-2	Two-parameter MLE	Two-parameter MLE
MAP-2, MLE-2	Two-parameter MAP	Two-parameter MLE
MAP-2, min. inert.	Two-parameter MAP	Minimum inertia
MAP-2, min. var.	Two-parameter MAP	Minimum variance
BAYESIAN	Expectation of threshold	Minimum expected entropy

Test Bed

Response Model

We implemented a realistic MEP model (Goetz et al., 2019) which includes previously found variability patterns such as amplifier noise, saturation, and neural variability behavior (Goetz et al., 2014, 2022; Goetz and Peterchev, 2012). The model's increased complexity compared to other numerical studies concerning threshold-estimation methods (Awiszus, 2003; Borckardt et al., 2006) was intended to avoid using the very same models

underlying the parameter estimation methods for the test procedure, which would risk numerical favoritism for those particular methods. The response model provides analog EMG amplitude values based on a stochastic recruitment curve (Figure 1A), with parameters mimicking real measurements performed at Columbia University (Peterchev et al., 2013). The analog amplitude response, i.e., the peak-to-peak MEP amplitude V_{PP} in units of Volt, was defined as

$$V_{\mathrm{PP}}(x) = \epsilon_{\mathrm{yadd},ij}\left(p_{1,j}, p_{8,j}, p_{9,j}\right) + 10^{\epsilon_{\mathrm{ymult},ij}(p_{6,j})} \cdot 10^{p_{2,j}\Big/\left(1+\frac{p_{3,j}}{\left(x-p_{5,j}+\epsilon_{\mathrm{xadd},ij}(p_{7,j})\right)^{p_{4,j}}}\right)} , \tag{22}$$

where the $x \in [0,1]$ is the stimulation amplitude in percent MSO, $p_{1,j}$ to $p_{9,j}$ are the individual response parameters for subject j obtained from an experimental population (Goetz et al., 2014); intra-individual variability sources $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$, and the additive noise $\epsilon_{\mathrm{yadd},ij}$ vary between subjects j and for pulses i along distributions with subject-dependent parameters. The distributions of recruitment parameters, intra-individual variability, and noise are given in the Appendix.

Model parameters were generated for a total of 25,000 virtual subjects. The *real* thresholds for each subject was determined from the corresponding *noiseless* model (see Appendix) by applying root finding methods to the response curve to identify the stimulation that generate a 50 µV response (Figure 1B). By applying MLE using 10,000 simulation–response pairs to each subject, the "real" parameters t and s were obtained and analyzed (Figure S1). The distribution of thresholds and slopes $p(t, s)$ can be fitted with a normal and lognormal distribution, respectively, with details given in the Appendix.

Simulation Routine

The test bed comprised all threshold detection algorithms with the above given parameter definitions, and the response model. In order to meet the real conditions of a threshold search, each algorithm was given information from the hotspot detection procedure, i.e., an amplitude level which was known to be surely above the threshold. To avoid ambiguity and circular definition between threshold and hotspot (Awiszus, 2014a), the latter was chosen as the stimulation level that generated a 500 µV response in the noiseless model, and was rounded to 1% MSO for the algorithms that use specific integer amplitudes (i.e., IFCN and MILLS–NITHI algorithms). The algorithms then use this amplitude as the starting point.

The responses for subsequent stimuli were provided by the full MEP model to the algorithms. Whereas the analog versions of the ROBBINS–MONRO-like algorithms directly process this soft information, all other methods quantized the response into the binary representation of above and below threshold first. The stochastic root-finding methods and parametric estimation algorithms were tested with a maximum of 200 steps each. The absolute and relative errors for each threshold estimate were analyzed with

$$\begin{aligned} \Delta x_{\mathrm{th}} &= x_{\mathrm{th,est}} - x_{\mathrm{th,real}} \\ \delta x_{\mathrm{th}} &= \Delta x_{\mathrm{th}} / x_{\mathrm{th,real}} \end{aligned} . \tag{23}$$

For comparison with previous studies in the literature, the absolute values of the errors $|\Delta x_{\mathrm{th}}|$ and $|\delta x_{\mathrm{th}}|$ were also analyzed, the latter of which is previously referred to as the relative error (Awiszus, 2003; Qi et al., 2011). The results mostly presents the relative errors δx_{th}, as they provide normalized distance of the threshold estimate from the true value expressed as percentage of the latter. The absolute error Δx_{th} overall have similar distributions and are shown in the supplement.

Computation and data analysis was performed with MATLAB (versions 2019a and 2020a, The Mathworks, Natick, MA, USA) on the Duke Compute Cluster, which runs Red Hat Enterprise Linux 8 on Cisco Systems UCS Blade Servers. The code is available online (DOI: 10.5281/zenodo.6483601, (Wang, 2022)).

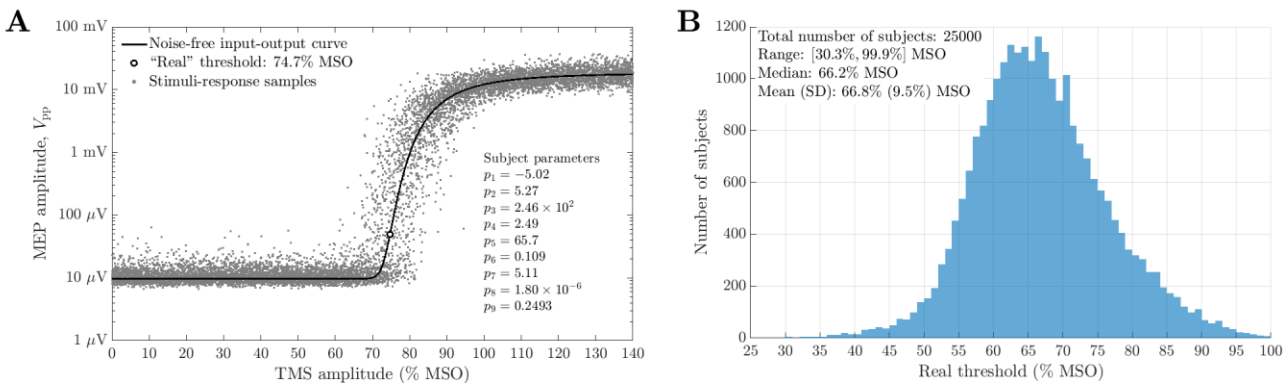

Figure 1: TMS response model.

A. An example of the TMS recruiting behavior that is mimicked by the model. This virtual stimulator provides analog values, which some of the algorithms use for a faster convergence. The used equations are independent from the tested black-box models in the parameter estimators in order to avoid model-determined advantages. Subsequent responses are not correlated. **B**. The real thresholds obtained from the noiseless model of the 25,000 virtual subjects shown in a histogram with bin width of 1% MSO.

Results

IFCN Relative-Frequency Method and MILLS–NITHI Two-Threshold Method

For the IFCN ten-out-of-twenty method with 1% MSO stepping size and a starting point of 500 µV (Figure 2A$_i$), the number of stimuli until the method terminates and reports a threshold value for a given subject had a wide distribution, ranging from as few as around ten to twenty steps (in which a single step below the starting amplitude was already considered subthreshold by the method, terminating the procedure practically immediately after the start) to over 400 steps. The number of stimuli had a median of 133 and mean and SD of 141.8±75.7 (Figure S3A). The relative errors δx_{th} overall had a relatively small bias (median 0.82%, mean and SD 0.92%±1.74%), skewing slightly towards the positive end due to more positive outliers (Figure S3B$_i$). Estimates obtained with fewer stimuli (within 40 stimuli) especially tended to overestimate the threshold due to early false-positive fulfillment of the termination. As the number of stimuli increased, the error distribution became more symmetric, becoming relatively stable between 100 and 250 stimuli, and only narrowed for thresholds obtained with over 250 steps. Increasing the stepping size to 2% MSO as widely done in practice (Figure 2A$_{ii}$) decreased the number of stimuli significantly (median 80, mean and SD 84.3±38.3, $p=6\times10^{-8}$, one-way analysis of variance, Table S1), on average requiring 57.5±42.6 fewer stimuli (median 54 fewer) and over 200 stimuli for very few subjects (<0.5%). The threshold accuracy as measured by the relative error (median 1.13%, mean and SD 1.23%±2.10%) was significantly ($p=6\times10^{-8}$) increased by 0.31±2.71%. The IFCN five-out-of-ten method with 1% MSO stepping size (Figure 2B$_i$) requires much fewer steps (median 65, mean and SD 69.4±38.0) compared to the ten-out-of-twenty method using either 1% or 2% MSO step size ($p=6\times10^{-8}$). The accuracy (median 0.92%, mean and SD 1.04%±2.22%) is significantly different from either ten-out-of-twenty method. And similarly, increasing the stepping size to 2% MSO (Figure 2B$_{ii}$) decreased the number of stimuli significantly (median 40, mean and SD 41.9±19.4, $p=6\times10^{-8}$), on average requiring 27.5±22.4 fewer stimuli (median 26 fewer). The relative error (median 1.04%, mean and SD 1.16%±2.60%) was also significantly increased (by mean and SD 0.13%±3.39%). The statistical significance for

errors, however, has little practical meaning given that they had similar ranges for both ten-out-of-twenty or five-out-of-ten methods and either step sizes.

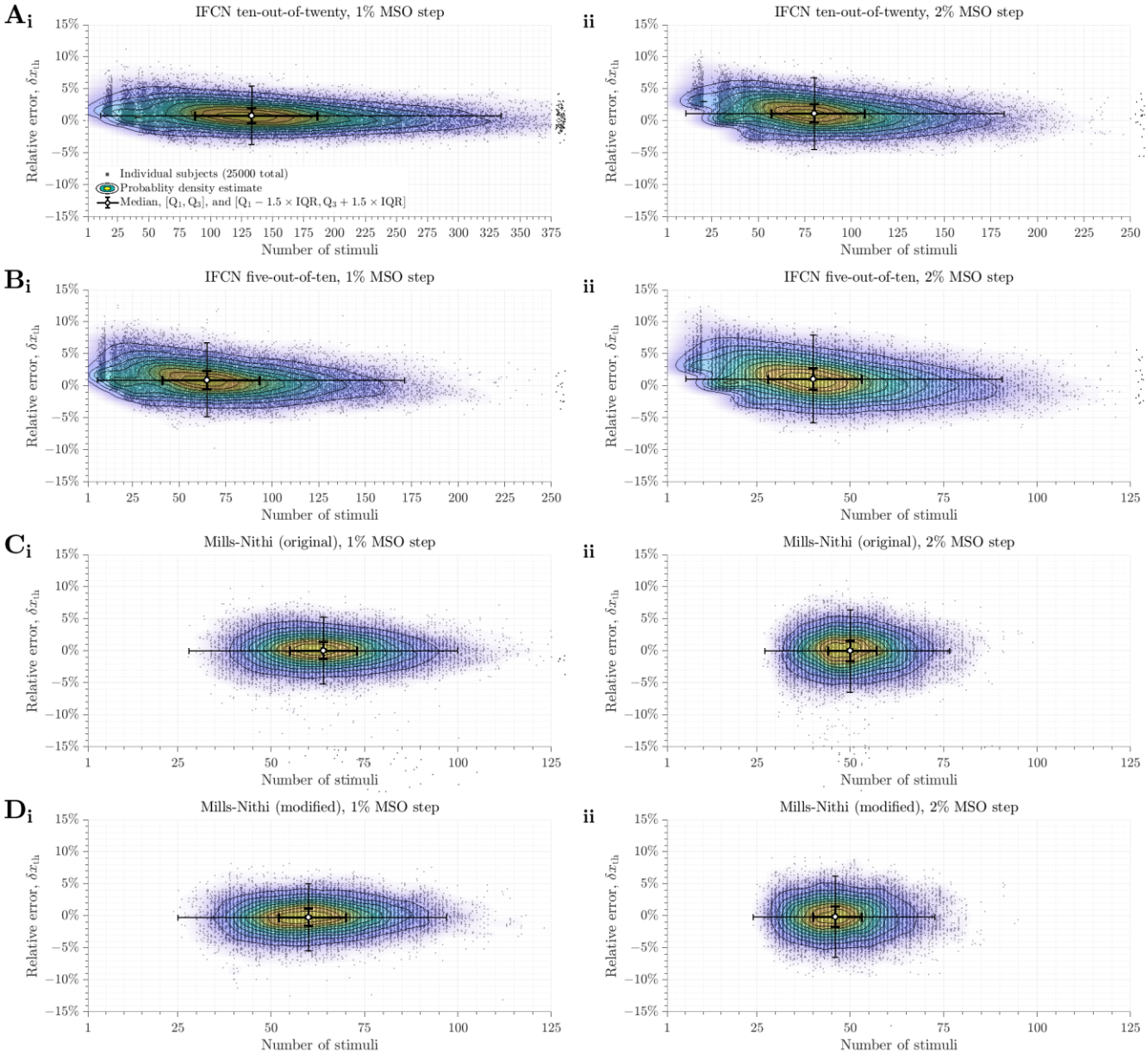

Figure 2: Distribution of relative threshold errors of the IFNC and MILLS–NITHI methods

The relative errors are plotted against the number of stimuli required for all the virtual subjects as dots. Several statistics are shown, with median as white circle, interquartile range (IQR) between first and third quartiles (Q_1 & Q_3) as thick error bars, and ranges extending to the lower and upper adjacent values—which are 1.5 times IQR below Q_1 and above Q_3 and cover approximately ± 2.7 standard deviations for normally-distributed data—as thin error bars. A probability density estimate based on a normal kernel function is also shown with the contour lines. **A.** IFNC ten-out-of-twenty method. **i.** 1% MSO step size. **ii.** 2% MSO step size. **B.** IFNC five-out-of-ten method. **i.** 1% MSO step size. **ii.** 2% MSO step size. **C.** Original MILLS–NITHI methods. **i.** 1% MSO step size. **ii.** 2% MSO step size. **D.** Modified MILLS–NITHI methods. **i.** 1% MSO step size. **ii.** 2% MSO step size. Note different horizontal scale in panels A_i, A_{ii}, and B_i versus the others.

The original MILLS–NITHI method with 1% MSO step size (Figure 2C_i) had a narrow distribution of number of stimuli (median 64, mean and SD 64.7±13.2), requiring at least 25 stimuli but rarely exceeding 100. Compared to the IFCN five-out-of-ten method with the same 1% MSO step size, the decrease in number of stimuli (by 4.7±39.2) is statistically significant, despite the significant overlap and a decrease of 1 only for the median. However, for any individual subject, the IFCN method could be much faster if the starting point for the latter (500 μV, which is typically around 115% threshold) is already close to the threshold, whereas the MILLS–NITHI method does not utilize the information obtained from the hotspot search. On the flipside, the MILLS–NITHI method does not result in extremely large numbers of stimuli. The errors were also centered close to zero (median 0.08%, mean and SD −0.01%±2.47%) with very small though statistically significant difference compared to the IFCN method (−0.91%±2.99%, $p=6\times10^{-8}$), but had a much wider distribution and more outliers. Increasing the step size to 2% MSO (Figure 2C_{ii}) reduced the number of steps and narrowed its distribution (median 50, mean and SD 50.6±9.4, a decrease with mean and SD of 14.1±16.0 and median of 14, $p=6\times10^{-8}$). Again, the accuracy (median 0.02%, mean and SD −0.08%±2.76%) was only slightly affected (−0.08%±3.63%, $p=8\times10^{-4}$). The modified MILLS–NITHI method (Figure 2D_i) utilized the same starting point as the IFCN methods. However, this additional information only modestly reduced the number of stimuli (median 60, mean and SD 61.2±13.2) on average by 3.5±18.6 steps (median 3) compared to the original method with the same 1% MSO step size. The modified method biased the errors towards negative values (median −0.19%, mean and SD −0.21%±2.03%), resulting in a small but significant change compared to the original method (a change of −0.22%±3.17%) and fewer outliers. Increasing the step size to 2% MSO (Figure 2D_{ii}) again reduced the number of steps and narrowed its distribution (median 46, mean and SD 46.6±9.3, a decrease with mean and SD of 14.3±15.8 and median of 14, $p=6\times10^{-8}$) but did not affect the accuracy (median −0.11%, mean and SD −0.11%±2.37%, change of 0.09±3.1%, p=1).

The distributions of absolute errors were similar to those the relative errors and are shown in Figures S2. Full statistics of the methods are given in Figure S3 and Table S1

Parametric Estimation

Unlike the IFCN and MILLS–NITHI methods, the parametric estimation methods generate a threshold estimate and thus an error at every step (relative error δx_{th} in Figure 3 and absolute error Δx_{th} in Figure S4). Full statistics are visualized every tenth step, showing the median, interquartile range (IQR), lower and upper adjacent values, and outliers outside the lower and upper adjacent values (Figure 3A). The lower and upper adjacent values are 1.5×IQR below and above the first and third quartiles (Q_1 and Q_3), respectively, corresponding to approximately ±2.7×SD or 99.3% coverage if the data is normally distributed. The median $|\delta x_{\mathrm{th}}|$ and the IQR of δx_{th} (Figure 3B, open and filled markers) characterize how close the errors approach zero and how wide the errors are distributed. The IQR of δx_{th}, and therefore the range between the lower and upper adjacent values as well, correlates well with the median $|\delta x_{\mathrm{th}}|$: the ratio between these two statistics after a few steps is approximately 2 and, except for the one-parameter stimulation methods, is very stable (Figure S5). Therefore, the median $|\delta x_{\mathrm{th}}|$ captures the distribution of the majority of errors, with the remaining outliers summarized as a percentage of the total number of subjects (Figure 3C). The IFCN five-out-of-ten method with 2% MSO step size, which obtains a threshold estimate with a median 40 steps and median $|\delta x_{\mathrm{th}}|$ is 1.85%, is used as reference.

The basic MLE method (Awiszus, 2003) which uses a single-parameter MLE for threshold estimation and stepping (Figure 3, dark blue lines) reaches the same error level as the IFCN reference with only 26 steps and the median $|\delta x_{\mathrm{th}}|$ at 40 steps is only two thirds that of the IFCN method (Table 4). The threshold estimate steadily improves with steps for non-outliers. For the large number of outliers (about 15%), however, the threshold estimate has a large negative error that does not improve at all. Because the basic MLE method just tests model parameters

(t, s) directly on the empirical line of $s = 0.07 \cdot t$ and fails to explore the full parameter space, it can perform poorly for the many subjects in our model that do not follow such a 0.07 relationship. Using a one-parameter MAP estimator for the threshold (Figure 3, light blue lines) does not improve the performance.

Table 4 Error convergence of parametric estimation methods compared to the reference IFCN five-out-of-ten method with 2%MSO step size

Method	Step reaching median $\lvert\delta x_{\mathrm{th}}\rvert$ of 1.85%	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 40
MLE-1, MLE-1	26	1.30%
MAP-1, MLE-1	26	1.29%
MLE-2, MLE-2	48	2.00%
MAP-2, MLE-2	18	1.10%
MAP-2, min. inert.	17	1.22%
MAP-2, min. var.	17	1.50%
BAYESIAN	13	1.02%

The two-parameter MLE methods (Figure 3, dark red lines) has a median $\lvert\delta x_{\mathrm{th}}\rvert$ slightly larger than the IFCN reference at 40 steps and reaches the error level of the latter slower (48 steps). This reduced speed of convergence reflects the exploration of the full parameter space: the likelihood function initially has a large plateau region that is absolutely flat, which reduces in size as more stimulation–response pairs are recorded. Only after a sufficient number of steps does the likelihood plateau narrow to generate a clear optimal estimate. As the slope parameter is no longer fixed to the threshold, the two-parameter MLE methods can match the wide range of behaviors of the subjects in our model and improve the threshold estimate for all subjects with very few outliers (<1%). Using a two-parameter MAP estimator for the threshold (Figure 3, orange lines) significantly improves the performance, reaching the error level of the IFCN method with only 18 steps. The prior information significantly influences the likelihood landscape, rendering one parameter pair (t, s) in the otherwise flat maximum plateau unique. The slightly increased percentage of error outliers is not a decrease in performance but rather reflects the improved and much tighter error distribution.

The two methods that used minimum-inertia and minimum-variance stepping and MAP threshold estimator (Figure 3, yellow and purple lines, respectively) both reach the median $\lvert\delta x_{\mathrm{th}}\rvert$ of the IFCN five-out-of-ten method with 2% MSO step size very fast at step 17, and their early convergence up to 20 steps is slighter better than that of the two-parameter MLE stepping with MAP estimator (Figure 3B, inset). Afterwards, the minimum-inertia method slows slightly and becomes comparable to the one-parameter MLE by AWISZUS after 40 steps. The minimum-variance method, however, plateaus after 40 steps at a relative high error level (although still lower than the IFCN reference). The marginal early advantage of these two methods is outweighed by their high computational cost to minimize the second-order moment (15).

Among all parametric stimulation methods, the KONTSEVICH–TYLER BAYESIAN method (Figure 3, green lines) reaches the IFCN method's median $\lvert\delta x_{\mathrm{th}}\rvert$ the fastest at step 13 and has the lowest error up to step 50. After that, the two-parameter MLE stepping with MAP estimator catches up and they have comparable performance. The BAYESIAN method achieves such good estimation by considering the entire function space of the probability density functions. Whereas MLE and MAP estimators evaluate the underlying likelihood function in every step anew based on the same key equation, the BAYESIAN approach adds information into the explicitly evaluated and stored distribution from the last iteration. Besides being the most computationally costly methods, the BAYESIAN method

is also highly sensitivity to rounding-error propagation and severely influenced by the sampling of the parameter space. This may have led to an unfair disadvantage in this comparison.

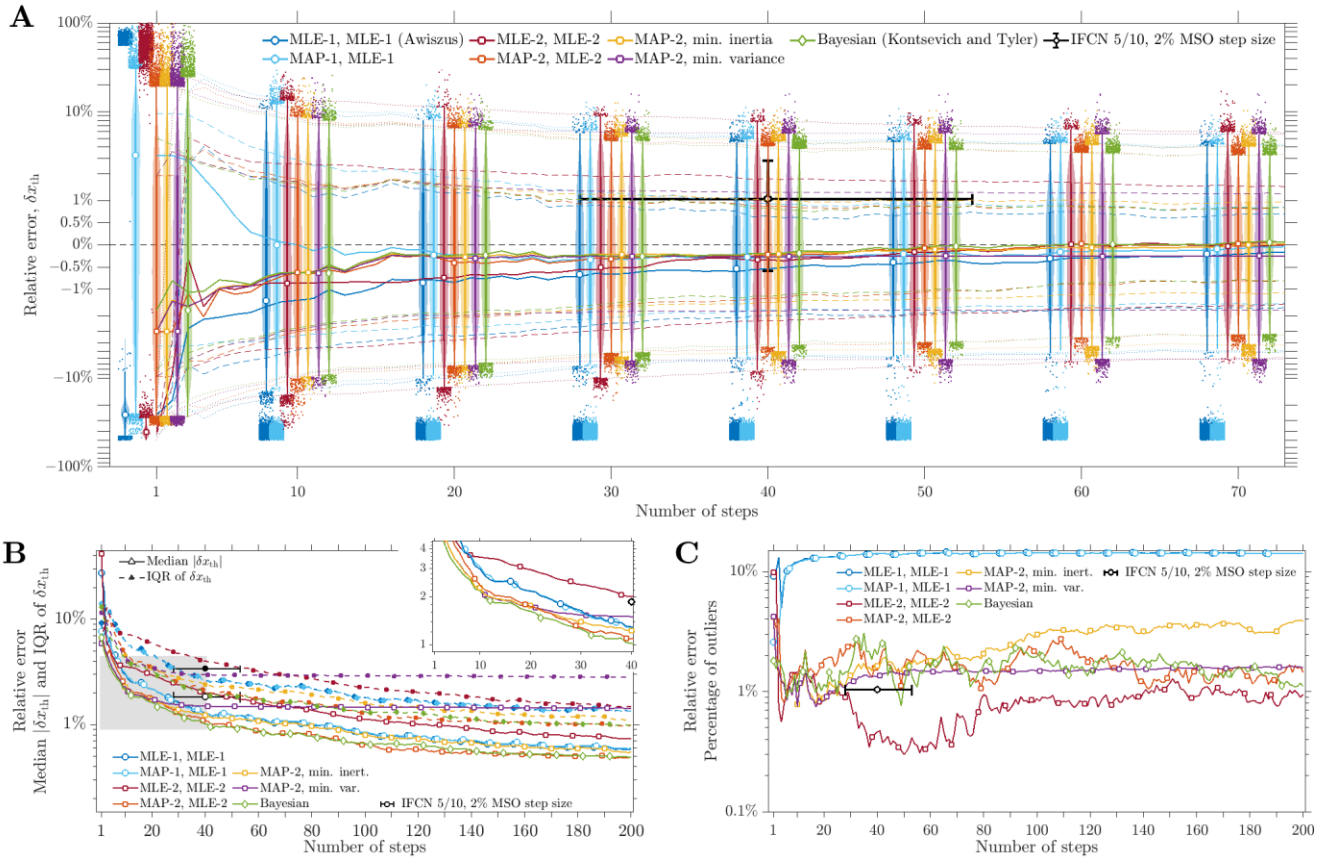

Figure 3: Relative threshold errors of the parametric estimation methods

A. Statistics are shown up to the seventieth steps as boxes-and-whiskers at the first step and every ten steps starting from the tenth step, with median as marker, interquartile range (IQR) as thick vertical line, upper and lower adjacent values as whiskers (thin vertical lines) extending 1.5×IQR below and above the first and third quartiles (Q_1 and Q_3), and outliers as dots, and as lines for every step with median as thick lines, Q_1 and Q_3 as dashed lines, and whiskers as dotted line. The log-lin-log mixed scale is used to better visualize the errors around zero, with linear scaling in the range of [−1%, 1%] and logarithmic scaling of amplitudes outside this range. The median and IQR of the IFCN five-out-of-ten method with 2%MSO step size are shown as black error bars as in Figure 2B for reference. **B.** The median of the absolute value of errors and the IQR of the errors (length of whiskers in panel A) on logarithmic scale. The inset shows the details in the first 20 steps (the gray area in the main panel). **C.** The percentage of outliers (dots in panel A) outside the upper and lower adjacent values (whiskers in panel A).

Stochastic Root-Finding Methods

The parameter-free stochastic root-finding methods are based on the threshold definition only and therefore have the important advantage over the parametric estimation approaches, which depend on a falsifiable or inaccurate model. Like the parametric methods, the root-finding methods also generate a threshold estimate and an error at every step (relative error δx_{th} in Figures 4, 5, S11A, S12A, and S13A, and absolute error Δx_{th} in Figures S6, S9, S11B, S12B, and S13B). All root-finding methods are provided with the same above-threshold amplitude from the

hotspot search and thus have the same distribution at the zeroth step (gray elements in Panels A or A_i and B_i of Figures 4, 5, S6, S9, and S11–S13). The ratio between the IQR of δx_{th} and the median $|\delta x_{\mathrm{th}}|$ (Figures S7 and S10) is also centered on 2 after a few steps, but shows larger variations compared to those of parametric estimation methods.

The distributions of all first-order analog methods except for ACS1-N become negatively biased within ten steps, i.e., the median is less than zero, the absolute values of the first quartile and lower adjacent value are larger than the third quartile and upper adjacent value, respectively, and there are more negative than positive outliers (Figures 4A). Three of the analog root-finding methods (ACS1, ACS3, and ACS3–1) with default parameters reach the accuracy level of the IFCN five-out-of-ten method with 2% MSO step size within half the number of steps (Table 5), whereas ACS1-N needs similar number of steps and ACS5 never achieves such level.

The non-adaptive ACS1-N (Figure 4, orange lines) has a monotonically declining median $|\delta x_{\mathrm{th}}|$ for the entire range of steps. The percentage of outliers, however, increases with the steps, showing that the errors does not improve for some subjects, which become a larger proportion of the total population as the errors of the other subjects in the population become more and more tightly distributed. The adaptive ACS1 (Figure 4, yellow lines) outperforms its non-adaptive counterpart by reaching a lower median $|\delta x_{\mathrm{th}}|$ faster and with distributions of δx_{th} being more negatively biased for ACS1 and having very few positive outliers.

The ACS-3 and ACS-5 methods (Figure 4, blue and green lines, respectively) with higher rate of convergence have a rapid exponential decay of the control sequence and do not further head for the threshold as soon as a_i is too small for a notable influence. They indeed plateau earlier, at around 40 and 20 steps, respectively, and at a higher median $|\delta x_{\mathrm{th}}|$ of 1.53% and 1.89%, respectively, albeit with a smaller portion of outliers as their distributions did not continue to tighten. In comparison, ACS3–1 (Figure 4, purple lines) switches from the fast geometric convergence that provides the rough estimate to a. s. harmonic convergence and performs very similarly compared to ACS1 in terms of median $|\delta x_{\mathrm{th}}|$. With increasing number of steps, the percentage of outliers increases for ACS3–1, though they are still less than those for either ACS1-N or ACS1.

Besides the convergence of the control sequence a_i, the root-finding methods' performance is also determined by the initial value a_0 of the sequence. The default value of a_0 (10) was chosen to approximate the typical slope of the recruitment curve but its definition was inevitably empirical and likely suboptimal. Examining the performance as a function of the initial step size indeed revealed that this was the case for almost all the root-finding methods. The change from overshoot and oscillatory behavior for large a_0 towards lengthy one-sided convergence for smaller numbers is clearly visible (Figure S8, panels A_i, B_i…, and I_i). Some methods do not have one single optimal initial step size, however, as the minimum of the median $|\delta x_{\mathrm{th}}|$ at a given step (Figure S8, panels A_{ii}, B_{ii}, …, and I_{ii}, black dashed line and dots) may not be obtained with the same initial step size. Instead of evaluating asymptotic behavior, we chose the optimal initial step size that minimized the error at step 40, the median number of steps of the IFCN five-out-of-ten method with 2% MSO step size. The performance indeed improved for all analog methods (Figure 5 and Table 5).

When the default initial step size is used, the digital versions of the root-finding methods (Figure S11, darker lines with triangle markers) could asymptotically perform either better than (DCS1-N and DCS1) or very similar to (DCS3 and DCS5) their analog counterpart in terms of median $|\delta x_{\mathrm{th}}|$. For the short-range behavior within 30 steps, however, the analog versions performed better, except for DCS1-N. Thus, the analog versions overall have faster initial convergence, utilizing the *soft* information, but as the root-finding approaches the target, the difference between the actual and target responses becomes very small ($|\Delta y_i| \ll 1$). The step towards the target $|\Delta x_i|$ is therefore smaller for the analog version ($|a_i \Delta y_i|$) compared to the digital version ($|a_i \delta y_i| = a_i$) and thus

approaches the target slower than the latter. With optimal initial step size, however, the difference between the analog and digital versions (Figure S12) becomes smaller for either short term or asymptotic behavior.

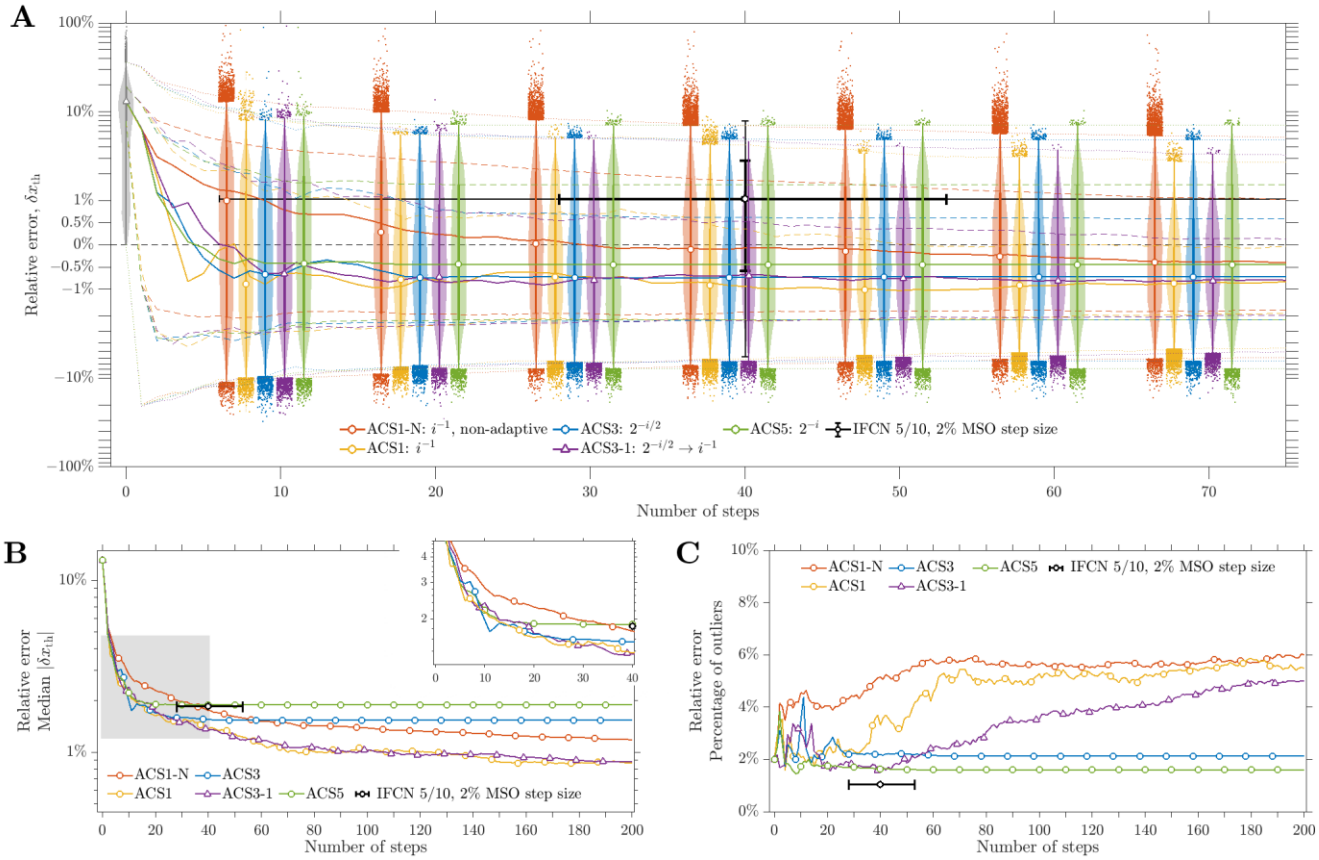

Figure 4: Statistics of the relative threshold errors of root-finding methods using first-order analog control sequence with default initial step size a_0

Similar to Figure 3, without the IQR of the errors in panel B and with linear scale in panel C.

Table 5 Optimal initial step size of the first-order root-finding methods and their error convergence compared to the reference IFCN five-out-of-ten method with 2%MSO step size

Method	Optimal a_0	Multiple of default value	Step reaching median $\lvert\delta x_{th}\rvert$ of 1.85%		Median $\lvert\delta x_{th}\rvert$ at step 40	
			Default	Optimal	Default	Optimal
ACS1-N	0.168	$\times 10^{0.4} \approx \times 2.51$	37	21	1.78%	1.29%
DCS1-N	0.084	$\times 10^{0.1} \approx \times 1.26$	17	14	1.20%	1.11%
ACS1	0.042	$\times 10^{-0.2} \approx \times 0.63$	16	17	1.39%	1.14%
DCS1	0.042	$\times 10^{-0.2} \approx \times 0.63$	20	16	1.13%	1.02%
ACS3	0.133	$\times 10^{0.3} \approx \times 2.00$	12	15	1.56%	1.33%
DCS3	0.084	$\times 10^{0.1} \approx \times 1.26$	17	17	1.53%	1.30%
ACS3–1	0.053	$\times 10^{-0.1} \approx \times 0.79$	20	15	1.37%	1.11%
ACS5	0.211	$\times 10^{0.5} = \times 3.16$	–	16	1.89%	1.77%
DCS5	0.211	$\times 10^{0.5} \approx \times 3.16$	–	21	1.87%	1.77%

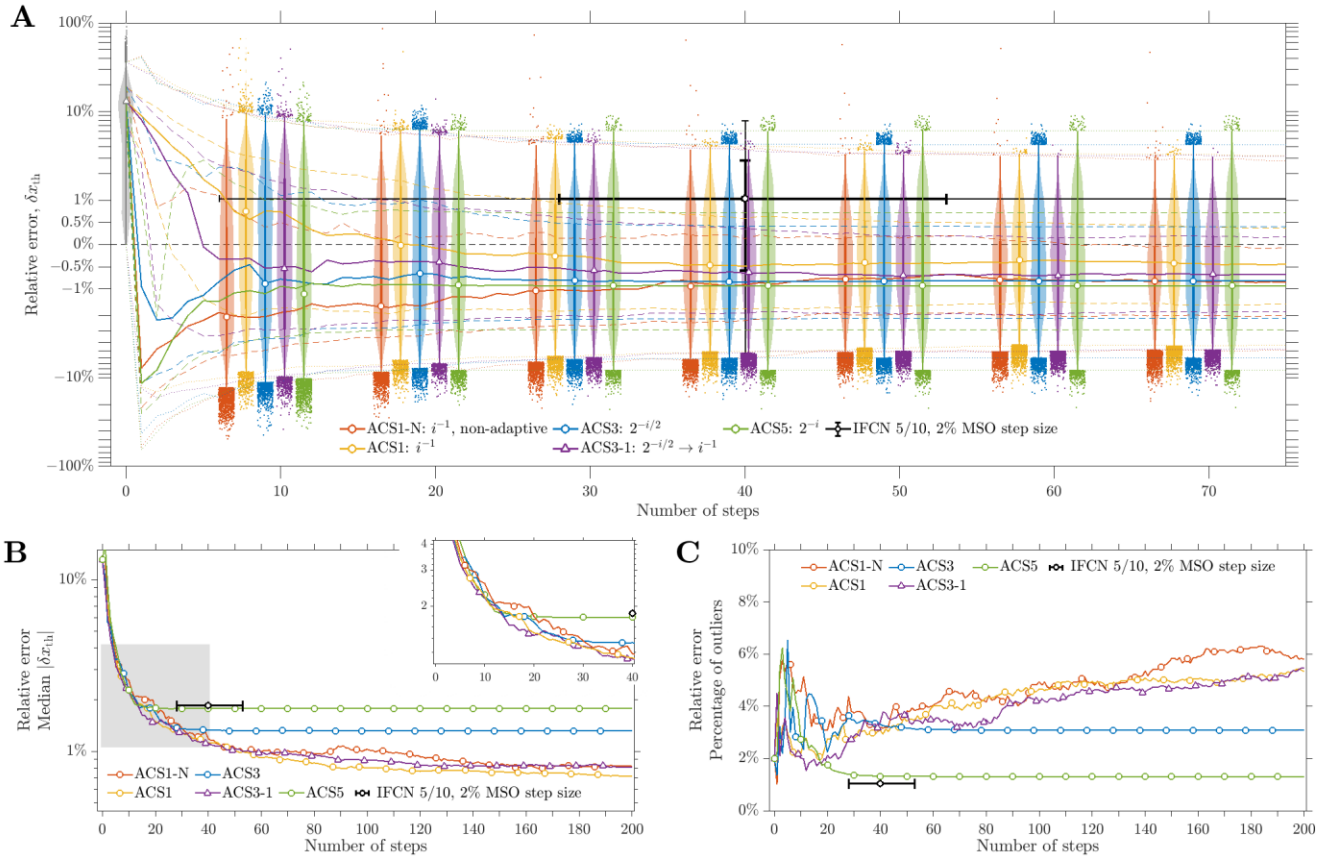

Figure 5: Statistics of the relative threshold errors of root-finding methods using first-order analog control sequence with optimal initial step size a_0
Similar to Figure 4.

The methods with second-order control sequences (Figure S13, lighter lines with square markers) again could perform either better (ACS2 and DCS4) or worse (ACS2-N and ACS6) than their first-order counterpart when the default parameters are used. Their performance depends on both the initial step size a_0 and the second-order weight w and the two parameters interact as demonstrated with ACS-2 (Figures S14 and S15). Large/positive w speed up the approach towards the threshold, but could generate oscillations similar to large a_0. Small/negative w, like small a_0, however, reduce the speed of convergence. The second-order weight can compensate a suboptimal initial step size to generate an optimal approach, and vice versa. Compared to the slope-like a_0 (9), an optimal w is difficult to determine the without empirical information or costly parameter sweeps. Therefore, the addition of a second-order term to the root-finding would be less preferable than optimizing the initial step size for the first-order methods.

Optimal Methods

We examined a variety of existing and new TMS thresholding methods and compared their performance in a virtual population of 25,000 subjects. The methods include eight based on relative frequencies using either a single threshold (IFCN methods) or two thresholds (MILLS–NITHI methods) and either 1% MSO or 2% MSO step size, seven parametric estimation methods, of which six use different combinations of threshold estimation and stepping rules and one is based on BAYESIAN estimation involving the probability density function itself, and thirteen

stochastic root-finding methods utilizing different convergence and stepping rules. Given the significant trade-off between different methods, a conclusive decision on a single optimal threshold estimation method may be difficult to provide at the moment (Figures 6 and S16). All relative-frequency methods have similar errors performance, and therefore the IFCN five-out-of-ten methods with 2% MSO step size with the minimum number of steps on average would be preferred. The two-parameter MAP estimation with MLE stepping demonstrates the best performance among the parametric estimators considering the speed of convergence, median error, and computational complexity. The KONTSEVICH–TYLER BAYESIAN estimation performs even better regarding the first two aspects, but with increased computational cost. The best stochastic root-finding methods using analog and digital control sequences are ACS3-1 and DCS1, respectively, irrespective of whether the optimal or default initial step size is used (Figures 6 and S16).

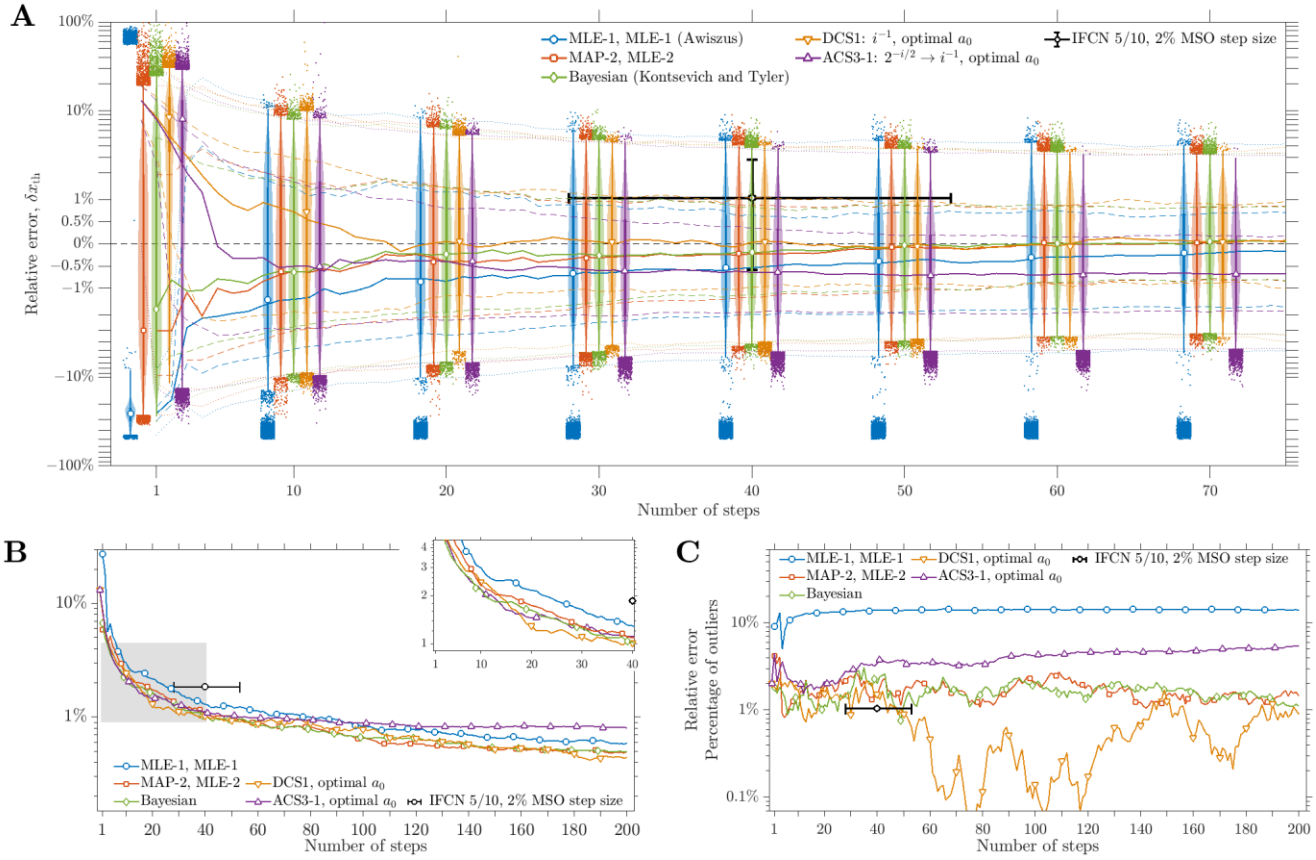

Figure 6: Statistics of the relative threshold errors of the best performing methods in each category
Similar to Figure 3, without the IQR of the errors in Panel B. The fastest relative-frequency method is the IFCN five-out-of-ten method with 2% MSO step size, whereas all relative-frequency methods have similar errors. The best parametric methods are two-parameter MLE stepping with MAP estimation (MAP-2, MLE-2) and KONTSEVICH–TYLER BAYESIAN estimation, with the single-parameter MLE methods by AWISZUS given for comparison. The best stochastic root-finding methods are ACS3-1 and DCS1 for analog and digital control sequence, respectively, here shown with optimal initial step size a_0.

Discussion

Despite their wide range of behavior and difference in performance, all methods performed remarkably accurate in terms of threshold error on the population level. Either Δx_{th} and δx_{th} (or $|x_{\mathrm{th}}|$ and $|\delta x_{\mathrm{th}}|$) have mean or median on the level of 1% MSO and 1%, respectively. In comparison, the pulse amplitude variability of TMS machines themselves have exactness (deviation from nominal value) and precision (variation between pulses) in comparable ranges (1% to 3%) depending on the device (Zacharias et al., 2019). The errors of individual subjects, however, have a large spread, and their relative error may not fall within the ±5% criteria for safety (Awiszus, 2011; Rossi et al., 2009). For some methods, the estimation may not improve with increasing number of TMS pulses for certain subjects (e.g., adaptive estimation using single parameter MLE stimulation and stepping) or the entire population (e.g., ACS3 and ACS5) after a certain number of steps.

We analyzed both the errors and their absolute values and in our results visualized the entire distribution or at least the outliers besides a wide range of statistics (median, Q_1, Q_3, IQR, upper and lower adjacent values, and a probability density distribution). In contrast, many previous studies only quantified the absolute values of the relative error (simply referring to it as the relative error) and/or presented mean and standard deviation only (Awiszus, 2003; Qi et al., 2011). The use of absolute values for errors provides the convenience to compare a positive metric that nicely decreases with stepping and allows the use of logarithmic scale for visualization. Whereas the median of the absolute values can capture the spread of the errors very well—our results showed that the ratio between the IQR of the errors and the median of the errors' absolute value is quite consistently around two for the methods examined—the absolute value of errors, however, do not capture the bias of the estimated thresholds or its symmetry around zero. Similarly, the mean and standard deviation do not capture outliers and skewness of the distributions. These are all important characteristics of the thresholding methods. For example, all four variations of the MILLS–NITHI methods have a negative bias and the original method has a large number of negative error outliers, contradicting its "mathematical safety" (Awiszus, 2011).

More importantly, existing estimation methods previously were tested using virtual subjects that have the same response model underlying the parameter estimation methods. Such tests of estimators against their own estimation model can be highly misleading. The binary GAUSSIAN response model (and the 0.07 ratio between slope and threshold) was heuristically suggested based on empirical data. This creates a statistically peculiar self-fulfilling prophecy, similar to the previously criticized circular definitions of hotspot and threshold (Awiszus, 2014a). We therefore used an independently derived MEP model from the literature, which was previously trained to the statistics of experimental data. The response model used here, although more sophisticated than most other approaches may still have limitations in representing all features and statistical properties of stimulation-evoked EMG responses due to so far limited study data (Darling et al., 2006; Kiers et al., 1993; Orth and Rothwell, 2004; Roy Choudhury et al., 2011). In addition, complex temporal effects, such as inter-pulse correlation caused by stimulation-induced short-term plasticity and also stimulation-independent excitability fluctuations (Möller et al., 2009) may become important, particularly for short inter-pulse intervals of only few seconds used in some labs. Exact knowledge of the physiologic system, however, is essential for the thresholding problem. This does not only influence the design of a realistic testbed, but also directly determines, for example, how well the GAUSSIAN model used in all parameter estimators suits the requirements or if another model, such as logistic distribution (Tranulis et al., 2006; Treutwein, 1995), is needed in order to avoid bias.

The performance of some estimation methods demonstrates substantial improvement over existing methods. For example, the two-parameter MAP and KONTSEVICH–TYLER BAYESIAN estimation methods all outperform the single- and two- parameter MLE methods. All parametric estimators, however, arouse justified criticism as

discussed that the underlying model could be inappropriate or in need of improvement. Furthermore, the computational cost should also be considered. The BAYESIAN method is computationally more expensive and the complexity may not be justified by the improvement in accuracy depending on the specific requirement of the study. Similarly, some of the parametric methods rely on accurate prior information of the model parameters (threshold and slope in the GAUSSIAN model) to achieve the presented accuracy and improvement over other methods. Such prior information on the population level could require significant effort to collect or may be completely unavailable.

The newly introduced non-parametric stochastic root-finding methods overall have only slightly higher errors compared to the best parametric estimation methods. Nevertheless, they do not rely on potentially incorrect models and thus could work for subjects whose MEP response curves do not follow the GAUSSIAN model. The rule for choosing and updating the step size a_i is also simple enough to be performed by hand or a calculator and do not necessarily rely on toolboxes (e.g., (Awiszus, 2003; Julkunen, 2019)). The best performing root-finding methods, e.g., DCS1 and ACS3–1, even with the default initial step size, are only slightly outperformed in terms of speed and accuracy by the best parametric estimation methods (two-parameter MLE stepping and MAP estimation and KONTSEVICH–TYLER BAYESIAN methods, both with prior information). The root-finding methods are a promising addition to the available threshold methods that motivate further studies and improvements.

Conclusion

This work analyzed five classes and a total of 28 variations of TMS thresholding methods and compared their performance, including their accuracy and speed. As was already criticized before, the relative frequencies methods may not be appropriate due to their large number of stimuli required and wide distribution of error. The accuracy is outperformed by parametric estimators and stochastic root-finding methods after a fraction of required stimuli already. As our MEP response model differs from the one underlying parametric estimation methods, some methods do not perform well for a subset of subjects in the population. Including all parameters in the estimation avoids such issues and incorporating prior information further boost their performance. Stochastic root-finding methods are comparable to adaptive estimation methods, but are much simpler to implement and do not rely on an underlying model.

Acknowledgments

This work was supported by grants No. R01NS117405 and No. RF1MH124943 from the National Institutes of Health of the United States of America; the content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency. Computational support was provided by the Duke Compute Cluster.

Author contribution

S. M. G. conceived, supervised, and secured funding for the study, and developed the theoretical framework and original simulation code. B.W. reconstructed and optimized the computational framework, performed the simulations, data analysis, and visualization. A.V.P. supervised and secured funding and computational resources for the study. B. W. and S. M. G. wrote the manuscript, and all authors revised, commented on, and approved the final version of the manuscript.

Code and data availability statement

The code and data that support the findings of this study are available upon reasonable request from the authors.

Conflict of interest

S. M. Goetz and A. V. Peterchev are inventors on patents and patent applications on TMS technology. Related to TMS technology, S. M. Goetz has received research funding from Magstim as well as Royalties from Rogue Research, and A. V. Peterchev has received research funding, travel support, patent royalties, consulting fees, equipment loans, hardware donations, and/or patent application support from Rogue Research, Tal Medical/Neurex, Magstim, MagVenture, Neuronetics, BTL Industries, and Advise Connect Inspire. B. Wang declares no relevant conflict of interest.

Appendix

Response model parameters

The individual recruitment parameters $p_{1,j}$ to $p_{7,j}$ show the following normal and lognormal distributions (Goetz et al., 2019):

$$\begin{aligned} p_{1,j} &\sim \mathcal{N}(\mu = -0.5818, \sigma^2 = 3.5705 \cdot 10^{-3}) \quad , \\ p_{2,j} &\sim \mathcal{N}(\mu = 4.5323, \sigma^2 = 0.12619) \quad , \\ \lg\begin{pmatrix} p_{3,j} \\ p_{4,j} \\ p_{5,j} \end{pmatrix} &\sim \mathcal{N}\left(\mu = \begin{bmatrix} 3.6466 \\ 0.42639 \\ 1.6665 \end{bmatrix}, \Sigma = \begin{bmatrix} 7.2027 & 0.56974 & -0.19746 \\ 0.56974 & 4.7426 \cdot 10^{-2} & -1.5113 \cdot 10^{-2} \\ -0.19746 & -1.5113 \cdot 10^{-2} & 7.1771 \cdot 10^{-2} \end{bmatrix}\right) \quad . \end{aligned}$$

The intra-individual variability sources $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$ follow normal distributions with subject-specific standard deviation at each step i:

$$\begin{aligned} \epsilon_{\mathrm{xadd},ij} &\sim \mathcal{N}\left(\mu = 0, p_{6,j}\right) \quad , \\ \epsilon_{\mathrm{ymult},ij} &\sim \mathcal{N}\left(\mu = -7, p_{7,j}\right) \quad . \end{aligned}$$

in which the standard deviations for subjects j follow lognormal distributions:

$$\begin{aligned} \lg\left(p_{6,j}\right) &\sim \mathcal{N}(\mu = -0.96453, \sigma^2 = 2.2759 \cdot 10^{-2}) \quad , \\ \lg\left(p_{7,j}\right) &\sim \mathcal{N}(\mu = \quad 0.68827, \sigma^2 = 2.3671 \cdot 10^{-2}) \quad . \end{aligned}$$

The additive noise $\epsilon_{\mathrm{yadd},ij}$ follows a generalized extreme-value (GEV) distribution, with subject-specific parameters

$$\epsilon_{\mathrm{yadd},ij} \sim \mathrm{GEV}(\mu = 10^{\,p_{1,j}}, \sigma = \ 1.55 \cdot 10^{-6} \cdot 1.5^{p_{8,j}}, k = \ 0.14 \cdot 2^{p_{9,j}}) \quad ,$$

where $p_{8,j}$ and $p_{9,j}$ are independent normal distributions $\mathcal{N}(0,1)$, and GEV distribution has a probability density function of

$$f_{\mathrm{GEV}} = \frac{1}{\sigma}\left(1 + \frac{k(x-\mu)}{\sigma}\right)^{-\left(1+\frac{1}{k}\right)} \cdot \exp\left(-\left(1 + \frac{k(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right) \quad .$$

The noiseless model has noise parameters $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$ set to zero and $\epsilon_{\mathrm{yadd},ij} = 10^{\,p_{1,j}}$, which approximates the mean of the GEV for the parameters $p_{1,j}$, $p_{8,j}$, and $p_{9,j}$.

Distribution of MLE threshold and slope

The MLE of threshold t and slope s (Figure S1D) for the GAUSSIAN model (11) can be fitted by a joint distribution

$$\begin{pmatrix} t_{\mathrm{MLE}} \\ \lg(s_{\mathrm{MLE}}) \end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} 0.65647 \\ -1.3028 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_t^2 & \rho\sigma_t\sigma_{\lg(s)} \\ \rho\sigma_t\sigma_{\lg(s)} & \sigma_{\lg(s)}^2 \end{bmatrix}\right) \quad ,$$

where $\sigma_t = 9.8215 \cdot 10^{-2}$, $\sigma_{\lg(s)} = 3.9674 \cdot 10^{-2}$, and the correlation coefficient $\rho = 0.0921$. Given the very small ρ, the distribution is simplified and approximated by independent normal and log-normal distributions, which are fitted as

$$\begin{aligned} t_{\mathrm{MLE}} &\sim \mathcal{N}(\mu = 0.65626, \sigma^2 = 9.5931 \cdot 10^{-3}) \quad , \\ \lg(s_{\mathrm{MLE}}) &\sim \mathcal{N}(\mu = -1.3027, \sigma^2 = 1.5731 \cdot 10^{-3}) \quad . \end{aligned}$$

References

Aberra, A. S., Wang, B., Grill, W. M., and Peterchev, A. V. Simulation of transcranial magnetic stimulation in head model with morphologically-realistic cortical neurons. *Brain Stimulation*, 13(1), pp. 175–189, 2020.

Anderson, B. S. and George, M. S. A review of studies comparing methods for determining transcranial magnetic stimulation motor threshold: observation of movement or electromyography assisted. *Journal of the American Psychiatric Nurses Association*, 15(5), pp. 304–313, 2009.

Antal, A., Nitsche, M. A., Kincses, T. Z., Lampe, C., and Paulus, W. No correlation between oving phosphene and motor thresholds: a transcranial magnetic stimulation study. *NeuroReport*, 15(2), pp. 297–302, 2004.

Awiszus, F. Chapter 2 TMS and threshold hunting. In *Supplements to Clinical Neurophysiology*, vol. 56, pp. 13–23. Elsevier, 2003.

Awiszus, F. Fast estimation of transcranial magnetic stimulation motor threshold: is it safe? *Brain Stimulation*, 4(1), pp. 58–59, 2011.

Awiszus, F. On relative frequency estimation of transcranial magnetic stimulation motor threshold. *Clinical Neurophysiology*, 123(11), pp. 2319–2320, 2012.

Awiszus, F. Of thresholds and “hot spots”. Quo vadis transcranial magnetic stimulation? *Clinical Neurophysiology*, 125(12), pp. 2451–2453, 2014a.

Awiszus, F. Using relative frequency estimation of transcranial magnetic stimulation motor threshold does not allow to draw any conclusions about true threshold. *Clinical Neurophysiology*, 125(6), pp. 1285–1286, 2014b.

Bawa, P. and Lemon, R. N. Recruitment of motor units in response to transcranial magnetic stimulation in man. *The Journal of Physiology*, 471(1), pp. 445–464, 1993.

Borckardt, J. J., Nahas, Z., Koola, J., and George, M. S. Estimating Resting Motor Thresholds in Transcranial Magnetic Stimulation Research and Practice: A computer Simulation Evaluation of Best Methods. *The Journal of ECT*, 22(3), pp. 169–175, 2006.

Capaday, C. On the variability of motor-evoked potentials: experimental results and mathematical model. *Experimental Brain Research*, 239(10), pp. 2979–2995, 2021.

Cotovio, G., Oliveira-Maia, A. J., Paul, C., Faro Viana, F., Rodrigues da Silva, D., Seybert, C., Stern, A. P., Pascual-Leone, A., and Press, D. Z. Day-to-day variability in motor threshold during rTMS treatment for depression: Clinical implications. *Brain Stimulation*, 14(5), pp. 1118–1125, 2021.

Darling, W. G., Wolf, S. L., and Butler, A. J. Variability of motor potentials evoked by transcranial magnetic stimulation depends on muscle activation. *Experimental Brain Research*, 174(2), pp. 376–385, 2006.

Deblieck, C., Thompson, B., Iacoboni, M., and Wu, A. D. Correlation between motor and phosphene thresholds: A transcranial magnetic stimulation study. *Human Brain Mapping*, 29(6), pp. 662–670, 2008.

Delyon, B. and Juditsky, A. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3(4), pp. 868–881, 1993.

Di Lazzaro, V., Profice, P., Ranieri, F., Capone, F., Dileone, M., Oliviero, A., and Pilato, F. I-wave origin and modulation. *Brain Stimulation*, 5(4), pp. 512–525, 2012.

Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594–604), pp. 309–368, 1922.

Goetz, S. M., Alavi, S. M. M., Deng, Z., and Peterchev, A. V. Statistical Model of Motor-Evoked Potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(8), pp. 1539–1545, 2019.

Goetz, S. M. and Deng, Z.-D. The development and modelling of devices and paradigms for transcranial magnetic stimulation. *International Review of Psychiatry*, 29(2), pp. 115–145, 2017.

Goetz, S. M., Howell, B., Wang, B., Li, Z., Sommer, M. A., Peterchev, A. V., and Grill, W. M. Isolating two sources of variability of subcortical stimulation to quantify fluctuations of corticospinal tract excitability. *Clinical Neurophysiology*, 138, pp. 134–142, 2022.

Goetz, S. M. and Kammer, T. Neuronavigation. In *The Oxford Handbook of TMS*, 2nd ed. Oxford University Press, Oxford, 2022.

Goetz, S. M., Li, Z., and Peterchev, A. V. Noninvasive Detection of Motor-Evoked Potentials in Response to Brain Stimulation Below the Noise Floor—How Weak Can a Stimulus Be and Still Stimulate. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4. IEEE, Honolulu, HI, USA, 2018.

Goetz, S. M., Luber, B., Lisanby, S. H., and Peterchev, A. V. A Novel Model Incorporating Two Variability Sources for Describing Motor Evoked Potentials. *Brain Stimulation*, 7(4), pp. 541–552, 2014.

Goetz, S. M. and Peterchev, A. V. A model of variability in brain stimulation evoked responses. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6434–6437, 2012.

Goetz, S., Whiting, P., and Peterchev, A. Threshold estimation with transcranial magnetic stimulation: algorithm comparison. *Clinical Neurophysiology*, 122, pp. S197, 2011.

Groppa, S., Oliviero, A., Eisen, A., Quartarone, A., Cohen, L. G., Mall, V., Kaelin-Lang, A., Mima, T., Rossi, S., Thickbroom, G. W., Rossini, P. M., Ziemann, U., Valls-Solé, J., and Siebner, H. R. A practical guide to diagnostic transcranial magnetic stimulation: Report of an IFCN committee. *Clinical Neurophysiology*, 123(5), pp. 858–882, 2012.

Halawa, I., Shirota, Y., Neef, A., Sommer, M., and Paulus, W. Neuronal tuning: Selective targeting of neuronal populations via manipulation of pulse width and directionality. *Brain Stimulation*, 2019.

Julkunen, P. Mobile Application for Adaptive Threshold Hunting in Transcranial Magnetic Stimulation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(8), pp. 1504–1510, 2019.

Jung, N. H., Delvendahl, I., Kuhnke, N. G., Hauschke, D., Stolle, S., and Mall, V. Navigated transcranial magnetic stimulation does not decrease the variability of motor-evoked potentials. *Brain Stimulation*, 3(2), pp. 87–94, 2010.

Kiers, L., Cros, D., Chiappa, K. H., and Fang, J. Variability of motor potentials evoked by transcranial magnetic stimulation. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 89(6), pp. 415–423, 1993.

Kobayashi, M. and Pascual-Leone, A. Transcranial magnetic stimulation in neurology. *The Lancet Neurology*, 2(3), pp. 145–156, 2003.

Kontsevich, L. L. and Tyler, C. W. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16), pp. 2729–2737, 1999.

Lai, T. L. Stochastic approximation. *The Annals of Statistics*, 31(2), pp. 391–406, 2003.

Lang, N., Harms, J., Weyh, T., Lemon, R. N., Paulus, W., Rothwell, J. C., and Siebner, H. R. Stimulus intensity and coil characteristics influence the efficacy of rTMS to suppress cortical excitability. *Clinical Neurophysiology*, 117(10), pp. 2292–2301, 2006.

Li, Z., Peterchev, A. V., Rothwell, J. C., and Goetz, S. M. *Detection of Motor-Evoked Potentials Below the Noise Floor: Rethinking the Motor Threshold*, pp. 2021.12.10.472126, 2021.

Mathias, J. P., Barsi, G. I., van de Ruit, M., and Grey, M. J. Rapid Acquisition of the Transcranial Magnetic Stimulation Stimulus Response Curve. *Brain Stimulation*, 7(1), pp. 59–65, 2014.

Mills, K. R. and Nithi, K. A. Corticomotor threshold to magnetic stimulation: Normal values and repeatability. *Muscle & Nerve*, 20(5), pp. 570–576, 1997.

Mishory, A., Molnar, C., Koola, J., Li, X., Kozel, F. A., Myrick, H., Stroud, Z., Nahas, Z., and George, M. S. The Maximum-likelihood Strategy for Determining Transcranial Magnetic Stimulation Motor Threshold, Using Parameter Estimation by Sequential Testing Is Faster Than Conventional Methods With Similar Precision. *The Journal of ECT*, 20(3), pp. 160–165, 2004.

Möller, C., Arai, N., Lücke, J., and Ziemann, U. Hysteresis effects on the input–output curve of motor evoked potentials. *Clinical Neurophysiology*, 120(5), pp. 1003–1008, 2009.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4), pp. 1574–1609, 2009.

Nielsen, J. Logarithmic Distribution of Amplitudes of Compound Muscle Action Potentials Evoked by Transcranial Magnetic Stimulation. *Journal of Clinical Neurophysiology*, 13(5), pp. 423–434, 1996.

Orth, M. and Rothwell, J. C. The cortical silent period: intrinsic variability and relation to the waveform of the transcranial magnetic stimulation pulse. *Clinical Neurophysiology*, 115(5), pp. 1076–1082, 2004.

Paninski, L. Asymptotic Theory of Information-Theoretic Experimental Design. *Neural Computation*, 17(7), pp. 1480–1507, 2005.

Pentland, A. Maximum likelihood estimation: the best PEST. *Perception & Psychophysics*, 1980.

Peterchev, A. V., Deng, Z.-D., and Goetz, S. M. Advances in Transcranial Magnetic Stimulation Technology. In *Brain Stimulation*, pp. 165–189. John Wiley & Sons, Inc, 2015.

Peterchev, A. V., Goetz, S. M., Westin, G. G., Luber, B., and Lisanby, S. H. Pulse width dependence of motor threshold and input–output curve characterized with controllable pulse parameter transcranial magnetic stimulation. *Clinical Neurophysiology*, 124(7), pp. 1364–1372, 2013.

Proakis, J. G. and Salehi, M. *Digital communications*. McGraw-Hill., 2008.

Qi, F., Wu, A. D., and Schweighofer, N. Fast estimation of transcranial magnetic stimulation motor threshold. *Brain Stimulation*, 4(1), pp. 50–57, 2011.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3), pp. 400–407, 1951.

Rossi, S., Antal, A., Bestmann, S., Bikson, M., Brewer, C., Brockmöller, J., Carpenter, L. L., Cincotta, M., Chen, R., Daskalakis, J. D., Di Lazzaro, V., Fox, M. D., George, M. S., Gilbert, D., Kimiskidis, V. K., Koch, G., Ilmoniemi, R. J., Pascal Lefaucheur, J., Leocani, L., Lisanby, S. H., Miniussi, C., Padberg, F., Pascual-Leone, A., Paulus, W., Peterchev, A. V., Quartarone, A., Rotenberg, A., Rothwell, J., Rossini, P. M., Santarnecchi, E., Shafi, M. M., Siebner, H. R., Ugawa, Y., Wassermann, E. M., Zangen, A., Ziemann, U., and Hallett, M. Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines. *Clinical Neurophysiology*, 2020.

Rossi, S., Hallett, M., Rossini, P. M., and Pascual-Leone, A. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical Neurophysiology*, 120(12), pp. 2008–2039, 2009.

Rossini, P. M., Barker, A. T., Berardelli, A., Caramia, M. D., Caruso, G., Cracco, R. Q., Dimitrijević, M. R., Hallett, M., Katayama, Y., Lücking, C. H., Maertens de Noordhout, A. L., Marsden, C. D., Murray, N. M. F., Rothwell, J. C., Swash, M., and Tomberg, C. 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. *Electroencephalography and Clinical Neurophysiology*, 91(2), pp. 79–92, 1994.

Rossini, P. M., Burke, D., Chen, R., Cohen, L. G., Daskalakis, Z., Di Iorio, R., Di Lazzaro, V., Ferreri, F., Fitzgerald, P. B., George, M. S., Hallett, M., Lefaucheur, J. P., Langguth, B., Matsumoto, H., Miniussi, C., Nitsche, M. A., Pascual-Leone, A., Paulus, W., Rossi, S., Rothwell, J. C., Siebner, H. R., Ugawa, Y., Walsh, V., and Ziemann, U. 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. *Clinical Neurophysiology*, 126(6), pp. 1071–1107, 2015.

Roy Choudhury, K., Boyle, L., Burke, M., Lombard, W., Ryan, S., and McNamara, B. Intra subject variation and correlation of motor potentials evoked by transcranial magnetic stimulation. *Irish Journal of Medical Science*, 180(4), pp. 873–880, 2011.

Sasaki, T., Shirota, Y., Kodama, S., Togashi, N., Sugiyama, Y., Tokushige, S., Inomata-Terada, S., Terao, Y., Ugawa, Y., Toda, T., and Hamada, M. Modulation of motor learning by a paired associative stimulation protocol inducing LTD-like effects. *Brain Stimulation*, 11(6), pp. 1314–1321, 2018.

Sommer, M., Alfaro, A., Rummel, M., Speck, S., Lang, N., Tings, T., and Paulus, W. Half sine, monophasic and biphasic transcranial magnetic stimulation of the human motor cortex. *Clinical Neurophysiology*, 117(4), pp. 838–844, 2006.

Stewart, L. M., Walsh, V., and Rothwell, J. C. Motor and phosphene thresholds: a transcranial magnetic stimulation correlation study. *Neuropsychologia*, 39(4), pp. 415–419, 2001.

Strasburger, H. and Treutwein, B. Fitting the psychometric function. *Perception & Psychophysics*, 61(1), pp. 87–106, 1999.

Tranulis, C., Guéguen, B., Pham-Scottez, A., Vacheron, M. N., Cabelguen, G., Costantini, A., Valero, G., and Galinovski, A. Motor threshold in transcranial magnetic stimulation: comparison of three estimation methods. *Neurophysiologie Clinique/Clinical Neurophysiology*, 36(1), pp. 1–7, 2006.

Treutwein, B. Adaptive psychophysical procedures. *Vision Research*, 35(17), pp. 2503–2522, 1995.

Turi, Z., Lenz, M., Paulus, W., Mittner, M., and Vlachos, A. Selecting stimulation intensity in repetitive transcranial magnetic stimulation studies: A systematic review between 1991 and 2020. *European Journal of Neuroscience*, 53(10), pp. 3404–3415, 2021.

Wang, B. *TMS_Thresholding*. MATLAB, Github, Zenodo, 2022.

Wassermann, E. M. Risk and safety of repetitive transcranial magnetic stimulation: report and suggested guidelines from the International Workshop on the Safety of Repetitive Transcranial Magnetic Stimulation, June 5–7, 1996. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 108(1), pp. 1–16, 1998.

Wassermann, E. M. Variation in the response to transcranial magnetic brain stimulation in the general population. *Clinical Neurophysiology*, 113(7), pp. 1165–1171, 2002.

Westin, G. G., Bassi, B. D., Lisanby, S. H., and Luber, B. Determination of motor threshold using visual observation overestimates transcranial magnetic stimulation dosage: Safety implications. *Clinical Neurophysiology*, 125(1), pp. 142–147, 2014.

Wichmann, F. A. and Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8), pp. 1293–1313, 2001.

Williams, D. *Probability with martingales*. Cambridge university press, 1991.

Zacharias, L. R., Peres, A. S. C., Souza, V. H., Conforto, A. B., and Baffa, O. Method to assess the mismatch between the measured and nominal parameters of transcranial magnetic stimulation devices. *Journal of Neuroscience Methods*, 322, pp. 83–87, 2019.

Zhong, G., Yang, Z., and Jiang, T. Precise Modulation Strategies for Transcranial Magnetic Stimulation: Advances and Future Directions. *Neuroscience Bulletin*, 37(12), pp. 1718–1734, 2021.