RESEARCH ARTICLE

Performance metrics for an application-driven selection and optimization of psychophysical sampling procedures

Mike D. Rinderknecht*, Olivier Lambercy, Roger Gassert

Rehabilitation Engineering Laboratory, Institute of Robotics and Intelligent Systems, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

* mike.rinderknecht@hest.ethz.ch

Abstract

When estimating psychometric functions with sampling procedures, psychophysical assessments should be precise and accurate while being as efficient as possible to reduce assessment duration. The estimation performance of sampling procedures is commonly evaluated in computer simulations for single psychometric functions and reported using metrics as a function of number of trials. However, the estimation performance of a sampling procedure may vary for different psychometric functions. Therefore, the results of these type of evaluations may not be generalizable to a heterogeneous population of interest. In addition, the maximum number of trials is often imposed by time restrictions, especially in clinical applications, making trial-based metrics suboptimal. Hence, the benefit of these simulations to select and tune an ideal sampling procedure for a specific application is limited. We suggest to evaluate the estimation performance of sampling procedures in simulations covering the entire range of psychometric functions found in a population of interest, and propose a comprehensive set of performance metrics for a detailed analysis. To illustrate the information gained from these metrics in an application example, six sampling procedures were evaluated in a computer simulation based on prior knowledge on the population distribution and requirements from proprioceptive assessments. The metrics revealed limitations of the sampling procedures, such as inhomogeneous or systematically decreasing performance depending on the psychometric functions, which can inform the tuning process of a sampling procedure. More advanced metrics allowed directly comparing overall performances of different sampling procedures and select the best-suited sampling procedure for the example application. The proposed analysis metrics can be used for any sampling procedure and the estimation of any parameter of a psychometric function, independent of the shape of the psychometric function and of how such a parameter was estimated. This framework should help to accelerate the development process of psychophysical assessments.

OPEN ACCESS

Citation: Rinderknecht MD, Lambercy O, Gassert R (2018) Performance metrics for an application-driven selection and optimization of psychophysical sampling procedures. PLoS ONE 13(11): e0207217. https://doi.org/10.1371/journal.pone.0207217

Editor: Jane Foster, McMaster University, CANADA

Received: April 23, 2018

Accepted: October 26, 2018

Published: November 28, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. No physical datasets were used in this work, only simulations that rely on large MATLAB structures and classes that require custom-made libraries. All method parameters, settings of the simulation, parametric models and software packages necessary for the replication are described, listed, or referenced in the Materials and Methods section of the manuscript.

Funding: This research was supported by the ETH Zurich Foundation in collaboration with Hocoma

AG, the Janggen-Pöhn Foundation, and the Swiss National Science Foundation through project 320030L_170163. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Estimating psychometric functions is an important topic, both in basic psychophysics research to investigate mechanisms of sensation and perception, but also in clinical assessments to diagnose sensory deficits, e.g., after neurological injuries. A psychometric function relates physical stimuli to the perception, respectively performance, of the subject in detection and discrimination tasks [1]. In order to estimate a psychometric function, stimuli of different magnitudes (also referred to as levels) have to be presented to the subject, who then has to rate the stimuli according to the paradigm used in the experiment (e.g., yes-no, same-different, reminder or two-alternative forced choice (2AFC) tasks) [2, 3].

The term *sampling procedure* usually encompasses a set of rules with procedure-specific parameters defining the levels and order of the presented stimuli. There exist many different sampling procedures. The classical sampling procedure is the method of constant stimuli (MOCS) [2] presenting stimuli at predefined (i.e., fixed-grid) levels spanning the perception range of interest. While this sampling procedure can be used to obtain the entire shape of the psychometric function, adaptive sampling procedures have been developed to quantify only specific features of a psychometric function (e.g., the perception threshold or slope of a sigmoidal psychometric function) (see [4, 5] for reviews). These range, among others, from relatively simple staircase [6–10], heuristic [11, 12], and stochastic approaches [13, 14] to Bayesian [15–17] and maximum-likelihood procedures [18–20].

An ideal sampling procedure should be precise and accurate (i.e., present low inherent method variability and be unbiased) to guarantee a high assessment reliability. In addition, sampling procedures should be as efficient as possible (i.e., low number of required trials to achieve a wanted precision) as, especially in clinical settings, time is scarce and costly (e.g., [21]). Thus, often, the number of trials is strictly limited due to time constraints or because lengthy experiments could be detrimental for the subject's attention and lead to mental fatigue [22]. Resulting time-dependent alteration of perception (i.e., drift of psychometric functions) can lead to misestimations of parameters [23–26]. Moreover, depending on the application scenario, the inter-subject variability may differ and prior knowledge on the distribution of the population of interest may be available or not. As a consequence, different values for the sampling procedure-specific parameters of adaptive sampling procedures defining the stimulus levels to be presented (e.g., around the threshold) may be needed for rapid convergence towards desired features of the psychometric function. Therefore, the question arises how sampling procedures and how sampling procedure-specific parameters should be selected for best performance in a specific application scenario.

As evaluating the performance of sampling procedures through a series of behavioral studies would be too time consuming, computer simulations offer a valuable alternative and powerful tool to simulate psychophysical experiments and to evaluate different sampling procedures and sampling procedure-specific parameters. Besides being used to investigate the process of fitting psychometric functions to psychophysical data [27–31], computer simulations have been widely used to simulate sampling procedures and quantify their properties, such as the efficiency [8, 11, 12, 18–20, 32–35]. To quantify the efficiency of a sampling procedure, a metric called *sweat factor* was proposed [12, 36]. The sweat factor is defined by the product of the variance of the estimates and the number of trials, in order to evaluate the relative benefit of a longer procedure (i.e., more trials) for a reduced measurement error. Various other approaches and metrics have also been proposed to evaluate the performance of sampling procedures. They commonly include, for example, *mean* or *bias* (e.g., [34, 35, 37]), *standard deviation* (e.g., [18, 34, 35]), or *settling accuracy* (e.g., [11, 20]). Others have additionally used *information gain* in bits [35] or *percentage usable* [37].

However, these performance metrics are commonly used as a function of the number of trials, which may not be the factor which can be acted upon in many application scenarios due to limited time for assessments. When calculated for one given maximum number of trials, the sweat factor's information content is actually confined to the variability and cannot provide additional information. Furthermore, many simulations sample only one or a very limited number of threshold or slope parameter values [8, 11, 18, 20, 33–35, 38]. As a matter of fact, the outcome for those metrics may depend on the actual parameter values (e.g., threshold and slope) of the psychometric function to be estimated, and performance may not be homogeneous across this parameter space. Thus, performance results are very likely not representative for other psychometric functions of the population of interest. Instead, constraining the simulations to a specific number of trials given by the requirements of the application and exploring the estimation performance for different psychometric functions covering the entire threshold/slope parameter space of the population would provide relevant insight when selecting and optimizing a sampling procedure for a specific application.

The aim of this paper is to take an application-driven approach and introduce an evaluation framework with a comprehensive set of metrics to analyze psychophysical procedures in terms of threshold and slope estimation performance. We suggest to use error measures widely used in motor control and learning studies [39] in order to describe bias and variability, and introduce *percentage within bounds* (*PCTw/iB*) curves, a practical measure depending on desired estimation tolerances which can be directly related to application requirements. Based on this concept, the *normalized area under the curve* (*nAUC*), spanning a surface in a specific threshold-slope parameter space, can be computed. With the *normalized volume under the surface* (*nVUS*) and the inhomogeneity σ, we propose measures to compare the performance across different procedures or settings. This framework should facilitate the selection and optimization of sampling procedures for specific applications.

Inspired by a real-world scenario—assessment of proprioceptive joint angle difference thresholds using a 2AFC paradigm in a clinical setting—the metrics are illustrated and discussed here on six different procedures using computer simulations: (i) MOCS [2], (ii) Weighted Up-Down method [8], (iii) slightly altered Parameter Estimation by Sequential Testing (PEST) [12, 40], (iv–v) standard and accelerated Stochastic Approximation (SA) Staircases [13, 14], and (vi) the Bayesian Ψ (PSI) method [15] to illustrate what kind of insights can be gained by the proposed framework.

2 Definition and parameters of the psychometric function

In the present work, the perception models consisted of psychometric functions $\psi(x)$ defining the proportion of correct responses at different stimulus levels x:

$$\psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda) F(x; \alpha, \beta) \quad , \tag{1}$$

with the generic threshold parameter α, generic slope parameter β, guessing rate γ, and lapse rate λ (taking into account stimulus-independent errors, or "lapses"). The guess rate γ depends on the psychophysical paradigm (e.g., yes-no: $\gamma = 0$, 2AFC: $\gamma = 0.5$). The lapse rate λ is often set to 0, to limit the complexity of computer simulations. For the generic sigmoid function $F(x; \alpha, \beta)$, a cumulative normal function $F_{Gauss}(x; \mu, \sigma)$ with a mean μ and standard deviation σ according to the following equation was chosen:

$$F_{Gauss}(x; \mu, \sigma) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right)\right) \quad , \tag{2}$$

where erf(x) is the standard definition of the error function:

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2} \mathrm{d}t \quad . \tag{3}$$

As a consequence of this choice and a lapse rate λ of 0, the generic threshold parameter α corresponds directly to the inflection point μ and the threshold, which is defined at $x_T = \psi^{-1}(P_t)$, being the target probability or proportion of correct responses. The generic slope parameter β is inversely proportional to the "spread" (i.e., standard deviation σ at this point). In order to have comparable values across different studies using various analytic functions, it has been recommended to use the maximum actual slope $Slope_\alpha$, instead of the slope parameter β depending on the type of cumulative distribution. This is achieved by taking the first derivative $d\psi/dx|_{x=\alpha}$ [41] according to:

$$Slope_\alpha = \frac{(1-\gamma-\lambda)}{\sqrt{2\pi}} \frac{1}{\sigma} \quad . \tag{4}$$

The parameters and psychometric functions are illustrated in Fig 1.

3 Performance metrics for psychophysical procedures

In order to quantify the estimation performance of a sampling procedure, the "real" value of the parameter in question (i.e., in this case the threshold) of the psychometric function to be estimated should be known. This is where computers simulations come in useful, as the "real" psychometric function can be modeled and is known. Furthermore, the psychophysical experiment using the sampling procedure should ideally be simulated multiple times to obtain a distribution of estimates for high statistical power. To better distinguish between the "real" psychometric function to be estimated and the psychometric function fitted to the data provided by a (simulated) experiment, the first is referred to as template $\psi_{i,j}^{\top}(x)$ (the symbol $\top$ denotes a template, and indices i, j the combination of threshold α and slope $Slope_\alpha$ values).

The most elementary analysis consists of quantifying estimation bias (accuracy) and estimation variability (precision). The following nomenclature of constant errors (CE = average

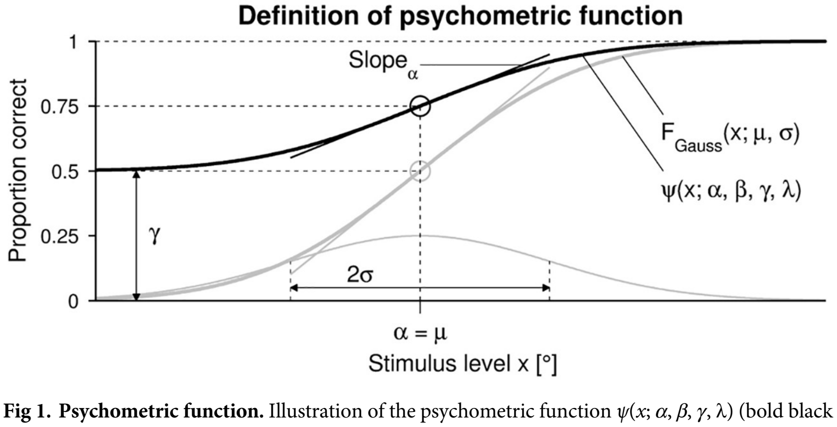

Fig 1. Psychometric function. Illustration of the psychometric function $\psi(x; \alpha, \beta, \gamma, \lambda)$ (bold black sigmoid) and related parameters. The underlying cumulative normal function $F_{Gauss}(x; \mu, \sigma)$ (bold gray sigmoid) and its underlying normal probability density function (gray) are also indicated. The slopes are indicated at the inflection points in the same color of the corresponding sigmoids. Note that the psychometric function is illustrated for the present application of proprioceptive joint angle assessments using a 2AFC paradigm, where the stimulus level x corresponds to the angular difference between two stimuli to be distinguished in a trial in degrees (°). In this application, γ was set to 0.5, λ to 0, and the difference threshold was defined at $x_T = \psi^{-1}(P_t)$, with $P_t = 0.75$.

https://doi.org/10.1371/journal.pone.0207217.g001

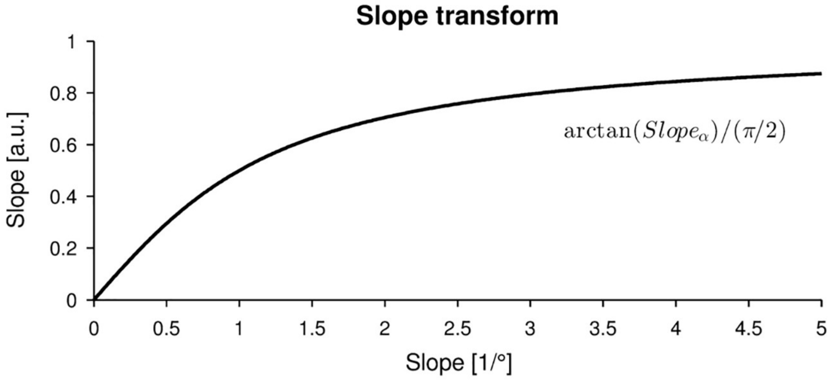

Fig 2. Slope transform. Function transforming the actual slope in stimulus units (in the present application of proprioceptive joint angle assessments in (1/°)) to arbitrary units (a.u.) within the range [0, 1]. This transformation prior to the calculation of the performance metrics is essential to make the estimation errors comparable for different slope values.

https://doi.org/10.1371/journal.pone.0207217.g002

signed errors) and variable errors (*VE* = standard deviations of errors) was used. This nomenclature is commonly used in motor control and learning [42]. In order to evaluate the estimation performance of sampling procedure, the threshold estimation error is computed by subtracting the threshold of the template $\psi_{i,j}^{\top}$ from the threshold of the estimated psychometric function. A positive error represents an overestimation (i.e., larger threshold or larger slope estimates compared to the template), whereas a negative error represents an underestimation (i.e. smaller threshold or smaller slope estimates). Before calculating the slope estimation error in the same way, an arctan($Slope_\alpha$)/(π/2) transform was applied (Fig 2). This transformed the slope space from [0, inf) to [0, 1] in arbitrary units (a.u.), where zero and one corresponded to a completely flat psychometric function and a perfect step function, respectively. Without applying this transform first, slope errors around large slopes (i.e., steep psychometric functions) diverge towards infinity, despite the psychometric functions looking almost identically. Calculating and plotting the *CE* and *VE* for a fine, two dimensional grid of threshold and slope parameter values allows to identify potential zones in the threshold-slope space, which may suffer from poorer estimation performance. This may provide insight on how sampling procedure-specific parameters could be tuned for the psychophysical assessment application.

Since the absolute errors (*AE* = average absolute errors) are a complex combination of *CE* and *VE* and can be predicted from them [42], direct examination of the *AE* becomes superfluous. However, it can be used in an application-driven approach to develop other higher level metrics building upon it. As the required precision depends on the application, it can be useful to describe the probability of a resulting estimation lying within an interval. Thus, the performance of a procedure could be expressed as the percentage of simulation results of threshold and slope estimates lying within a tolerance interval around the template values, respectively, as a function of the interval size. This *percentage within bounds* (*PCTw/iB*) function is related to absolute errors as illustrated in Fig 3(A)–3(C). A faster-rising *PCTw/iB* function would correspond to a stimulus selection method with higher performance. This metric would allows selecting the optimal sampling procedure given a required maximal error. Note that this metric takes into account both accuracy and precision, but provides more information relevant to the application compared to the elementary absolute error.

Similar to the receiver operating characteristic curve, the performance of the sampling procedure can be quantified by the area under the curve (AUC) (i.e., definite integral under the *PCTw/iB*-curve). Since the AUC is not bounded in the case of the threshold and could not be

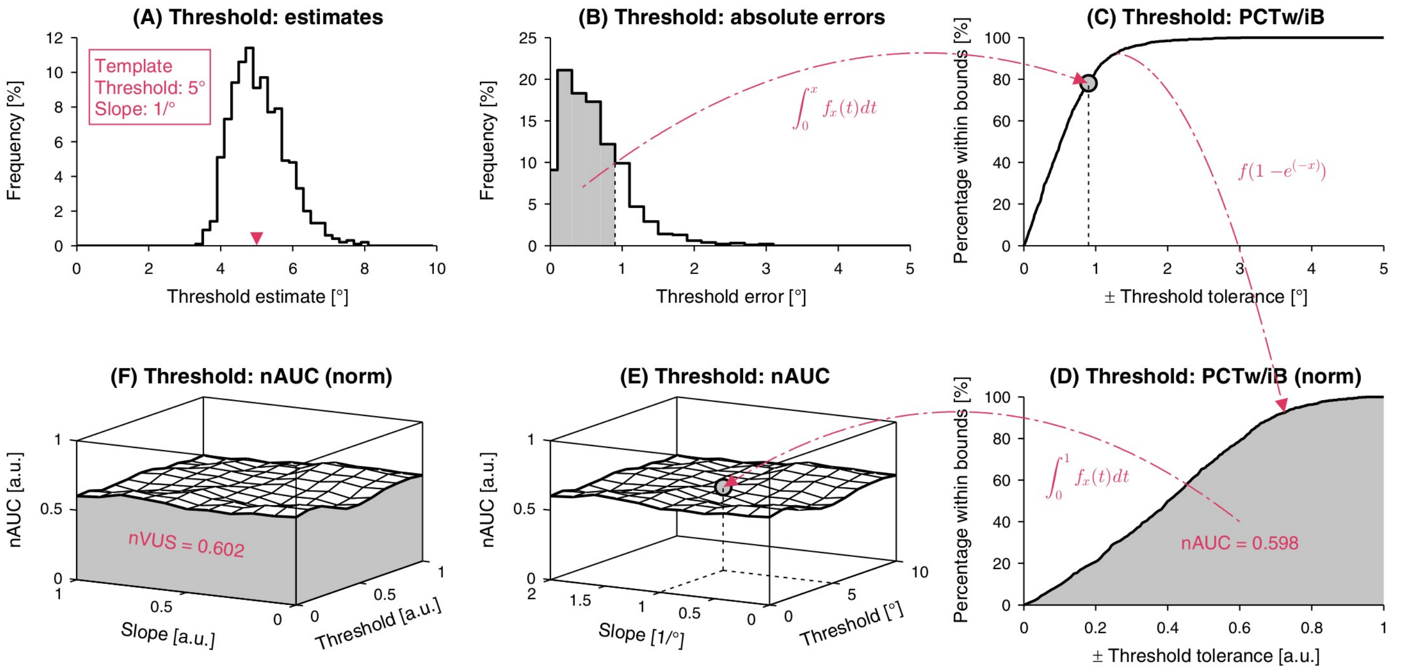

Fig 3. Explanation of performance metrics to characterize psychophysical sampling procedures. **(A)** Distribution of threshold estimates for a given template. **(B)** Distribution of the absolute errors. **(C)** *Percentage within bounds* (*PCTw/iB*) as a function of the tolerance interval. Each point of this function is generated by integrating the area of the absolute error distribution from zero to the respective error, respectively tolerance, value. **(D)** After a nonlinear transform of the *x* axis to [0, 1], the *normalized area under the curve* (*nAUC*) can be calculated (1 corresponding to perfect estimation for this template). **(E)** When the *nAUC* is computed for different templates within a threshold and slope parameter space, the performance can be visualized as a surface in a three-dimensional space. **(F)** After normalizing the parameter space axes to [0, 1], the *normalized volume under the surface* (*nVUS*) can be calculated (1 corresponding to perfect estimation across all templates). Note that the metrics are illustrated for the present application of proprioceptive joint angle assessments using a 2AFC paradigm, for a psychometric function template $\psi^{\top}_{5^{\circ},1/^{\circ}}$.

https://doi.org/10.1371/journal.pone.0207217.g003

calculated without defining an arbitrary upper bound on the tolerance interval size, a nonlinear transform $1 - e^{(-x)}$ is applied to the tolerance interval axis x of the *PCTw/iB*-curve. With this transform, the positive semi-infinite support [0,inf) is transformed to [0, 1] and the *normalized AUC* (*nAUC*) $\in [0, 1]$ can be calculated (Fig 3(D)). Due to the transform applied to the slopes before calculating the errors, the tolerance interval axis is already bounded and normalized to [0, 1]. Therefore, this nonlinear transform is not required anymore and the *nAUC* can be calculated directly for the slopes. The precision of *nAUC* can be improved by increasing the number of simulated runs for each template $\psi^{\top}_{i,j}$. A repeatedly perfect estimation of the threshold or slope parameter would result in a *nAUC* of one. This metric takes all estimates into account without having to calculate the parametric statistics, such as the arithmetic mean for the average absolute error. This metric is still template-dependent and can be used for more high-level metrics quantifying performance over the complete threshold and slope parameter space.

The performance of a procedure may vary depending on the threshold and slope of the psychometric function in question. Therefore, the *PCTw/iB* and corresponding *nAUC* for the threshold and slope estimation have to be calculated for each template $\psi^{\top}_{i,j}$. The *nAUC* can be visualized as a surface in a three-dimensional space (Fig 3(E)). For a given threshold and slope parameter space the *normalized volume under the surface* (*nVUS*) can be calculated after linearly normalizing the threshold and slope axes to [0, 1] (Fig 3(F)). As a result, the *nVUS* is also $\in [0, 1]$ and can be used to compare different procedures or method settings, as long as the same application-dependent parameter space range is used. The accuracy of $nVUS$ can be improved by simulating a denser grid of templates $\psi_{i,j}^{\top}$. A $nVUS$ of one corresponds to perfect estimations for all threshold and slope combinations within the evaluated parameter space. This metric can be used for an overall performance comparison across sampling procedures.

In addition to the $nVUS$, the performance variability can be evaluated: The inhomogeneity of the $nAUC$ across the parameter space can be described by calculating the standard deviation σ of all the $nAUC$ values for the different templates $\psi_{i,j}^{\top}$. This parameter σ should not be confused with the parameter of $F(x; \mu, \sigma)$. To calculate the standard deviation, each axis of the parameter space should be linearly sampled to avoid bias towards $nAUC$ values where the sampling density is higher. In case the sampling of the simulated parameter space is not linear along one or both axes, the $nAUC$ surface has to be resampled and interpolated along the axes in question.

To have an overall performance measure for each threshold and slope estimate as a function of the accepted estimation tolerance, for each tolerance interval the $PCTw/iB_{\pm Tol}$ can be presented as a surface for the parameter space, similar to the $nAUC$. From this surface, the overall tolerance-dependent performance can be calculated, as done for the $nVUS$ and σ.

4 Computer simulations

In order to illustrate and discuss the performance evaluation metrics described in the previous section for different procedures, a computer simulation based on a concrete application example was implemented.

4.1 Application example: Assessment of proprioceptive difference thresholds

Accurate and sensitive assessments of proprioception may be used for diagnosis, prognosis and treatment planning [43] for patients with somatosensory deficits affecting the upper limbs (e.g., after neurological injuries and diseases). However, clinical assessments, such as the up-down finger proprioception test [44], provide mostly dichotomous or ordinal scales and may thus be used for screening, but not for assessing functional improvements [45]. The combination of psychophysical procedures to estimate psychometric functions with robotic technology would offer more reproducible assessments with a higher resolution. There have been few studies exploring this approach for the assessment of the upper limb [40, 46–52]. So far MOCS has been predominantly used, with experiments typically lasting about 45 min [46–49]. However, in order to achieve clinical utility, the number of trials and the assessment duration should be reduced to below 15 min without compromising the quality of the outcome measures. This may be achieved by using adaptive procedures (e.g., as shown in a pilot study comparing MOCS and the adaptive PEST experimentally [40]).

The problem with such experimental validations is that the "real" psychometric functions are unknown. Therefore, the estimation performance and efficiency of the sampling procedures cannot be directly quantified. This can be addressed in computer simulations, where the actual psychometric function templates are known. Moreover, this allows for optimization of sampling procedure-specific parameter settings for a specific threshold and slope range of the population of interest. So far, experimental results on the difference threshold of angular joint position revealed values ranging from 1 to 5°, approximately, for elbow, wrist, and finger joints in healthy subjects [40, 46–48, 50–53]. However, in patients with proprioceptive deficits those values are higher and may go up to around 10° [54].

The quality of estimation of difference thresholds can be affected by the used experimental paradigm. While psychophysical experiments based on paradigms such as yes-no, reminder, and same-different can provide quantitative results, they are contaminated by effects of the decision criterion (i.e., response bias) [2, 3]. Despite some literature claiming that the two-alternative forced-choice (2AFC) paradigm requires a two- to three-fold number of trials for a given precision compared to the Yes-No paradigm [32, 55], 2AFC addresses the previously mentioned limitations, as it is expected to be a more sensitive, more objective and almost bias-free alternative [3].

Assessing proprioception at a specific joint using the 2AFC paradigm requires a two-interval design: two different stimuli (i.e., two flexion or extension movements are consecutively presented) before the subject rates which angle, respectively movement from a reference position, was larger. In the present work the difference between the two angles of one trial is referred to as level *x*. Prior work showed that one trial including response time lasts about 15 s [40]. Thus within an acceptable assessment duration of 15 min around 60 trials can be administered.

These criteria set the stage for the following computer simulations, illustrating the proposed evaluation metrics for a real-world application example.

4.2 Methods

4.2.1 Templates. To cover the full population, a set of 240 subject templates $\psi^{\top}_{i,j}$ was created with 40 different thresholds linearly spaced within the range [0.25°, 10°] (index *i*) and slopes $Slope_{\alpha} \in \{0.0625, 0.125, 0.25, 0.5, 1, 2\}/^{\circ}$ (index *j*). The guess rate γ was 0.5 and the lapse rate λ was zero for all templates. As an example, the template $\psi^{\top}_{5^{\circ},0.25/^{\circ}}$ would have a threshold of 5° and a slope $Slope_{\alpha}$ of 0.25/°. Examples of modeled psychometric functions with a threshold of 5° and different slopes are visualized in Fig 4.

4.2.2 Simulation process. The response of a simulated subject for a given stimulus level *x* was generated by comparing a randomly generated number between 0 and 1 to $\psi^{\top}_{i,j}(x)$. A larger random number compared to $\psi^{\top}_{i,j}(x)$ corresponded to a false response, and a smaller random number compared to $\psi^{\top}_{i,j}(x)$ to a correct response.

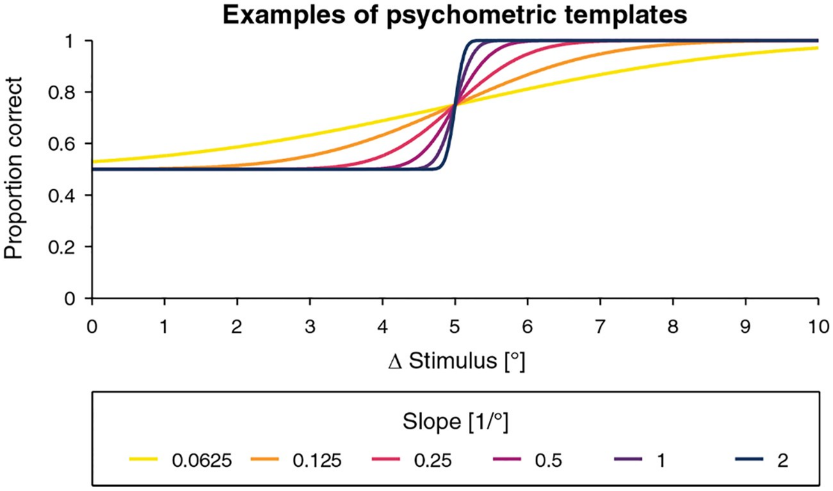

Fig 4. Examples of psychometric templates. Examples of psychometric templates $\psi^{\top}_{5^{\circ},j}$ with a threshold of 5° and different slopes used in the computer simulations.

https://doi.org/10.1371/journal.pone.0207217.g004

Six different sampling procedures were simulated. A brief explanation of the sampling procedures and follows in the section below, and the used sampling procedure-specific parameter values are reported. The parameter values were empirically chosen based on experience and reasonable outcomes for this application and population. Thus, we do not claim that the parameter values result in optimal performance for each sampling procedure. Each sampling procedure was simulated for a length of 60 trials, and each template was simulated 1000 times for each psychophysical procedure, leading to a total of 6×240000 simulated psychophysical experiments.

Even if some adaptive methods directly provide a threshold estimate, they may not always reach convergence within the given maximum number of trials and therefore threshold estimates may not be very accurate. Thus, the threshold and the slope parameters were estimated by fitting $\psi(x)$ to the proportion of correct responses at stimulus levels x using a Maximum Likelihood criterion implemented in the Palamedes MATLAB routines [56]. Moreover, this way the parameter estimation process was identical for all examined methods. The computer simulations, as well as the following data analysis, were performed in MATLAB R2014a (MathWorks, Natick, MA, USA).

4.2.3 Sampling procedures. **4.2.3.1 MOCS** The method of constant stimuli [2] is the simplest sampling procedure, where a set of stimulus levels is predefined and presented multiple times in random order. The set of stimulus levels used in this simulation consisted of 12 levels spaced equally $\in [0.75^\circ, 9^\circ]$. Each level was presented 5 times.

4.2.3.2 Weighted Up-Down In contrast to the Simple Up-Down [7] and the Transformed Up-Down [9] methods, the Weighted Up-Down method proposed by Kaernbach [8] can converge to any desired point on the psychometric function using the equilibrium condition $step_{up}$ $P_t = step_{down}(1 - P_t)$ for the convergence point $x_t = \psi^{-1}(P_t)$. Thus, for a target probability $P_t =$ 75%, the ratio $step_{up}/step_{down}$ results in ${}^{1}/_{3}$. Each correct or incorrect response leads to a decrease, respectively increase, of the stimulus level according to the following rules:

- -3 $step_{unit}$ after 3 correct responses,
- $+1$ $step_{unit}$ after 2 correct and 1 incorrect response,
- $+2$ $step_{unit}$ after 1 correct and 1 incorrect response,
- $+3$ $step_{unit}$ after 1 incorrect response,

where $step_{unit}$ was 0.5° and the start level x_0 was 5.5°.

4.2.3.3 PEST (log) The adaptive procedure called Parameter Estimation by Sequential Testing was introduced by [12]. The desired convergence point (percentage of correct responses) can be selected, no prior assumptions on the subject's psychometric function are required, and it is based on a set of heuristic rules defining step sizes as follows:

- The step size is halved on every direction reversal.
- The first and second step in the same direction are of same size.
- The third step is double the second if the step immediately preceding the last reversal resulted from a doubling, or same otherwise.
- The fourth and additional steps in the same direction are the double of their predecessor.

According to the Wald sequential likelihood-ratio test, the level is maintained if $N_{correct} \in (N_{total} \times P_t \pm W)$ and changed otherwise. $N_{correct}$ and N_{total} correspond to the number of correctly responded trials and the total number of subsequent trials at the same level. A value of $P_t = 0.75$ leads to a convergence towards 75% correct responses in a 2AFC experiment. The deviation limit W of the sequential test was set to $W = 1$, as suggested in [12]. This parameter defines the trade-off between quick (highly dynamic behavior of PEST) and powerful (slower level changes but with higher statistical confidence) decisions. PEST requires two starting parameters: the start level x_0 and the $step_{start}$, which were set to 5.5° and 2°, respectively. Despite PEST having three termination conditions in addition to a total maximum number of trials (i.e., maximum number of consecutive trials at the same level x or a step below a minimum threshold $step_{min}$), they were not used in the present simulations to reach always 60 trials.

PEST is often used in psychoacoustic experiments where auditory stimulus levels are given in dB [4, 12, 19]. However, when PEST is applied to an experiment estimating a joint angle difference threshold, zero crossings of the level x when continuously decreasing the level leads to potential problems and undesired behavior of the algorithm (convergence towards the upper or lower difference threshold, and reduction of efficiency through temporally divergence from the threshold). To address these issues, a logarithmic mapping f: $Stimulus \rightarrow PEST$, $f(x) = \log x$ between the stimulus domain in degrees and the PEST domain was introduced [40]. Consequently, the stimulus levels always remain positive, even if the PEST level performs zero crossings. Because the mapping depends on absolute values, the mapping functions $f(x)$ and $f^{-1}(x)$ cannot be directly applied to a step. Instead, a step has to be regarded as a vector and the mapping functions have to be applied to the initial and terminal points separately, after which the two values are subtracted to define the step length in the specific domain (i.e., *Stimulus*- or *PEST*-domain).

4.2.3.4 SA Staircases (standard) The standard Stochastic Approximation (SA) Staircases [14] can converge to any target probability P_t using asymmetric upward and downward steps. The step size is defined by the following rule and decreases with the number of trials n:

$$x_{n+1} = x_n - \frac{step_{unit}}{n}(z_n - P_t) \quad , \tag{5}$$

where z_n is the binary response (0 incorrect, 1 correct) at trial n. The start level x_0 and $step_{unit}$ were set to 5.5° and 4°, respectively, with $P_t = 0.75$.

4.2.3.5 SA Staircases (accelerated) Kesten proposed an accelerated version of the Stochastic Approximation (SA) Staircases [13]. The first two trials of the procedure are identical to the standard SA Staircase. For the subsequent trials ($n > 2$) the step size is changed only when the response changes according to the following rule:

$$x_{n+1} = x_n - \frac{step_{unit}}{2 + m_{shift}}(z_n - P_t) \quad , \tag{6}$$

where m_{shift} is the number of shifts in response category (i.e., reversals). The start level x_0 and $step_{unit}$ were set to 5.5° and 4°, respectively, with $P_t = 0.75$.

4.2.3.6 PSI method The Ψ method was developed by [15] to estimate the threshold and slope: The posterior probability distribution across the two-dimensional space (threshold and slope) of psychometric functions is updated following Bayes' rule. Subsequently, the psychometric function is estimated by computing the mean of the posterior probability distribution. The next level is defined by a one-step ahead minimum search of an entropy-based cost function in order to optimize information gain. The detailed equations are presented in [15]. The threshold grid used in the simulations was [0.25°, 10°] in steps of 0.125°. The slope ($Slope_\alpha$) grid consisted of 12 logarithmically spaced values from 0.0625/° to 2/°. The guessing rate γ and lapse rate λ were set to 0.5 and 0. Levels were restricted to $x \in [0.25°, 10°]$ in steps of 0.125°.

4.3 Results

4.3.1 Example sequences. A set of procedure sequence examples (from one simulation run) for the six different tested procedures illustrating stimulus placement and adaptiveness (where applicable) are shown in Fig 5 in combination with the template function $\psi^{\top}_{5^\circ,0.25/^\circ}$ and the resulting estimated psychometric function.

4.3.2 Constant and variable errors of threshold and slope estimates. For each psychometric template and procedure, the *CE* and *VE* were calculated for the estimation of the threshold (Figs 6 and 7) and the slope (Figs 8 and 9). In these figures, it can be observed that for all sampling procedures except MOCS the absolute value of the threshold *CE* was below 0.1° for the largest part of the threshold and slope parameter space. For the PSI method, the *CE* was around 0.01°. In the case of MOCS and Weighted Up-Down, the threshold *CE* presented ripples depending on the threshold axis with absolute biases up to 0.4° and 0.2°, respectively. All methods showed decreasing performance towards the boundaries of the parameter space, especially for the smallest slopes. The standard SA Staircases showed a large negative bias for increasing thresholds and decreasing slopes (*CE* up to −1.5°) and large positive bias for low thresholds and small slopes (*CE* up to 1.2°). Similar but less pronounced effects could be found for the accelerated version of the SA Staircases. For the PSI method, the *CE* increased up to 0.8° for low thresholds.

The *VE* of the threshold showed much less consistent results compared to the *CE* with more oscillations along the threshold axis. The *VE* increased severely for small slopes in all methods. For thresholds larger than around 6°, *VE* started to increase non-monotonically to more than 1° for both standard and accelerated SA Staircases. For MOCS and the PSI method this was only the case for thresholds larger than 9°. The Weighted Up-Down method was the only sampling procedure not showing effects depending on the threshold value, neither for the central region of the parameter space, nor for the boundary. The PSI method showed the lowest *VE* of around 0.1° for almost the entire parameter space, compared to the other sampling procedures.

The slope *CE* showed an overestimation of the slope in the whole parameter space for all simulated sampling procedures. Similar to the threshold *CE*, slope *CE* ripples were found for MOCS and Weighted Up-Down. Both SA Staircases methods showed increasing bias for large thresholds and small slopes, whereas the standard SA Staircases also showed an increased bias for low thresholds. MOCS, Weighted Up-Down and PEST showed increasing overestimation for lower slopes, but bias decreased again for slopes smaller than 0.5/°. The most homogeneous *CE* across the parameter space (around 0.1) could be found for the PSI method and it was the only procedure for which bias decreased monotonically with smaller slopes. For illustration, a *CE* of 0.2 corresponds to an estimated slope of 1/° for a slope template of 0.5/°.

The slope *VE* showed similar behavior across the parameter space as the slope *CE*. For all methods the *VE* lay between 0.15 and 0.3. The main difference compared to the slope *CE* was the decreasing variability for small thresholds for MOCS and the standard SA Staircases.

4.3.3 Percentage within bounds. The *PCTw/iB* curves are presented for each sampling procedure in Fig 10 for the threshold estimates and Fig 11 for the slope estimates. In general, the *PCTw/iB* curves for the threshold estimates followed the shape of exponential cumulative density functions. All methods had in common that for smaller slopes the *PCTw/iB* curve increases less rapidly, and the variability across thresholds was lower compared to the variability across different slopes. PEST, the PSI method, and the accelerated SA Staircases demonstrated the best performance and smallest variability across thresholds (e.g., reaching up to between 80% and 90% for a slope of 2/° and a threshold tolerance of ±0.1°). However, for slopes of 0.0625/° even these methods achieved only between 60% and 70% within a tolerance

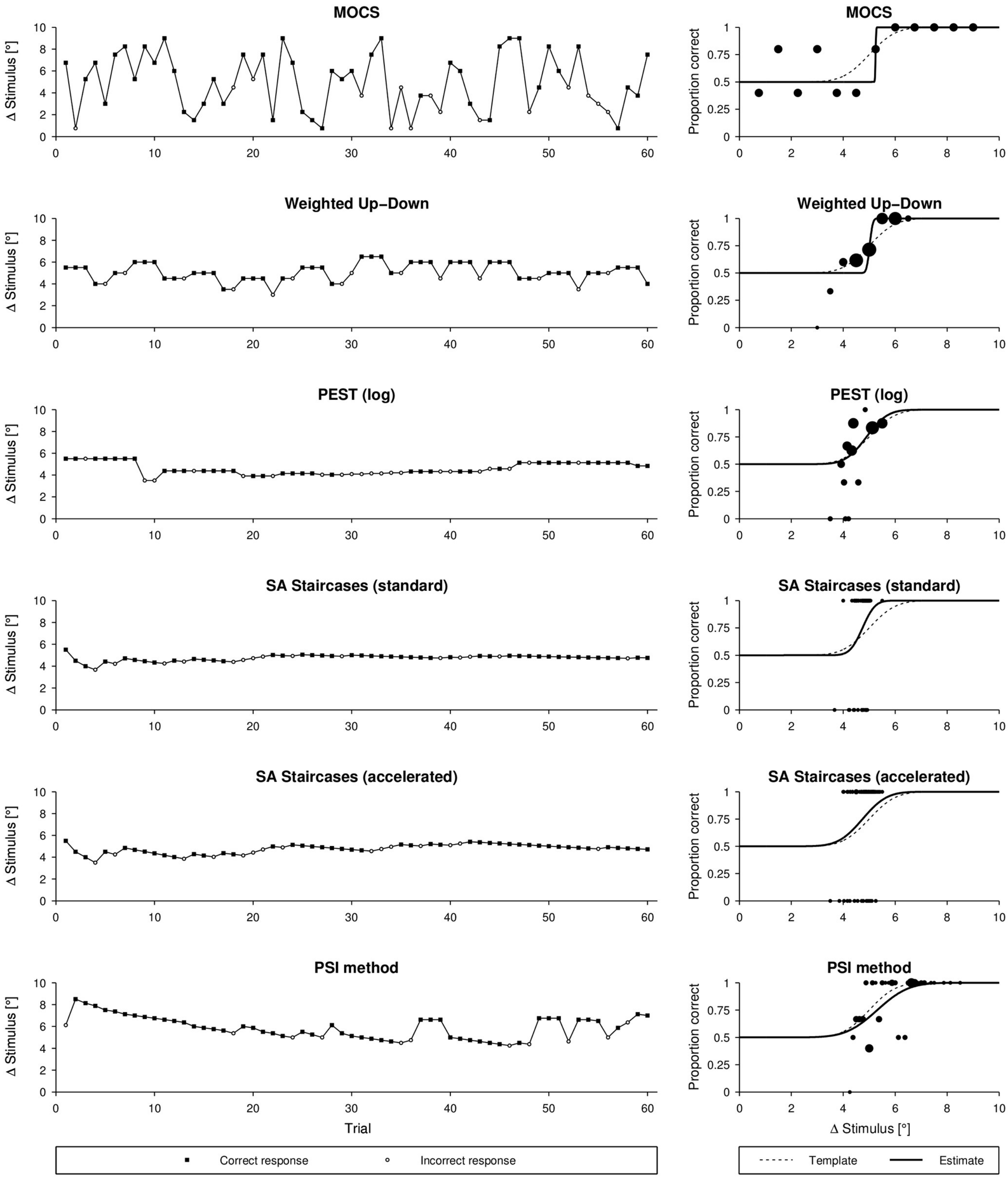

Fig 5. Examples of simulated sequences and psychometric functions for different sampling procedures. Examples for a template $\psi^{\top}_{5^{\circ},0.25/^{\circ}}$ and comparison of the psychometric function of the template with the resulting fit using a Maximum Likelihood criterion. The size of the black dots indicates the number of repetitions at same stimulus level.

https://doi.org/10.1371/journal.pone.0207217.g005

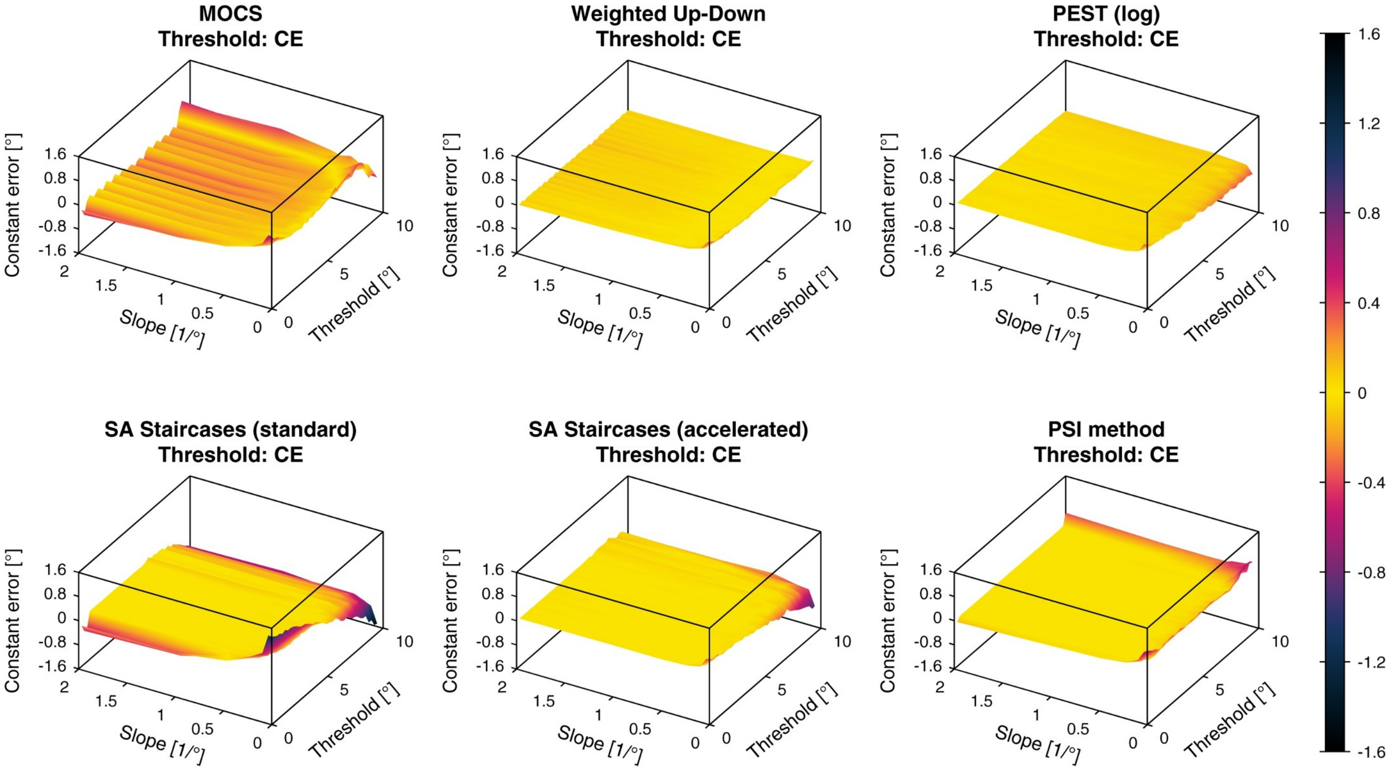

Fig 6. Constant errors of threshold estimates. Constant errors (*CE*) of the threshold estimates for the six different simulated sampling procedures. The color bar indicated the *CE* values in (°).

https://doi.org/10.1371/journal.pone.0207217.g006

interval of ±1°. On the other hand, MOCS and standard SA Staircases showed the largest variability and lowest ratio of estimates for given tolerance bounds.

In contrast to the *PCTw/iB* curves for the threshold estimates, the derivative of the *PCTw/iB* curves did not monotonically decrease for the slope estimates. For all procedures except both SA Staircases, the *PCTw/iB* for small slope tolerance intervals was higher for smaller slopes compared to larger slopes, but was the other way around for larger intervals (i.e., larger than ±0.2). In the case of both SA Staircases, *PCTw/iB* was higher for larger slopes independently of the slope tolerance interval. The PSI method showed the best overall performance and reached 80% for all slopes for a tolerance of ±0.3. As for the threshold *PCTw/iB*, MOCS and standard SA Staircases showed the largest variability.

The tolerance-dependent overall *PCTw/iB* curves are presented for each sampling procedure in Fig 12 for the threshold estimates and Fig 13 for the slope estimates. For the threshold estimates, overall *PCTw/iB* reached 80% for all methods except MOCS for a tolerance interval of ±0.3°. PEST, accelerated SA Staircases, and the PSI method showed similar performance and outperform the others. The overall *PCTw/iB* for the slope was the highest for the PSI method and the accelerated SA Staircases, mostly independent of the slope tolerance interval size.

4.3.4 Normalized area under the curve. The *nAUC* values computed individually for each combination of the two-dimensional parameter space are shown as surfaces for each sampling procedure in Fig 14 for the threshold estimates and Fig 15 for the slope estimates. These *nAUC* surfaces reflected the area under the *PCTw/iB* curves, showing a decrease of *nAUC* for smaller slopes for threshold estimates across all methods. In addition to Fig 10 the *nAUC*

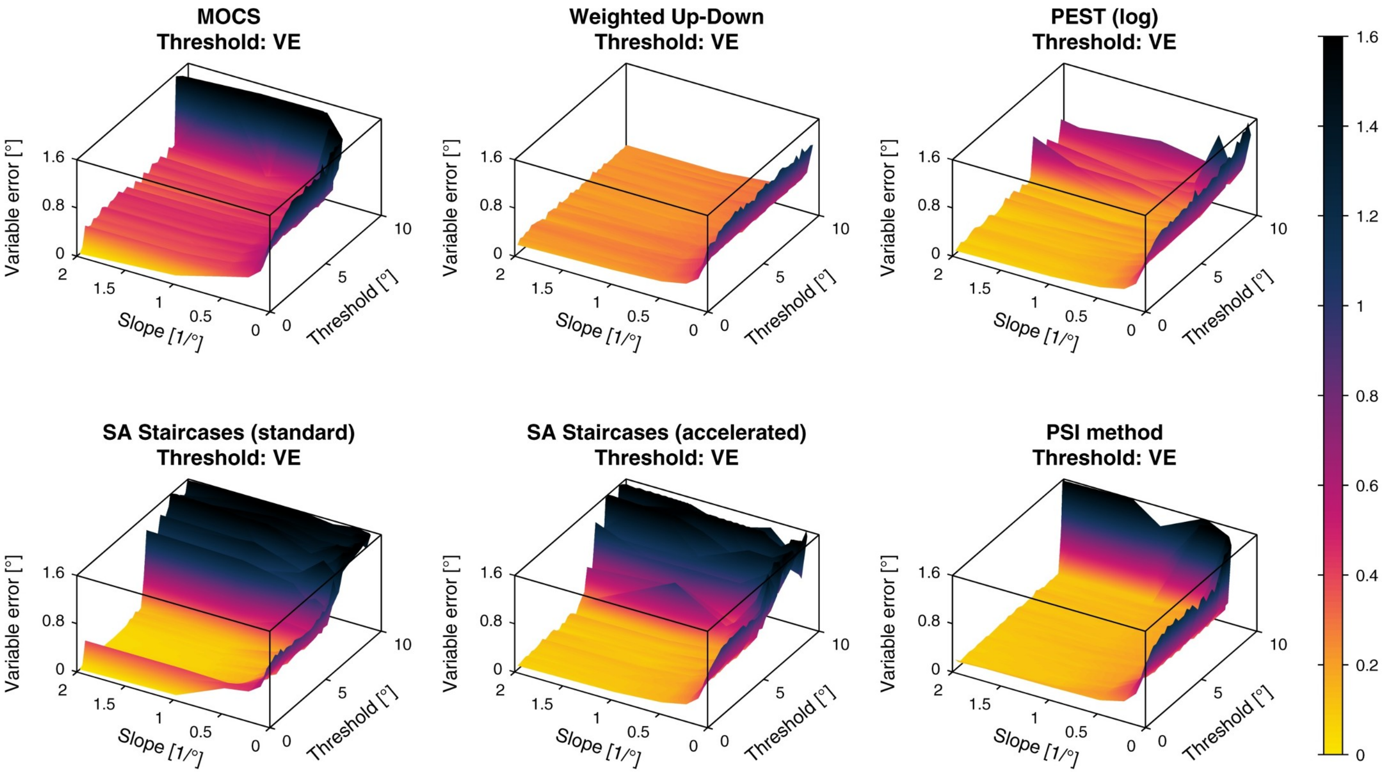

Fig 7. Variable errors of threshold estimates. Variable errors (*VE*) of the threshold estimates for the six different simulated sampling procedures. The color bar indicated the *VE* values in (°).

https://doi.org/10.1371/journal.pone.0207217.g007

surfaces revealed threshold-dependent differences in performance, such as visible for both versions of the SA Staircases. For the slope estimates, both versions of the SA Staircases showed monotonically decreasing *nAUC* for smaller slopes, whereas in particular MOCS and Weighted Up-Down showed a large increase of performance for slopes below 0.5/°.

4.3.5 Normalized volume under the surface. The *nVUS* and inhomogeneity parameter σ are listed for each sampling procedure in Fig 14 for the threshold estimates and Fig 15 for the slope estimates. To compare the overall performance of the methods, these metrics are presented as ellipses (with center *nVUS* and half-axis σ, for both threshold and slope) in Fig 16. For the threshold estimates, PEST, accelerated SA Staircases, and the PSI method performed almost identically in *nVUS* (around 0.88) and σ (around 0.08). The overall best slope estimates were provided by the PSI method (*nVUS* = 0.85, $\sigma = 0.03$) followed by the accelerated SA Staircases. MOCS showed by far the worst performance for both threshold (*nVUS* = 0.72, $\sigma = 0.07$) and slope (*nVUS* = 0.72, $\sigma = 0.07$) estimates, and the standard SA Staircases showed the largest inhomogeneities (threshold: $\sigma = 0.12$, slope: $\sigma = 0.08$).

4.4 Discussion

The aim of this work was to introduce metrics to quantify the performance of psychophysical sampling procedures in application-driven simulations. The usefulness of the proposed metrics for in-depth analysis to identify how the procedures or their parameters could be tuned to potentially improve estimates and for choosing the best procedure for a specific application given some requirements (e.g., tolerance interval for the estimates) was illustrated using

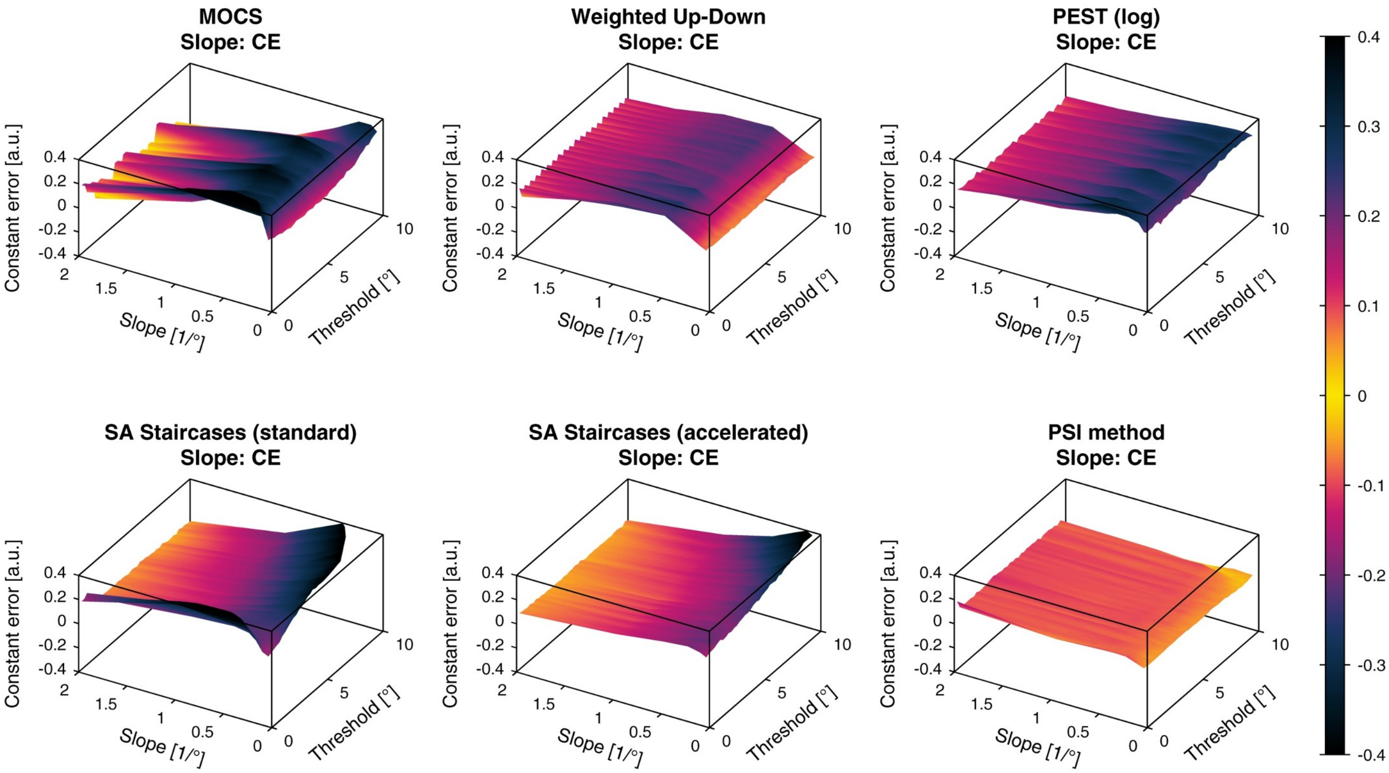

Fig 8. Constant errors of slope estimates. Constant errors (*CE*) of the slope estimates for the six different simulated sampling procedures. The color bar indicated the *CE* values in (a.u.).

https://doi.org/10.1371/journal.pone.0207217.g008

computer simulations. These included the simulation of six different sampling procedures using values and ranges from a real-world application example.

4.4.1 Strengths of this analysis framework. The efficiency in terms of variability multiplied by the number of trials [12, 36] has been evaluated for many sampling procedures using computer simulations [8, 11, 12, 18–20, 32–35]. This approach aims at evaluating the benefit of adding more trials for better estimates. While this may be beneficial in some applications or for benchmarking, this analysis approach is incomplete and of limited use for real applications, where a sampling procedure should be selected and tuned to a specific distribution of psychometric functions of the population to be assessed. Furthermore, the performance of sampling procedures is often investigated for a single psychometric function or for a very limited number of threshold and slope parameters [8, 11, 18, 20, 33–35, 38]. As a matter of fact, this definition of efficiency corresponds to the square of the *VE* multiplied by a constant number for a given number of trials. Thus, when computed for a specific psychometric function, the efficiency would correspond to a single point on the *VE*-surface plots (e.g., Fig 7).

As shown in the present work, the performance of sampling procedures can vary considerably for different parameter values of psychometric functions. Therefore, results may not be representative for other values. As it has been noted by Klein: "*It is always a good idea to carry out Monte Carlo simulations of one's experimental procedures, looking for the unexpected.*" [57]. With the presented analysis method we could demonstrate the importance of simulating a wide range of psychometric functions covering the full parameter space of interest (i.e., threshold and slope) and illustrate the benefits of the different metrics for selecting and improving psychophysical sampling procedures.

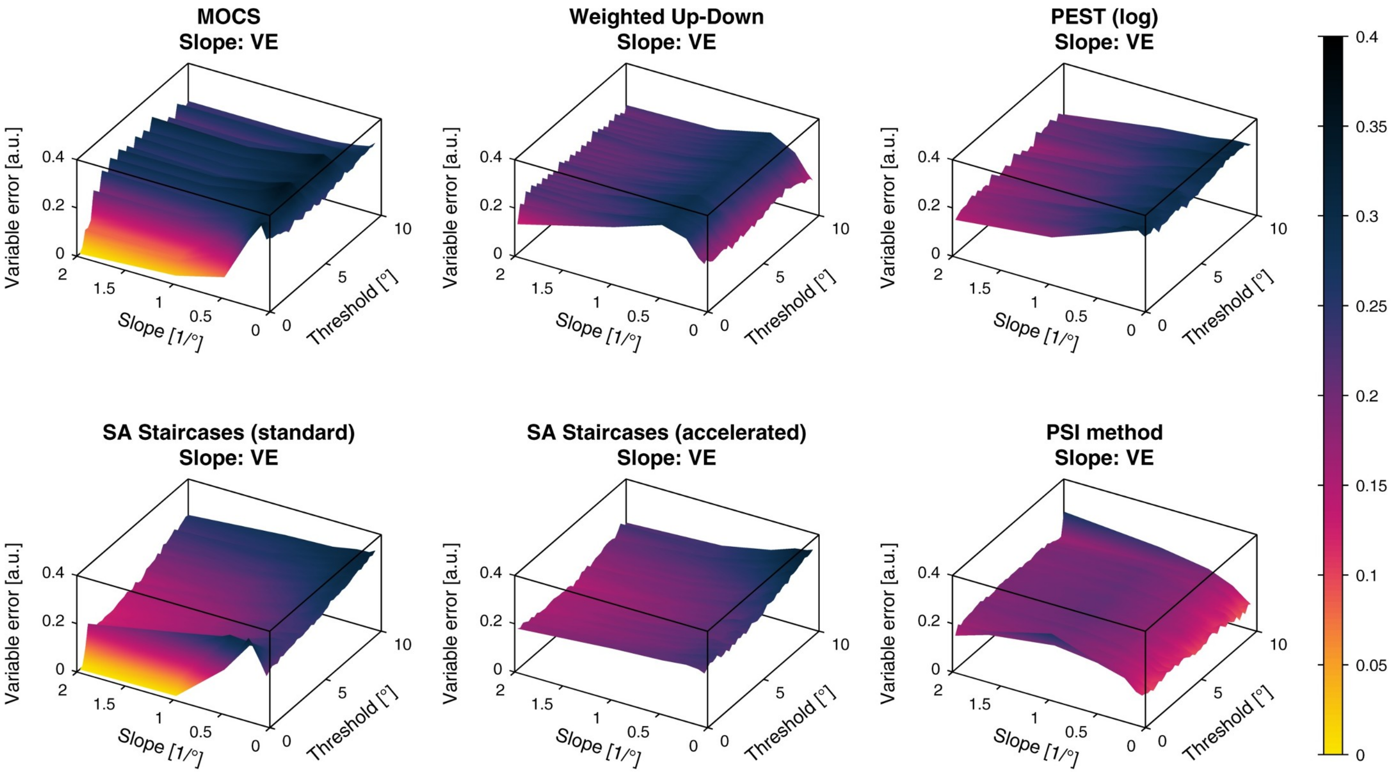

Fig 9. Variable errors of slope estimates. Variable errors (*VE*) of the slope estimates for the six different simulated sampling procedures. The color bar indicated the *VE* values in (a.u.).

https://doi.org/10.1371/journal.pone.0207217.g009

The raw errors can describe bias (*CE*) and variability (*VE*). Their representation in the parameter space allows to detect suboptimal performance, for example, due to poor selection of parameters of a sampling procedure or decreases in performance towards the boundaries of the parameter space. By applying the transformation to arbitrary units on the slopes, errors in slope estimates can also be calculated and analyzed the same way, without penalizing slope errors in steep psychometric functions. For applications where a certain tolerance of estimation errors can be accepted, a new metric (*PCTw/iB*) based on the absolute errors (*AE*) was introduced. This metric is useful to assess the probability of the estimated parameter falling into a defined tolerance interval, and can be plotted for different psychometric functions. One advantage of the *PCTw/iB* compared to the average *CE* and *VE*, or the sweat factor [12, 36] is that this metric is robust against large outliers. As is the case in Fig 10, the plots can show very poor performance for small slopes. However, this visual representation of simulated templates may be misleading when deducing the overall performance. This has been addressed by post-hoc linear resampling of the parameter space of $PCTw/iB_{\pm Tol}$ for each tolerance, from which the overall performance (tolerance-dependent overall *PCTw/iB*) can be calculated (Fig 12). This plot reveals that the overall performance for this parameter space is higher than what could have been misinterpreted based on Fig 10. Thus, despite large threshold estimate variability for small slopes, the best simulated methods provide around 80% of the threshold estimates within a tolerance interval ±0.2°, which in the present application is around one order of magnitude smaller than the proprioceptive difference thresholds of healthy subjects (ranging from around 1° to 5°) [40, 46–48, 50–53]. The *PCTw/iB* is similar to the concept of a "usability index" introduced by [37], describing the percentage of times that a sampling

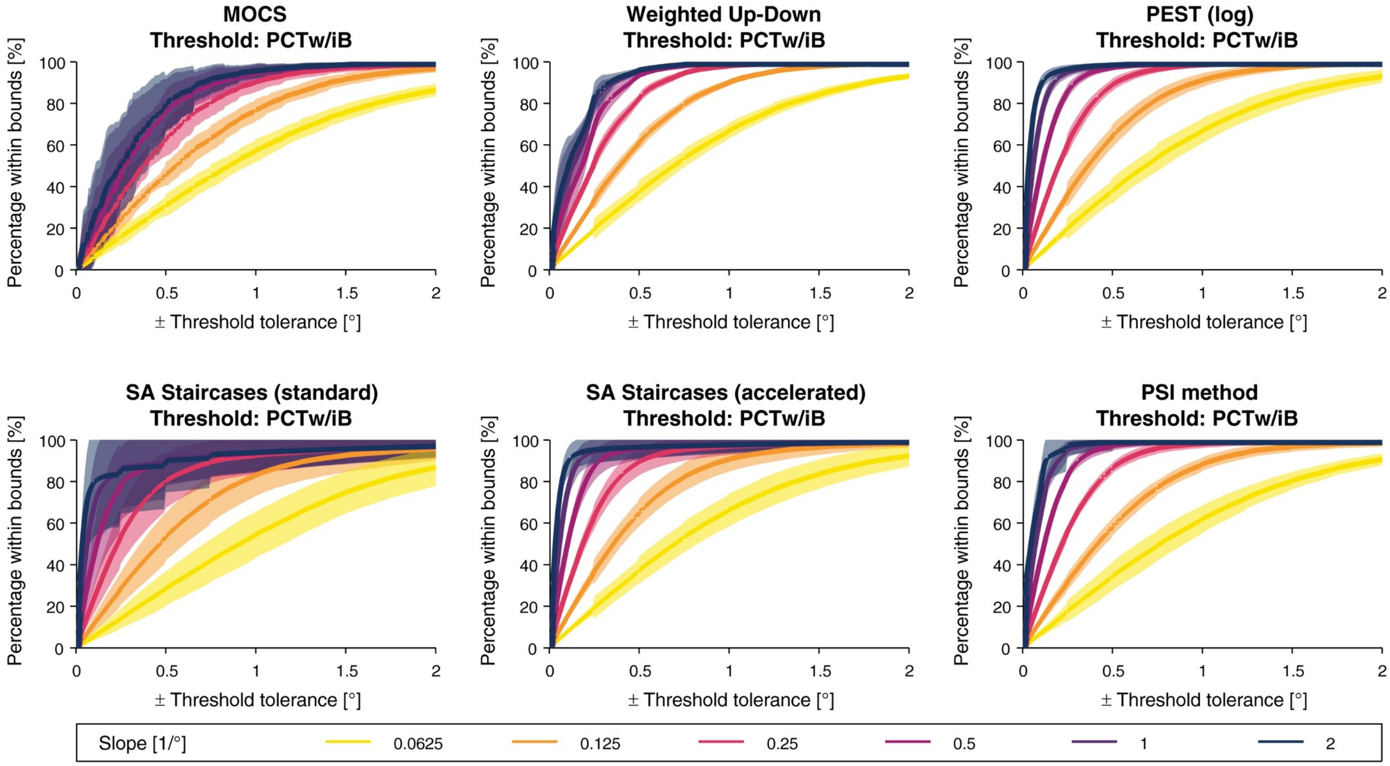

Fig 10. Percentage within bounds for the threshold estimates. For each slope value of the template parameter set, the *percentage within bounds* (*PCTw/iB*) function for threshold estimation are shown with the mean (bold line) and standard deviation (shaded band) across all modeled thresholds.

https://doi.org/10.1371/journal.pone.0207217.g010

procedure produces "usable data". While authors present the "usability index" as a function of number of trials, our metric is a function of the acceptable tolerance and has a relation to the specific application examples. To deduce higher-level overall performance metrics, the *nAUC* surface (based on the integral of the normalized *PCTw/iB*) can be computed. Since the *nAUC* is remotely based on the *AE*, which is a composite of *CE* and *VE*, the *nAUC* surfaces are similar to the inverse of the *CE* and *VE* surfaces, but normalized to [0, 1]. Whereas the *nAUC* is calculated for each simulated psychometric function, the general performance metric (*nVUS*) and inhomogeneity parameter (σ) are independent of the parameter space, and can be used to select the optimal sampling procedure and to compare different sets of sampling procedure-specific parameters. The latter two metrics can, in case of threshold and slope to be estimated, be visualized as ellipses allowing a summarizing view on the performance for both parameters (Fig 16). Furthermore, this analysis framework is independent of the parametric version of the psychometric function and paradigm used, and can thus also be used, for example, for yes-no experiments with psychometric functions with $\gamma = 0$. Moreover, it can also be extended to the estimation of the lapse rate.

4.4.2 Comparison of psychophysical sampling procedures for the application example. Before discussing differences in the performance between sampling procedures, it should be noted that the choice of the parameters of the sampling procedures was based on experimental experience and that they were not systematically optimized to achieve the highest possible performance. It is, therefore, not possible to make claims about consistent superiority of any sampling procedure, and results are only an example of performances that could be

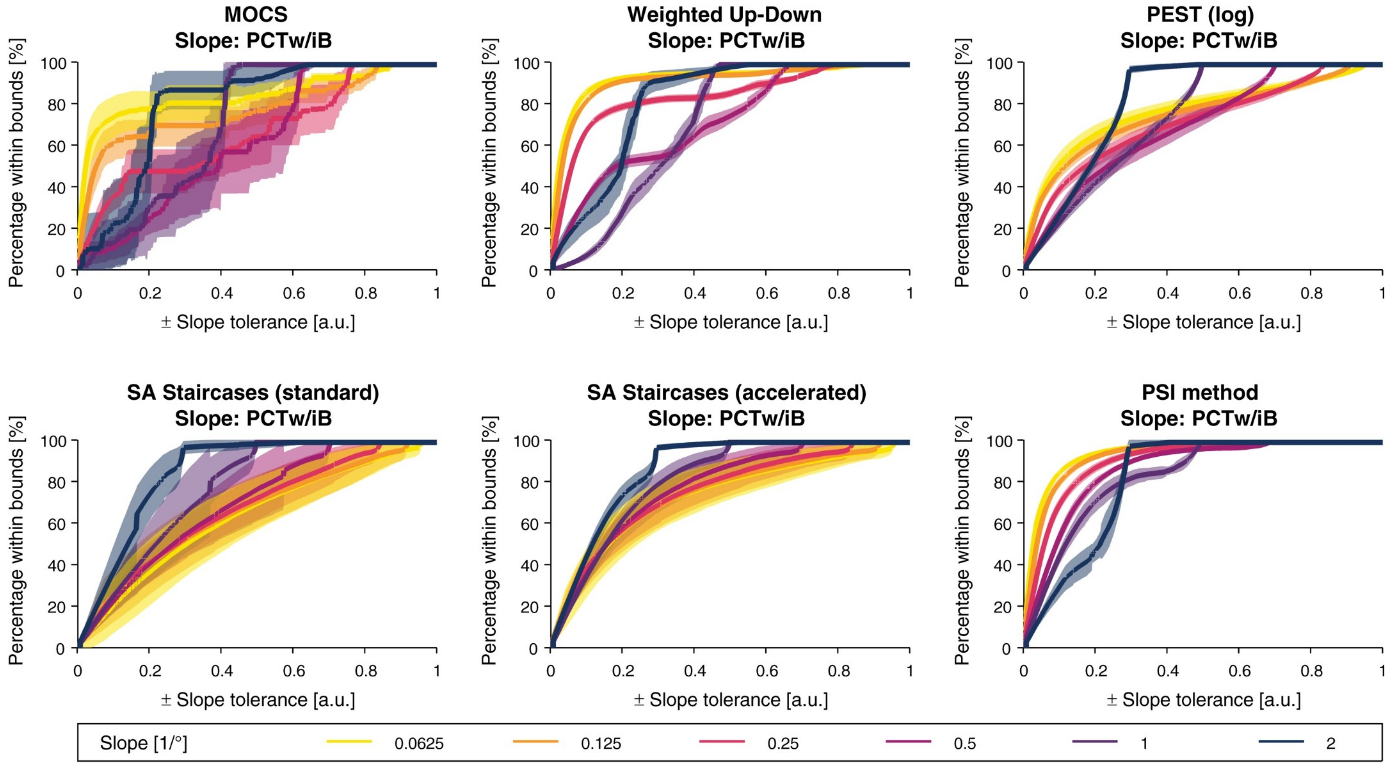

Fig 11. Percentage within bounds for the slope estimates. For each slope value of the template parameter set, the *percentage within bounds* (*PCTw/iB*) function for slope estimation are shown with the mean (bold line) and standard deviation (shaded band) across all modeled thresholds.

https://doi.org/10.1371/journal.pone.0207217.g011

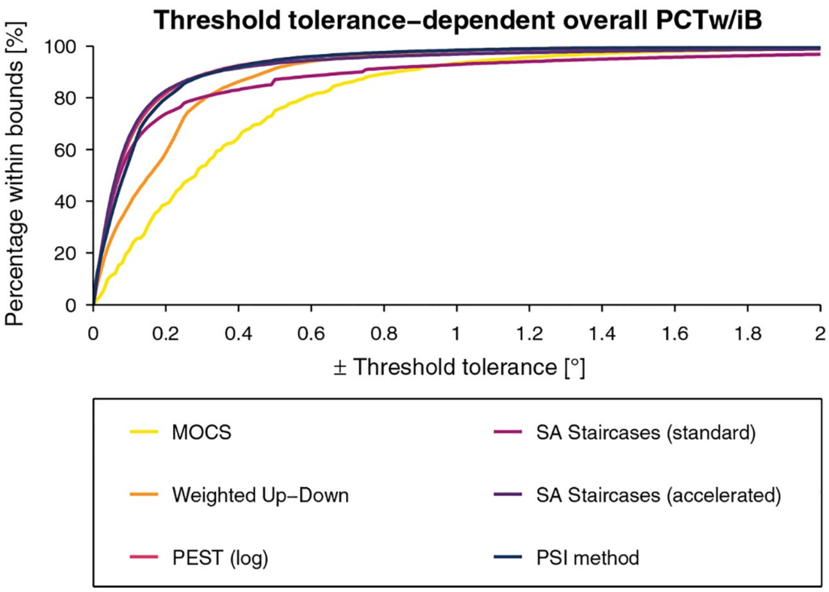

Fig 12. Overall percentage within bounds for the threshold estimates. Overall *percentage within bounds* (*PCTw/iB*) function for threshold estimation for a linearly resampled parameter space.

https://doi.org/10.1371/journal.pone.0207217.g012

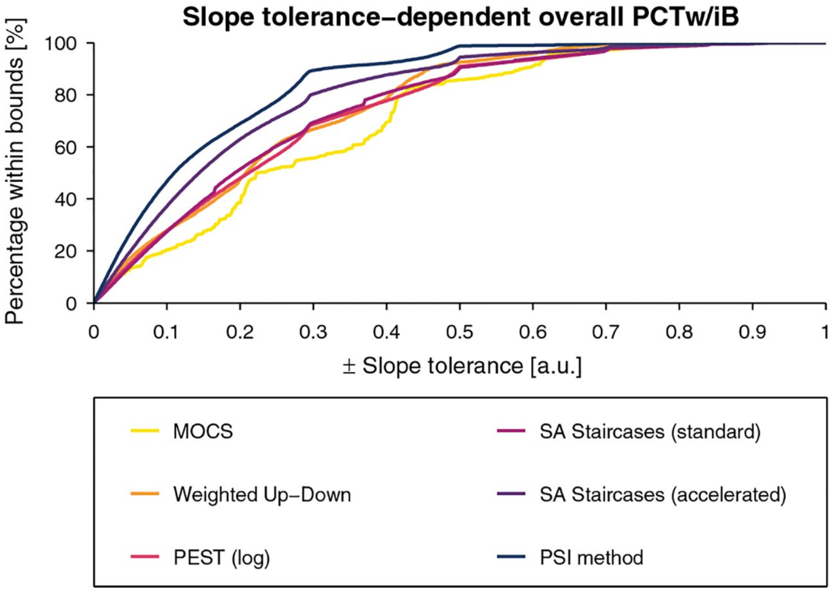

Fig 13. Overall percentage within bounds for the slope estimates. Overall *percentage within bounds* (*PCTw/iB*) function for slope estimation for a linearly resampled parameter space.

https://doi.org/10.1371/journal.pone.0207217.g013

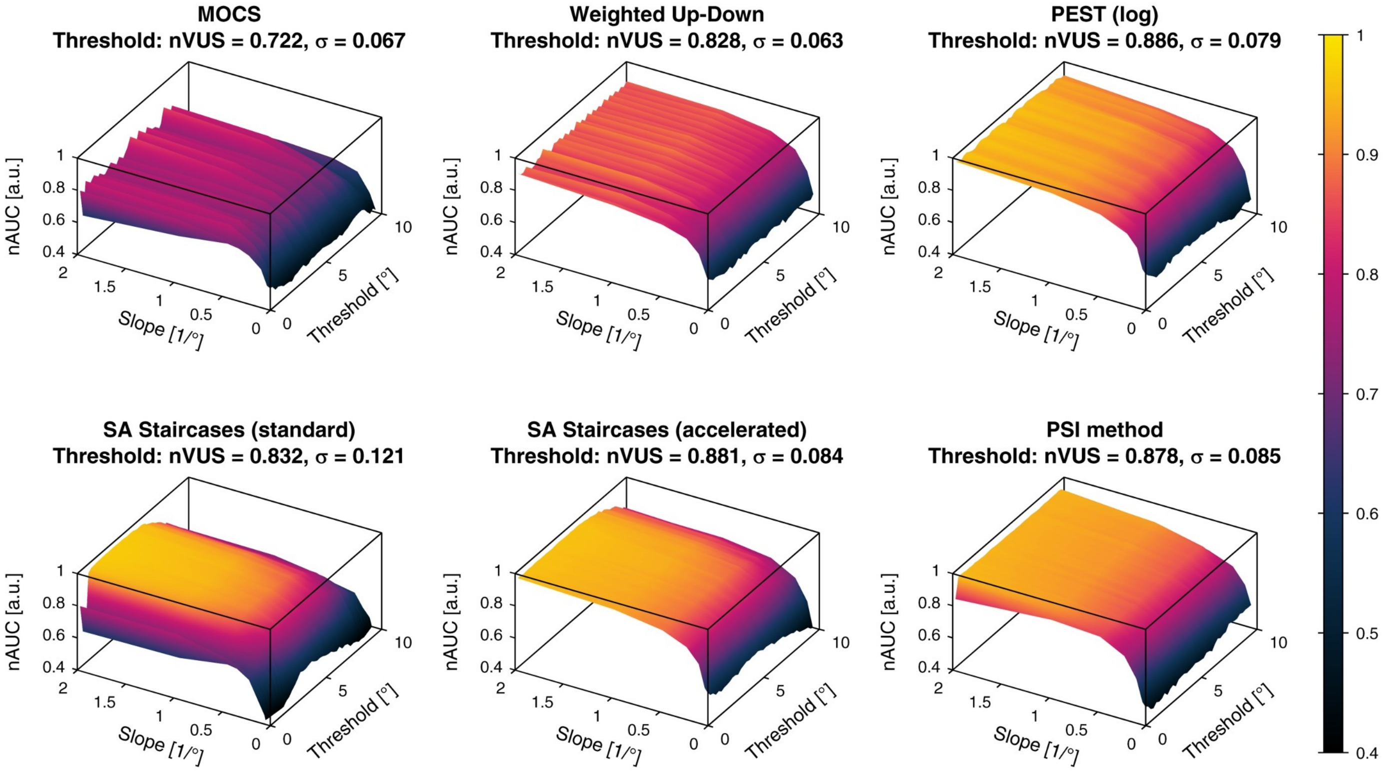

Fig 14. Normalized area under the curve for the threshold estimates. The *normalized area under the curve* (*nAUC*, represented by surfaces for the parameter space) and corresponding *normalized volume under the surface* (*nVUS*) and inhomogeneity parameter *σ* are shown for the threshold estimation. *nAUC* and *nVUS* values of 1 correspond to a perfect threshold estimation. The color bar indicates the *nAUC* values in (a.u.).

https://doi.org/10.1371/journal.pone.0207217.g014

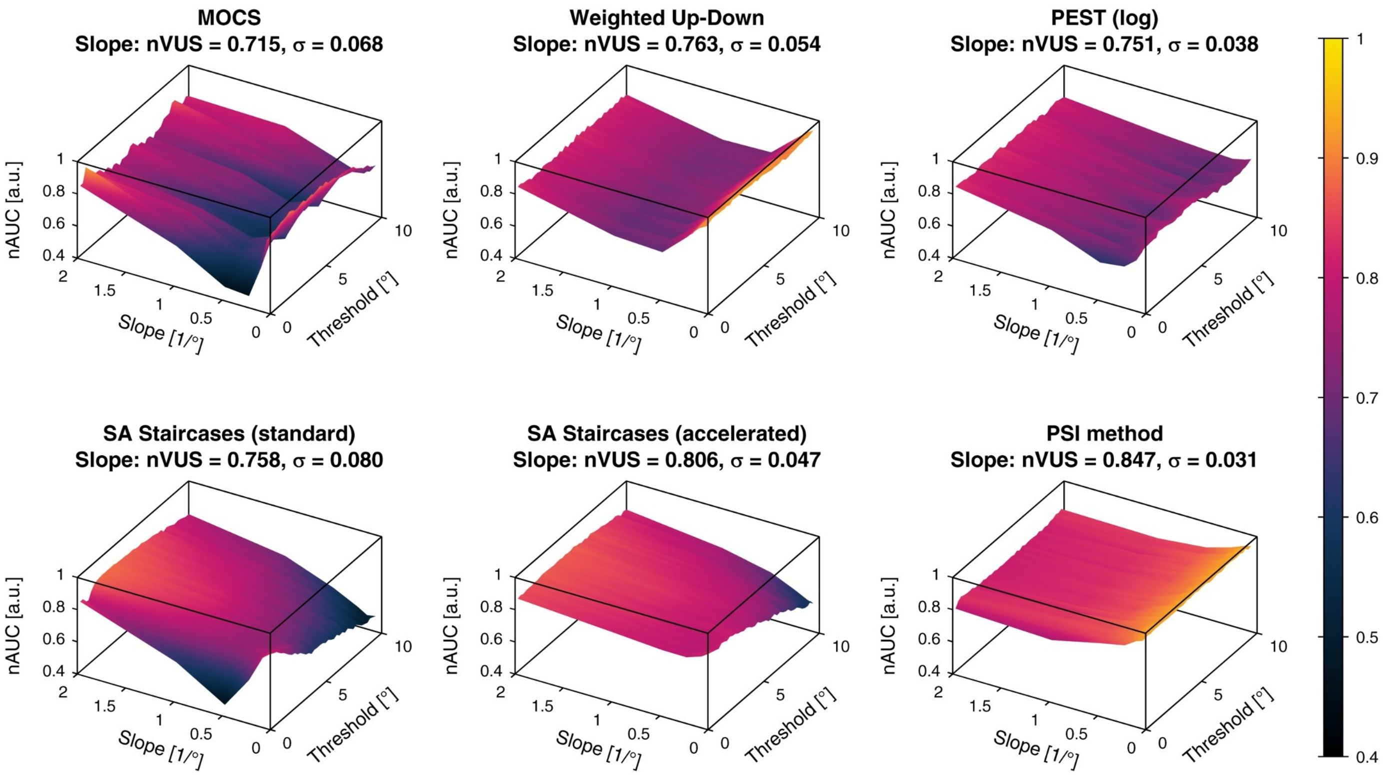

Fig 15. Normalized area under the curve for the slope estimates. The *normalized area under the curve* (*nAUC*, represented by surfaces for the parameter space) and corresponding *normalized volume under the surface* (*nVUS*) and inhomogeneity parameter *σ* are shown for the slope estimation. *nAUC* and *nVUS* values of 1 correspond to a perfect slope estimation. The color bar indicates the *nAUC* values in (a.u.).

https://doi.org/10.1371/journal.pone.0207217.g015

obtained. However, they serve well to illustrate insights, which can be gained by using the proposed analysis framework, and identify actions to address limitations of sampling procedures.

In this application example and given the chosen parameters of the sampling procedures, PEST provides the best overall threshold estimates (i.e., highest *nVUS*), as visible in Fig 16. Yet, the PSI method and the accelerated SA Staircases perform similarly well to PEST and distinctively better than the other sampling procedures. This can be confirmed in Fig 12 showing the tolerance-dependent estimation performance. Interestingly, the accelerated SA Staircases performs better than, for example, the Weighted Up-Down method, despite large *VE* and increased bias (*CE*) for high thresholds, whereas the latter shows more constant performance across the threshold parameter space (Fig 7). These high threshold *VE* oscillations (appearing for both SA Staircases) exist only for thresholds larger than the start level and may arise from the asymmetric descending and ascending step sizes. A similar, yet not so strong, effect can be noticed for PEST. Since with the logarithmic version of PEST steps become larger when ascending towards higher levels, it requires more trials to converge towards a threshold with the same precision. However, as the number of available trials is limited by administration time requirements of this application, variability of the estimates increases. Thus, a recommendation to improve the performance of these three procedures would be to select a higher start level and converge towards the threshold with descending steps. The reason for the Weighted Up-Down method not showing such decreases in performance for high thresholds, in contrast to the other adaptive methods, may be the constant step size which is well chosen with regards to the parameter space of the simulated templates. Unsurprisingly, the estimation performance

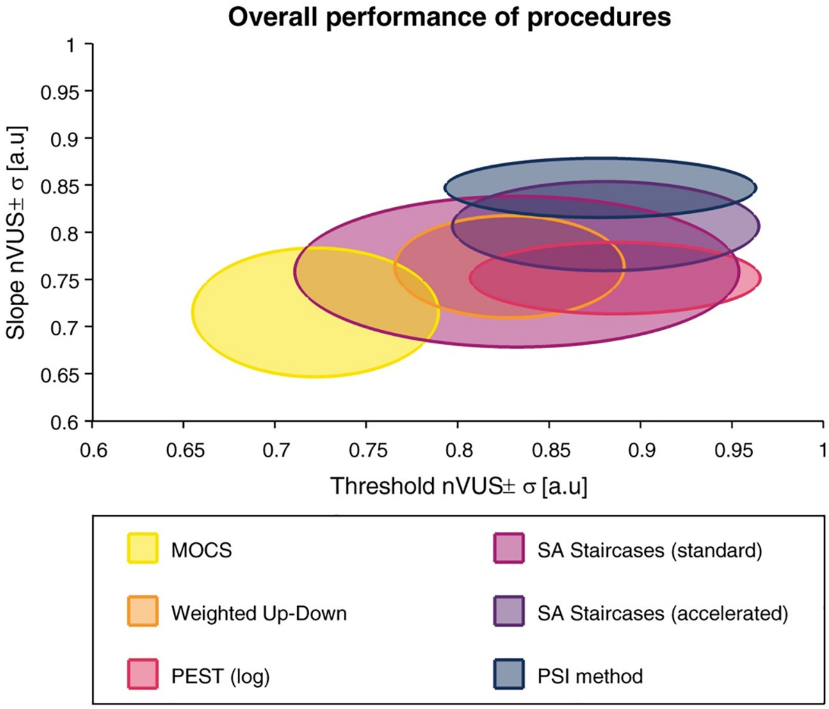

Fig 16. Overall performance of the six procedures. Two-dimensional plot showing the overall performance of the six methods based on the *normalized volume under the surface* (*nVUS*) and inhomogeneity parameter *σ* for the threshold and slope. The ellipses are centered at *nVUS* and the half-axes correspond to *σ*. Higher *nVUS* values and smaller half-axed indicate better estimation performance.

https://doi.org/10.1371/journal.pone.0207217.g016

of MOCS shows ripples in both *CE* and *VE* (Figs 6 and 7). This reflects the limited number of different stimulus levels and presentations leading to equal estimations of psychometric functions for similar templates creating error ripples depending on the threshold. Both MOCS and the PSI method present increased *VE* at the boundary of the parameter space (i.e., high thresholds). This is attributable to the boundaries of the grid-like stimulus levels and the grid of psychometric functions used for the posterior probability distribution of the PSI method. As, for example, in the case of MOCS, the maximum stimulus level presented is 9˚. If the threshold of the template to be estimated is higher than this maximum level, there is no data at levels with high performance available for the Maximum Likelihood fit of the psychometric function. Furthermore, each level is presented only five times, resulting in a discretization of the proportions of correct responses with steps of 0.2. This is not sensitive enough to properly estimate the psychometric function just based on data from low performance levels. Thus, to improve MOCS and the PSI method for this application, the threshold grids should be expanded beyond the parameter space, which would resolve the issues at the boundary. Across all sampling procedures, the threshold estimation performance decreases when the psychometric functions to be estimated have a small slope. This is an inherent problem of psychophysical assessments, as a slowly rising psychometric function increases uncertainty and thus introduces higher variability in answers. Moreover, it becomes more difficult for threshold tracking methods to converge towards the threshold. Therefore, all sampling procedures perform poorly for small slopes, as visible in Fig 10.

While PEST, accelerated SA Staircases, and the PSI method perform similarly well in estimating thresholds, the PSI method outperforms the other methods in estimating the slope of the psychometric function (Figs 16 and 13). The PSI method was also in particular developed to estimate both threshold and slope [15]. Thus, this sampling procedure places the stimulus levels in a way to optimize the estimation of both parameters (i.e., also at levels of high and low percentage of correct responses). This is also reflected in the "jumps" of the stimulus levels of the sequence example shown in Fig 5. However, it should be noted that, when choosing new stimulus levels, the PSI method can take advantage of using the parametrized psychometric function, which in simulations often follows the same parametrization as the psychometric function generating the responses (i.e., the psychometric function to be sampled). As a consequence, the PSI method might have an advantage that may result in artificially inflated performance. To test how much the results are affected, the presented metrics framework could again be used. In contrast to the PSI method, the other adaptive methods (in particular PEST and both SA Staircases) show faster asymptotic behavior towards the threshold to be estimated. The exploration of the stimulus levels with the latter sampling procedures depends solely on the selected start level and the initial step size, and is thus not optimized to cover a wide range of the psychometric function. The accelerated SA Staircases procedure is more aggressive than the standard version (because the denominator in the equation for calculating the new level is smaller, resulting in larger steps) and thus leads to more oscillations around the threshold, which helps to estimate the slope. Therefore, if the slope is of interest, a sampling procedure which places trials at well separated levels should be used [57]. However, if the differences between the levels are too large in comparison to the slope (which can be the case for MOCS and the Weighted Up-Down method if sampling procedure-specific parameters are not well chosen), steep slopes are difficult to estimate because the sampling procedure may not place levels in the "slope region" of the psychometric function. In addition, due to its non-adaptiveness (i.e., predefined and fixed stimulus levels), MOCS places too many trials in regions of little information content (i.e., of very low and very high percentage of correct responses) (Fig 5). Besides the proportion of correct responses suffering from low resolution, these trials do not significantly contribute to the Maximum Likelihood fit of the psychometric function, as by definition of the psychometric function $\psi(x)$ the asymptotic values are already defined (i.e., 0.5 and 1). Thus, depending on whether the slope of the psychometric function to be estimated is small or steep, non-adaptive and adaptive fixed-grid-level procedures (e.g., MOCS, Weighted Up-Down) may perform better than threshold tracking procedures (e.g., PEST, standard and accelerated SA Staircases), and vice versa, as can be inferred from Fig 11. The PSI method can also be considered an adaptive fixed-grid-level procedure. While the variability of the slope estimates improves for small slopes, the differences are small across the parameter space, as shown by the small value of the inhomogeneity metric σ. Similar to the threshold estimates, MOCS shows threshold-dependent ripples in the performance of the slope estimation, also resulting in large inhomogeneity, due to suboptimal and coarse distribution of stimuli.

Many of our findings are in line with the literature, though the results from the simulations are difficult to compare quantitatively due to a large number of different variants of sampling procedures, analysis methods mostly focusing on overall efficiency, and other sampling procedure-specific parameter values and applications. In the literature, most sampling procedures were compared to MOCS, and all discourage from using MOCS, as it is inefficient (i.e., suffers from higher variability) compared to adaptive methods [35, 37, 58] and provides more biased results [59]. Especially in scenarios where there is no prior information about the parameters of the psychometric function to be estimated (which is often the case in clinical assessments with a large range of conditions), stimuli may (initially) be placed far from the region of interest. Methods using stochastic approximation have also been investigated, with the accelerated SA Staircase procedure showing better performance compared to the standard version and to fixed step size staircase methods [38]. An additional advantage of the SA Staircase procedures is that no assumptions of shape or parameters of the psychometric function are required, and is thus less susceptible to parameter mismatches [5, 38]. The accelerated SA Staircase procedure was also shown to have a similar performance as mean-Bayesian methods, which seem to have near-optimal performance [5]. As in the present work, it was suggested to use a clear suprathreshold initial stimulus intensity (i.e., level at high proportion of correct responses) and a large initial step size, which should lead to a good performance independent of the slope of the psychometric function [38]. If the full shape of the psychometric function is desired, fast threshold tracking methods with decreasing step sizes should be avoided [37]. This could also be shown in the present simulations, although it is mostly valid for small slopes only.

4.4.3 Limitations. One limitation of this analysis framework is the need to representatively sample the parameter space and acquire enough information to observe the behavior of different sampling procedures (e.g., threshold-dependent ripples induced by MOCS) as well as to run enough repetitions of the simulations to obtain representative results. As simulations have been run only 1000 times for each template and sampling procedure, the computed metrics need to be regarded as stochastic variables, and their non-deterministic nature may compromise the interpretation of the results. However, using a jackknife resampling technique (computing a distribution of metric values by systematically leaving out each simulation run once) it could be shown that the maximum standard deviation of threshold $nAUC$ across all templates remained below 0.001 for all six sampling procedures. This is around two orders of magnitude smaller than the inhomogeneity parameter σ, proving that 1000 simulation runs are sufficient.

Since the shape of the psychometric function may not follow the mathematical model $\psi(x)$ used when fitting the psychometric function, this might introduce systematic errors. However, this is a general concern of hybrid procedures (using heuristic adaptive sampling procedures combined with fitting of parametric psychometric functions) and methods assuming a particular shape for the stimulus selection. Nevertheless, the presented metrics could also be used with estimates resulting directly from the adaptive sequence instead of estimates from the fitting process. Depending on the perception modality assessed and the sampling procedure, these two types of estimates may highly correlate (e.g., [40]).

Since the performance metrics are intended to be used for an application-driven evaluation of sampling procedures, the present quantitative illustration is limited to the example of proprioceptive function, and a non-exhaustive set of different sampling procedures and method parameters was presented. There exist other commonly used sampling procedures (e.g., QUEST [17]) that could be evaluated with this framework. It should be noted that this framework does not allow deducing the performance of one sampling procedure from another in absence of a theoretical analysis of the sampling procedures. Therefore, new simulations would be required for other sampling procedures.

To simplify the simulations and the presentation of the results, the guess rate $\gamma = 0.5$ and lapse rate $\lambda = 0$ were held constant for the templates as well as for the fitting process instead of using free parameters. Since it has been shown that not taking lapses into account may bias the threshold estimates [30], and that perception can be non-stationary during experiments due to, for example, inattention, learning, or change in decision criteria [23, 25, 60–62], more realistic simulations should take these factors into account, and more elaborate sampling procedures [16] or methods to address these specific challenges could be used in combination with the fitting process or the sampling procedure [26, 30]. Since the perception thresholds and slopes of the sampled population may not be linearly distributed, as assumed in this work when resampling the $PCTw/iB_{\pm Tol}$ or $nAUC$ space, a more realistic distribution (e.g., log–normal distribution for parameters with semi-infinite positive support) could be based on some prior knowledge. This could be achieved by defining an n-dimensional (in the present case two-dimensional for the threshold and slope parameters) density function in the parameter space to correct *nVUS* with a weighting factor (e.g. multiplying the volume sections by a factor) or to calculate the inhomogeneity parameter σ based on a non-linear resampled space (i.e., resampling density according to the population distribution).

5 Conclusions

This work introduces a set of novel metrics to evaluate the performance of psychophysical sampling procedures in estimating parameters of psychometric functions quantitatively, using application-driven simulations. The analysis framework can be used for any type of sampling procedure and parameters to be estimated (e.g., threshold, slope, lapse rate), and is independent of the parametric versions of psychometric functions (e.g., normal, logistic, Weibull) and the application-specific parameter spaces. In summary, the illustrative analysis of a simulation based on a scenario of proprioceptive assessment using these metrics allowed identifying sub-optimal parameter choices for different simulated sampling procedures and deriving suggestions on how to improve the methods. Furthermore, the optimal sampling procedure could be identified, which could be tuned and analyzed in a second iterative step using the same metrics framework. Thus, these metrics allow a deeper understanding of the strengths and limitations of the sampling procedures, facilitate the parameter tuning, and showed that it is important to evaluate the procedures for different psychometric functions with metrics beyond efficiency.

Acknowledgments

The authors would like to thank J. Egloff and S. Huber for the acquisition of the behavioral data as well as W. L. Popp for fruitful discussions.

Author Contributions

Conceptualization: Mike D. Rinderknecht, Olivier Lambercy, Roger Gassert.

Data curation: Mike D. Rinderknecht.

Formal analysis: Mike D. Rinderknecht.

Funding acquisition: Mike D. Rinderknecht, Olivier Lambercy, Roger Gassert.

Investigation: Mike D. Rinderknecht.

Methodology: Mike D. Rinderknecht.

Project administration: Mike D. Rinderknecht.

Resources: Mike D. Rinderknecht.

Software: Mike D. Rinderknecht.

Supervision: Olivier Lambercy, Roger Gassert.

Validation: Mike D. Rinderknecht.

Visualization: Mike D. Rinderknecht.

Writing – original draft: Mike D. Rinderknecht.

Writing – review & editing: Mike D. Rinderknecht, Olivier Lambercy, Roger Gassert.

References

1. Gescheider G. Psychophysics: The Fundamentals. New Jersey: Lawrence Erlbaum Associates; 1997.

2. Gescheider G. Psychophysics: Method, Theory, and Applications. New Jersey: Lawrence Erlbaum Associates; 1985.
3. Macmillan NA, Douglas Creelman C. Detection Theory: A User's Guide. New Jersey: Lawrence Erlbaum Associates; 2005.
4. Leek MR. Adaptive procedures in psychophysical research. Perception & Psychophysics. 2001; 63(8):1279–1292. https://doi.org/10.3758/BF03194543
5. Treutwein B. Adaptive psychophysical procedures. Vision Research. 1995; 35(17):2503–2522. https://doi.org/10.1016/0042-6989(95)00016-X PMID: 8594817
6. Cornsweet TN. The staircase-method in psychophysics. The American journal of psychology. 1962; p. 485–491. https://doi.org/10.2307/1419876 PMID: 13881416
7. Dixon WJ, Mood AM. A method for obtaining and analyzing sensitivity data. Journal of the American Statistical Association. 1948; 43(241):109–126. https://doi.org/10.1080/01621459.1948.10483254
8. Kaernbach C. Simple adaptive testing with the weighted up-down method. Perception & Psychophysics. 1991; 49(3):227–229. https://doi.org/10.3758/BF03214307
9. Levitt H. Transformed up-down methods in psychoacoustics. The Journal of the Acoustical Society of America. 1971; 49(2B):467–477. https://doi.org/10.1121/1.1912375
10. Tyrrell RA, Owens DA. A rapid technique to assess the resting states of the eyes and other threshold phenomena: the modified binary search (MOBS). Behavior Research Methods, Instruments, & Computers. 1988; 20(2):137–141. https://doi.org/10.3758/BF03203817
11. Findlay J. Estimates on probability functions: A more virulent PEST. Attention, Perception, & Psychophysics. 1978; 23:181–185. https://doi.org/10.3758/BF03208300
12. Taylor MM, Douglas Creelman C. PEST: Efficient estimates on probability functions. The Journal of the Acoustical Society of America. 1967; 41:782. https://doi.org/10.1121/1.1910407
13. Kesten H. Accelerated stochastic approximation. The Annals of Mathematical Statistics. 1958; p. 41–59. https://doi.org/10.1214/aoms/1177706705
14. Robbins H, Monro S. A stochastic approximation method. The annals of mathematical statistics. 1951; p. 400–407. https://doi.org/10.1214/aoms/1177729586
15. Kontsevich LL, Tyler CW. Bayesian adaptive estimation of psychometric slope and threshold. Vision Research. 1999; 39(16):2729–2737. https://doi.org/10.1016/S0042-6989(98)00285-5 PMID: 10492833
16. Prins N. The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). Journal of vision. 2013; 13(7). https://doi.org/10.1167/13.7.3 PMID: 23750016
17. Watson AB, Pelli DG. QUEST: A Bayesian adaptive psychometric method. Perception & Psychophysics. 1983; 33(2):113–120. https://doi.org/10.3758/BF03202828
18. Green DM. A maximum-likelihood method for estimating thresholds in a yes-no task. The Journal of the Acoustical Society of America. 1993; 93(4):2096–2105. http://dx.doi.org/10.1121/1.406696.
19. Hall JL. Hybrid adaptive procedure for estimation of psychometric functions. The Journal of the Acoustical Society of America. 1981; 69:1763. https://doi.org/10.1121/1.385912 PMID: 7240589
20. Pentland A. Maximum likelihood estimation: The best PEST. Attention, Perception, & Psychophysics. 1980; 28(4):377–379. https://doi.org/10.3758/BF03204398
21. Gresham G, Duncan P, Stason W, Adams H, Adelman A, Alexander D, et al. Post-stroke rehabilitation: Assessment, referral, and patient management. Quick Reference Guide for Clinicians, Number 16. Journal of Pharmacoepidemiology. 1996; 5(2):35–63.
22. Sullivan JE, Hedman LD. Sensory dysfunction following stroke: Incidence, significance, examination, and intervention. Top Stroke Rehabil. 2008; 15(3):200–217. https://doi.org/10.1310/tsr1503-200 PMID: 18647725
23. Doll RJ, Veltink PH, Buitenweg JR. Observation of time-dependent psychophysical functions and accounting for threshold drifts. Attention, Perception, & Psychophysics. 2015; 77(4):1440–1447. https://doi.org/10.3758/s13414-015-0865-x
24. Fründ I, Haenel NV, Wichmann FA. Inference for psychometric functions in the presence of nonstationary behavior. J Vis. 2011; 11(6). https://doi.org/10.1167/11.6.16 PMID: 21606382
25. Leek MR, Hanna TE, Marshall L. An interleaved tracking procedure to monitor unstable psychometric functions. The Journal of the Acoustical Society of America. 1991; 90(3):1385–1397. https://doi.org/10.1121/1.401930 PMID: 1939903
26. Rinderknecht MD, Ranzani R, Popp WL, Lambercy O, Gassert R. Algorithm for improving psychophysical threshold estimates by detecting sustained inattention in experiments using PEST. Attention, Perception, & Psychophysics. 2008; 80(6):1629–1645. https://doi.org/10.3758/s13414-018-1521-z

27. O'Regan J, Humbert R. Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. Perception & Psychophysics. 1989; 46(5):434–442. https://doi.org/10.3758/BF03210858

28. Treutwein B, Strasburger H. Fitting the psychometric function. Perception & Psychophysics. 1999; 61(1):87–106. https://doi.org/10.3758/BF03211951

29. Prins N. The psychometric function: The lapse rate revisited. Journal of Vision. 2012; 12(6):25. https://doi.org/10.1167/12.6.25 PMID: 22715196

30. Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics. 2001; 63(8):1293–1313. https://doi.org/10.3758/BF03194544

31. Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence intervals and sampling. Perception & Psychophysics. 2001; 63(8):1314–1329. https://doi.org/10.3758/BF03194545

32. King-Smith PE, Grigsby SS, Vingrys AJ, Benes SC, Supowit A. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. Vision Research. 1994; 34(7):885–912. https://doi.org/10.1016/0042-6989(94)90039-6 PMID: 8160402

33. Madigan R, Williams D. Maximum-likelihood psychometric procedures in two-alternative forced-choice: evaluation and recommendations. Perception & Psychophysics. 1987; 42(3):240–249. https://doi.org/10.3758/BF03203075

34. Simpson WA. The step method: A new adaptive psychophysical procedure. Perception & Psychophysics. 1989; 45(6):572–576. https://doi.org/10.3758/BF03208065

35. Watson AB, Fitzhugh A. The method of constant stimuli is inefficient. Perception & Psychophysics. 1990; 47(1):87–91. https://doi.org/10.3758/BF03208169

36. Taylor MM. On the efficiency of psychophysical measurement. The Journal of the Acoustical Society of America. 1971; 49(2):505–508. https://doi.org/10.1121/1.1912379

37. García-Pérez MA, Alcalá-Quintana R. Sampling plans for fitting the psychometric function. Span J Psychol. 2005; 8(2):256–289. https://doi.org/10.1017/S113874160000514X PMID: 16255393

38. Faes L, Nollo G, Ravelli F, Ricci L, Vescovi M, Turatto M, et al. Small-sample characterization of stochastic approximation staircases in forced-choice adaptive threshold estimation. Perception & Psychophysics. 2007; 69(2):254–262. https://doi.org/10.3758/BF03193747

39. Schmidt RA, Lee T. Motor control and learning. 5th ed. Champaign, IL: Human Kinetics; 2011.

40. Rinderknecht MD, Popp WL, Lambercy O, Gassert R. Experimental Validation of a Rapid, Adaptive Robotic Assessment of the MCP Joint Angle Difference Threshold. In: Auvray M, Duriez C, editors. Haptics: Neuroscience, Devices, Modeling, and Applications. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 3–10. Available from: http://dx.doi.org/10.1007/978-3-662-44196-1_1.

41. Strasburger H. Converting between measures of slope of the psychometric function. Perception & Psychophysics. 2001; 63(8):1348–1355. https://doi.org/10.3758/BF03194547

42. Schutz RW, Roy EA. Absolute error: The Devil in Disguise. J Mot Behav. 1973; 5(3):141–153. https://doi.org/10.1080/00222895.1973.10734959 PMID: 23961744

43. Pumpa LU, Cahill LS, Carey LM. Somatosensory assessment and treatment after stroke: An evidence-practice gap. Aust Occup Ther J. 2015; 62(2):93–104. https://doi.org/10.1111/1440-1630.12170 PMID: 25615889

44. Lincoln NB, Crow JL, Jackson JM, Waters GR, Adams SA, Hodgson P. The unreliability of sensory assessments. Clin Rehabil. 1991; 5(4):273–282. https://doi.org/10.1177/026921559100500403

45. Hillier S, Immink M, Thewlis D. Assessing Proprioception: A Systematic Review of Possibilities. Neurorehabil Neural Repair. 2015; 29(10):933–949. https://doi.org/10.1177/1545968315573055 PMID: 25712470

46. Brewer BR, Fagan M, Klatzky RL, Matsuoka Y. Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. Neural Systems and Rehabilitation Engineering, IEEE Transactions on. 2005; 13(1):1–11. https://doi.org/10.1109/TNSRE.2005.843443

47. Tan HZ, Srinivasan MA, Reed CM, Durlach NI. Discrimination and identification of finger joint-angle position using active motion. ACM Transactions on Applied Perception (TAP). 2007; 4(2):10. https://doi.org/10.1145/1265957.1265959

48. Lambercy O, JuÃ¡rez Robles A, Kim Y, Gassert R. Design of a robotic device for assessment and rehabilitation of hand sensory function. In: Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on. Zurich, Switzerland; 2011. p. 1–6. Available from: http://dx.doi.org/10.1109/ICORR.2011.5975436.

49. Simo L, Botzer L, Ghez C, Scheidt RA. A robotic test of proprioception within the hemiparetic arm post-stroke. J Neuroeng Rehabil. 2014; 11:77. https://doi.org/10.1186/1743-0003-11-77 PMID: 24885197

50. Elangovan N, Herrmann A, Konczak J. Assessing proprioceptive function: evaluating joint position matching methods against psychophysical thresholds. Phys Ther. 2014; 94(4):553–561. https://doi.org/10.2522/ptj.20130103 PMID: 24262599

51. Cappello L, Elangovan N, Contu S, Khosravani S, Konczak J, Masia L. Robot-aided assessment of wrist proprioception. Front Hum Neurosci. 2015; 9:198. https://doi.org/10.3389/fnhum.2015.00198 PMID: 25926785

52. Rinderknecht MD, Lambercy O, Raible V, Liepert J, Gassert R. Age-based model for metacarpophalangeal joint proprioception in elderly. Clin Interv Aging. 2017; 12:635–643. Available from: http://dx.doi.org/10.2147/CIA.S129601.

53. Tan HZ, Srinivasan MA, Eberman B, Cheng B. Human factors for the design of force-reflecting haptic interfaces. Dynamic Systems and Control. 1994; 55(1):353–359.

54. Rinderknecht MD, Lambercy O, Raible V, Büsching I, Sehle A, Liepert J, Gassert R. Reliability, validity, and clinical feasibility of a rapid and objective assessment of post-stroke deficits in hand proprioception. Journal of NeuroEngineering and Rehabilitation. 2018; 15(1). Available from: http://dx.doi.org/10.1186/s12984-018-0387-6.

55. McKee SP, Klein SA, Teller DY. Statistical properties of forced-choice psychometric functions: Implications of probit analysis. Perception & Psychophysics. 1985; 37(4):286–298. https://doi.org/10.3758/BF03211350

56. Prins N, Kingdom FAA. Palamedes: Matlab routines for analyzing psychophysical data.; 2009. Available from: http://www.palamedestoolbox.org.

57. Klein SA. Measuring, estimating, and understanding the psychometric function: A commentary. Perception & Psychophysics. 2001; 63(8):1421–1455. https://doi.org/10.3758/BF03194552

58. Turpin A, Jankovic D, McKendrick A. Identifying steep psychometric function slope quickly in clinical applications. Vision Research. 2010; 50(23):2476–2485. http://dx.doi.org/10.1016/j.visres.2010.08.032.

59. Taylor MM, Forbes SM, Douglas Creelman C. PEST reduces bias in forced choice psychophysics. The Journal of the Acoustical Society of America. 1983; 74:1367. https://doi.org/10.1121/1.390161 PMID: 6643848

60. Watson CS. Time course of auditory perceptual learning. Ann Otol Rhinol Laryngol Suppl. 1980; 89(5 Pt 2):96–102. PMID: 6786201

61. Hall JL. A procedure for detecting variability of psychophysical thresholds. The Journal of the Acoustical Society of America. 1983; 73(2):663–667. https://doi.org/10.1121/1.388958 PMID: 6841806

62. Cohen MR, Maunsell JHR. When attention wanders: how uncontrolled fluctuations in attention affect performance. J Neurosci. 2011; 31(44):15802–15806. https://doi.org/10.1523/JNEUROSCI.3063-11.2011 PMID: 22049423