Enhancing Simulations With Intra-Subject Variability for Improved Psychophysical Assessments

Mike D. Rinderknecht[1], Olivier Lambercy[1], and Roger Gassert[1]
[1]ETH Zurich, Zurich, Switzerland

Psychometric properties of perceptual assessments, like reliability, depend on the stochastic properties of psychophysical sampling procedures resulting in method variability, as well as inter- and intra-subject variability. Method variability is commonly minimized by optimizing sampling procedures through computer simulations. Inter-subject variability is inherent in the population of interest and cannot be acted upon. In contrast, intra-subject variability introduced by confounds cannot be simply quantified from experimental data, as this data also includes method variability. Therefore, this aspect is generally neglected when developing assessments. Yet, comparing method variability and intra-subject variability could give insights on whether effort should be invested in optimizing the sampling procedure, or in addressing potential confounds instead. We propose a new approach to estimate and model intra-subject variability of psychometric functions by combining computer simulations and behavioral data, and to account for it when simulating experiments. The approach was illustrated in a real-world scenario of proprioceptive difference threshold assessments. The behavioral study revealed a test-retest reliability of 0.212. Computer simulations lacking intra-subject variability predicted a reliability of 0.768, whereas the new approach including an intra-subject variability model lead to a realistic estimate of reliability (0.207). Such a model also allows computing the theoretically maximally attainable reliability (0.552) assuming an ideal sampling procedure. Comparing the reliability estimates when exclusively accounting for method variability versus intra-subject variability reveals that intra-subject variability should be reduced by addressing confounds and that only optimizing the sampling procedure may be insufficient to achieve a high reliability. The new approach also allows accelerating the development of assessments by simulating the converging behavior of the reliability confidence interval with a large number of subjects and retests without requiring additional experiments. Having such a tool of predictive value is especially important for target populations where time is scarce, such as for assessments in clinical settings.

Keywords: behavioral data, method variability, parameter estimation by sequential testing (PEST), psychophysics, proprioception assessment, reliability

1 Introduction

The development of assessments of human perception thresholds (e.g., visual, auditory, tactile, or proprioceptive stimuli) is a challenging field, as these require good psychometric and clinimetric properties such as high reliability, for both research and clinical applications. The selection and optimization of psychophysical assessments is, in general, a lengthy, iterative, and cumbersome process where different psychophysical methods need to be tested and their parameters tuned (Rinderknecht et al., in preparation). Evaluating such procedures requires time and financial resources, as it involves repeated assessment of a large number of subjects. This may present a serious hurdle for the development of reliable assessments, especially for sample populations where available time is scarce and recruitment is difficult or expensive (e.g., neurological patients).

When evaluating and optimizing psychophysical methods (e.g., for a high test-retest reliability), different factors play an essential role: method variability as well as inter- and intra-subject variability. While inter-subject variability clearly has an effect on reliability (Streiner and Norman, 2008), it is given by the population of interest and cannot be acted upon.

Mike D. Rinderknecht, Olivier Lambercy, and Roger Gassert are with the Rehabilitation Engineering Laboratory, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland.

Correspondence concerning this article should be addressed to Mike D. Rinderknecht, Rehabilitation Engineering Laboratory, Department of Health Sciences and Technology, ETH Zurich, Lengghalde 5, 8092 Zurich, Switzerland. Phone: +41 44 510 72 38, e-mail: mike.rinderknecht@hest.ethz.ch

Previous works have suggested that a lack of correlation, respectively agreement, between methods tested on the same subjects may originate either from inherent method variability (i.e., based on the stochastic process, the statistical properties of the method, and number of trials) or from intra-subject variability (Rinderknecht et al., 2014). As both method and intra-subject variability are confounded in the outcome measure of a perception assessment, it is impossible to discern one factor from the other and quantify them independently through behavioral studies.

The detection or discrimination capability of physical stimuli is often assumed to resemble a sigmoidal psychometric function (Gescheider, 1997; Macmillan and Douglas Creelman, 2005) based on which the subject generates responses in a stochastic process. Therefore, perception and psychophysical procedures (i.e., complete perception experiments) can be modeled. As a matter of fact, the method variability as well as other performance metrics such as bias and efficiency can be quantified using computer simulations and have been widely investigated for various procedures (Taylor and Douglas Creelman, 1967; Taylor, 1971; Findlay, 1978; Pentland, 1980; Hall, 1981; Madigan and Williams, 1987; Simpson, 1989; Watson and Fitzhugh, 1990; Kaernbach, 1991; Green, 1993; King-Smith et al., 1994)(Rinderknecht et al., in preparation).

In contrast, intra-subject variability is difficult to estimate and cannot be directly quantified based on experimental or simulated data only. Therefore, intra-subject variability has received little attention so far, and is generally neglected in computer simulations. As a consequence, simulations of psychophysical experiments are hardly realistic, and results are not representative. Better knowledge about human perception and the ability to model the intra-subject variability would offer many possibilities, such as comparing, selecting, and tuning different psychophysical methods in simulated scenarios corresponding closely to the real application and population of interest. Furthermore, model-based extrapolation to a larger number of trials or increased sample size, for example to explore their impact on reliability, could then be performed purely in simulation. This could significantly speed up the development and testing of psychophysical assessment procedures.

The aim of this work is twofold: firstly, to present an approach to quantify intra-subject variability, and secondly, to apply and illustrate the approach by creating a general model of intra-subject variability—in this case of proprioceptive perception at the wrist assessed in a two-alternative forced-choice (2AFC) setting. To estimate the intra-subject variability for different parameters of the psychometric function, a dataset with repeated measures from a behavioral study is required. Based on this experimental data, models of the subject's psychometric function are created to simulate the same population. We propose to add individual, statistical noise models on the different parameters (threshold and slope) of the psychometric functions to simulate intra-subject variability. The level of intra-subject variability (i.e., noise) on the different parameters can be quantified by matching the test-retest reliability of the simulated experiment with the test-retest reliability of the behavioral data and by maximizing the similarity between the distributions of outcome measures.

2 Materials and methods

2.1 Behavioral data

2.1.1 Subjects

Thirty-three healthy young subjects ($N_{\text{subjects}} = 33$) were recruited and participated in this experiment (age mean ± SD: 24.1 ± 3.4 years, 20 male and 13 female, 27 right handed, 5 left handed, and 1 ambidextrous). Handedness was assessed with the Edinburgh Handedness Inventory (Oldfield, 1971). Exclusion criteria comprised sensory and motor deficits affecting normal wrist and hand function, as well as any history of neurological or wrist injury. Prior to participating in the experiment, all subjects gave written informed consent. The study was approved by the institutional ethics committee of the ETH Zurich (EK 2015-N-03).

2.1.2 Protocol of the proprioceptive assessment

Each trial of the assessment aiming at estimating the difference threshold of the angular position at the right wrist joint consisted of the consecutive presentation of two different angles and the subsequent judgment by the subject which of the two presented movements was larger (two-interval 2AFC paradigm (Macmillan and Douglas Creelman, 2005)). The movements were applied to the passive wrist with a one degree-of-freedom robotic wrist interface (Chapuis et al., 2010). The movements always started from the resting position (hand aligned with forearm, 0°) and went into flexion direction (maximum 40°). The two presented angles were always centered around a reference of 20°. The difference between the two angles (referred to as level) was defined by an adaptive sampling procedure named Parameter Estimation by Sequential Testing (PEST) (Taylor and Douglas Creelman, 1967). PEST was used with a logarithmic adaptation for positive-only stimuli to avoid an undesired behavior of the algorithm due to zero crossings (Rinderknecht et al., 2014). This adaptive algorithm takes the judgments (also referred to as responses) of past trials into account and changes the difference between the angles accordingly, using heuristic rules to approach the difference threshold as rapidly as possible. The same proprioceptive assessment has been previously used and described in more detail in other studies with a different robotic device for the assessment of the metacarpophalangeal joint (Rinderknecht et al., 2014, 2017, 2018)(Rinderknecht et al., submitted). The same movement timing characteristics and parameters for the PEST algorithm were used in the present experiment, except for the maximum number of trials (start level $x_0 = 5.5°$, start step $\Delta x_0 = \pm 2°$, target performance $P_t = 75\%$, Wald sequential likelihood ratio test parameter $W = 1$, minimum step $\Delta x_{\text{min}} = \pm 0.1°$, maximum trials at same level $trials_{\text{max@}x} = 20$, maximum trials in total $trials_{\text{max}} = 120$). To avoid any visual or auditory cues (e.g., noise emitted by the motor), the tested arm was occluded from vision and white noise was played over headphones during the whole experiment. Each subject performed the assessment in five session on different days (from 1 to 4 days between sessions, with a maximum of 7 days from the first to the last session).

2.1.3 Estimation of the psychometric function

Based on the data from the assessment sequence (i.e., difference between the two presented angles and corresponding response of the subject), the proportion of correct responses can be calculated for the different levels x to fit a sigmoidal psychometric function $\psi(x)$ (**Figure 1**) using a Maximum Likelihood criterion implemented in the Palamedes MATLAB routines (Prins and Kingdom, 2009):

$$\psi\,(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\gamma-\lambda)\,F_{\text{Gauss}}\left(x;\mu,\sigma^2\right) \quad , \qquad (1)$$

with $F_{\text{Gauss}}(x;\mu,\sigma^2)$ a sigmoidal cumulative Gaussian function. In the present work, the threshold parameter α corresponds to the mean μ of the underlying Gaussian function, and the slope parameter β is inversely proportional to the standard deviation σ:

$$\beta = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma} \quad . \qquad (2)$$

The guess rate parameter γ was fixed to 0.5 (according to the 2AFC paradigm), and the lapse rate parameter λ was allowed to vary $\in [0, 0.1]$. This has been shown to reduce estimation bias introduced by isolated scattered lapses (Wichmann and Hill, 2001). The actual slope (first order derivative) of $\psi(x)$ at the inflexion point α is

$$\beta_{\text{inflection}} = \frac{(1-\gamma-\lambda)}{\sqrt{2\pi}}\frac{1}{\sigma} \quad . \qquad (3)$$

This definition of the slope carries as units one over the units of the stimulus, in the present work [1/°], and can be used to compare the slope values across studies using different types of sigmoidal functions $F(x)$ (Strasburger, 2001). To do arithmetic calculations on the slope (e.g., arithmetic mean), it is reasonable to normalize the slope with the following nonlinear function to a range [0, 1] with arbitrary units [a.u.]:

$$\beta_{\text{inflection [a.u.]}} = \frac{\arctan(\beta_{\text{inflection}})}{\pi/2} \quad . \qquad (4)$$

If this nonlinear transformation is not performed, errors in slope estimation can diverge towards infinite for two almost identically steep psychometric functions, which would lead to a distortion when comparing to errors in shallow psychometric functions.

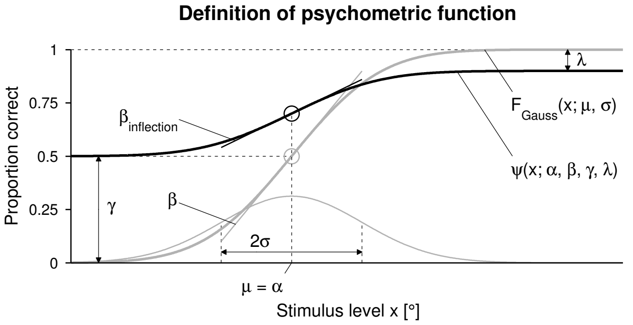

Figure 1. Definition of psychometric function and its parameters. Psychometric function $\psi(x;\alpha,\beta,\gamma,\lambda)$ (bold black sigmoid) and cumulative Gaussian function $F_{\text{Gauss}}(x;\mu,\sigma)$ (bold gray sigmoid) in the case of a two-alternative forced choice (2AFC) task. The thin gray curve is the underlying Gaussian probability density function. The inflection points are indicated as circles in the respective colors.

Based on computer simulations, the estimation quality of psychophysical sampling procedures can be calculated for the parameters of the estimated psychometric function depending on the real parameters (Rinderknecht et al., in preparation). Following this work, the estimation performance of PEST was evaluated with computer simulations using the same parameter values as used in the present behavioral study. The variable estimation error cannot be corrected for. However, the average bias (i.e., constant estimation error) can be removed after fitting the psychometric function with the Maximum Likelihood criterion. While PEST can be considered a bias-free sampling procedure for the threshold estimates, the slope estimation bias showed a strong dependence on the real slope and was approximatively corrected by using the following equation:

$$\beta_{\text{inflection, b.c. [a.u.]}} = \beta_{\text{inflection [a.u.]}}^{\;2.381} \quad . \qquad (5)$$

A further estimation bias in psychophysical experiments with human subjects can arise from longer inattention periods, as loss of attention may alter perception (Leek et al., 1991; Fründ et al., 2011; Cohen and Maunsell, 2011). A method to detect and remove such inattention periods in PEST sequences has recently been proposed (Rinderknecht et al., 2018). This method has shown to reduce estimation errors by up to around 75% and was applied *post-hoc* on the PEST sequences recorded in the behavioral study before fitting the psychometric function.

2.2 Computer simulations

2.2.1 Population model and templates

A model of the population distribution was created for each parameter of the psychometric function based on the averaged parameters (across measurements) of the psychometric functions obtained in the behavioral study. For the threshold α and lapse rate λ, the arithmetic mean was calculated for each subject (across the five measurements) to obtain an improved estimate of the subject's real psychometric function. The same was done for the slope β, however, β was first converted to the slope at inflection $\beta_{\text{inflection}}$ (with the five corresponding lapse rates of the individual subject), normalized ($\beta_{\text{inflection [a.u.]}}$), and the bias was removed ($\beta_{\text{inflection, b.c. [a.u.]}}$) before averaging across the five measurements. Subsequently, the slope was converted back with the inverse transformations (with the averaged lapse of the individual subject). Averaging was not necessary for the guess rate γ, as it was always fixed to 0.5. From these empirical parameter distributions a set of simulated perception models (also referred to as templates $\psi(x)^T$) was randomly sampled. To differentiate between psychometric functions and their parameters originating from the behavioral study and the simulated psychometric functions, the symbol T was added for variables referring to simulated templates (e.g., α^T). The number of templates $\psi^T(x)$ was set to be identical to the number of assessed subjects in the behavioral study ($N_{\text{templates}} = 33$).

2.2.2 Noise model

The threshold including intra-subject variability was modeled with a log–normal distribution with a support $[0, +\infty)$:

$$\alpha^T_{\text{noise}} \sim \text{Lognormal}(\mu_N, \sigma^2_N) \quad . \tag{6}$$

To avoid bias when introducing noise, the mean μ_{noise} was defined to be the threshold of the template:

$$\mu_{\text{noise}} := \alpha^T \quad . \tag{7}$$

The standard deviation σ_{noise} of the variability was controlled with the parameter $\nu_\alpha \in [0, +\infty)$:

$$\sigma_{\text{noise}} := \nu_\alpha \quad , \tag{8}$$

The two parameters of the log–normal distribution were calculated using μ_{noise} and the desired σ_{noise}:

$$\mu_N = \log\left(\frac{\mu_{\text{noise}}}{\sqrt{1+\frac{\sigma^2_{\text{noise}}}{\mu^2_{\text{noise}}}}}\right) \quad \text{and} \tag{9}$$

$$\sigma_N = \sqrt{\log\left(1+\frac{\sigma^2_{\text{noise}}}{\mu^2_{\text{noise}}}\right)} \quad . \tag{10}$$

The slope including intra-subject variability was modeled with a beta distribution with a support $[0, 1]$:

$$\beta^T_{\text{inflection [a.u.], noise}} \sim \mathcal{Be}(\alpha_{\mathcal{Be}}, \beta_{\mathcal{Be}}) \quad . \tag{11}$$

The mean μ_{noise} of $\mathcal{Be}(\alpha_{\mathcal{Be}}, \beta_{\mathcal{Be}})$ was defined to correspond to the normalized slope at the inflection of the template:

$$\mu_{\text{noise}} := \beta^T_{\text{inflection [a.u.]}} \quad . \tag{12}$$

The standard deviation σ_{noise} of the variability was controlled with the parameter $\nu_\beta \in (0, 1]$ serving as a scaling parameter:

$$\sigma_{\text{noise}} := \nu_\beta \hat{\sigma}_{\text{noise}} \quad , \tag{13}$$

where $\hat{\sigma}_{\text{noise}}$ is the maximum possible value for σ_{noise} to avoid a U-shaped distribution. This can be guaranteed with at least one of the parameters $\alpha_{\mathcal{Be}}$ or $\beta_{\mathcal{Be}} \geq 1$, leading to:

$$\hat{\sigma}_{\text{noise}} = \max\left(\mu_{\text{noise}}\sqrt{\frac{1-\mu_{\text{noise}}}{1+\mu_{\text{noise}}}} \quad , \quad \sqrt{\frac{\mu_{\text{noise}}\left(\mu^2_{\text{noise}} - 2\mu_{\text{noise}} + 1\right)}{2-\mu_{\text{noise}}}}\right) \quad . \tag{14}$$

With μ_{noise} and σ_{noise}, the two parameters of the beta distribution $\mathcal{Be}(\alpha_{\mathcal{Be}}, \beta_{\mathcal{Be}})$ could be calculated:

$$\alpha_{\mathcal{Be}} = \frac{\mu_{\text{noise}}\left(-\mu^2_{\text{noise}} + \mu_{\text{noise}} - \sigma^2_{\text{noise}}\right)}{\sigma^2_{\text{noise}}} \quad \text{and} \tag{15}$$

$$\beta_{\mathcal{Be}} = \alpha_{\mathcal{Be}}\left(\frac{1-\mu_{\text{noise}}}{\mu_{\text{noise}}}\right) \quad . \tag{16}$$

No noise was modeled on the lapse rate λ^T and on the guess rate $\gamma^T = 0.5$. The psychometric functions to be used for the simulated PEST sequences were of the form $\psi^T_{\text{noise}}(x; \alpha^T_{\text{noise}}, \beta^T_{\text{noise}}, \gamma^T, \lambda^T)$. For the threshold, 16 equally distributed noise levels $\nu_\alpha \in [0, 1.5]$, and for the slope, 14 noise levels $\nu_\beta \in [0, 1]$ with a twice as high grid density $\in [0.7, 1]$, were simulated.

2.2.3 Procedure

For each combination of ν_α and ν_β, the PEST sequence of the 2AFC task was simulated five times for the whole set of templates $\Psi^T_{\nu_\alpha,\nu_\beta}$. For each single simulated sequence, new random variables α^T_{noise} and $\beta^T_{\text{inflection [a.u.], noise}}$ were drawn from the log–normal and beta distributions, respectively, simulating intra-subject variability across the five measurements. The identical PEST parameters as in the behavioral study were used for the computer simulations. Responses to a specific level x were simulated by comparing a randomly generated number $\in \mathcal{U}(0, 1)$ to $\psi^T_{\text{noise}}(x)$ of the respective template. A smaller random number generated a correct response, and a larger random number a false response.

The simulation of the whole set $\Psi^T_{\nu_\alpha,\nu_\beta}$ was repeated for each combination of ν_α and ν_β $N_{\text{simulations}} = 1000$ times with new randomly sampled parameters (i.e., α^T, β^T, γ^T, λ^T) from the population distribution models.

The psychometric functions from the simulated PEST sequences were estimated identically to the behavioral study, including the bias correction. The only difference lay in the inattention correction algorithm (Rinderknecht et al., 2018), which was not applied on the simulated data. It was assumed that significant biases from potential inattention periods in the behavioral study were already corrected for before creating the population model for the templates. Thus, as no inattention periods were modeled in the simulations, there was no need to apply the algorithm. The computer simulations and estimations of the psychometric function were performed entirely in MATLAB R2014a.

2.3 Data analysis

Test-retest reliability of the estimated thresholds from the five measurements of the behavioral study was quantified by computing the ICC(2,1) intraclass correlation coefficient r (two-way layout with random effects for absolute agreement) (Shrout and Fleiss, 1979) and its 95% confidence interval (CI) (Lexell and Downham, 2005; de Vet et al., 2006).

Identically, for each set $\Psi^T_{\nu_\alpha,\nu_\beta}$, distributions of $N_{\text{simulations}}$ values for the reliability of the estimated thresholds as well as its lower and upper CI bounds for each combination of ν_α and ν_β were generated. From these $N_{\text{simulations}}$ reliability values, the arithmetic mean $\bar{r}^T_{\nu_\alpha,\nu_\beta}$ was calculated. In this two-dimensional noise space an iso-reliability contour where the reliability of the simulated experiment matched the reliability of the behavioral study ($\bar{r}^T_{\nu_\alpha,\nu_\beta} = r$) was calculated (set of ν_α and ν_β pairs). To obtain a smoother contour, the reliability surface was interpolated with a spline on a finer grid (by halving the grid intervals three times in each dimension).

To find which ν_α and ν_β pair of the iso-reliability contour corresponds the best to the intra-subject variability of the behavioral study, for each of the $N_{\text{simulations}}$ per pair, histograms of the parameters of the estimated psychometric functions from the computer simulation were compared to histograms of the parameters of the psychometric functions originating from the behavioral data. This was done by calculating the cosine similarity between the two vectors of histogram bin counts ($\mathbf{h}$ and $\mathbf{h}^T$, for the behavioral and simulated data, respectively) for the parameters α, $\beta_{\text{inflection, b.c. [a.u.]}}$, and λ:

$$similarity^T_i = \cos(\theta_i) = \frac{\mathbf{h}_i \cdot \mathbf{h}^T_i}{||\mathbf{h}_i||_2 ||\mathbf{h}^T_i||_2} \quad \forall i \in \{\alpha, \beta, \lambda\} \quad , \quad (17)$$

where a similarity of 1 represents identical histograms. Note that by using this similarity metric the histograms do not need to be additionally normalized. The following bin sizes were used for α, $\beta_{\text{inflection, b.c. [a.u.]}}$, and λ: 0.25, 0.05, and 0.005. To obtain an overall similarity, the three calculated similarities were multiplied with each other.

$$s^T_{\nu_\alpha,\nu_\beta} = \prod_{i \in \{\alpha,\beta,\lambda\}} similarity^T_i \quad , \quad (18)$$

From these $N_{\text{simulations}}$ overall similarity values, the arithmetic mean $\bar{s}^T_{\nu_\alpha,\nu_\beta}$ was calculated. The iso-reliability contour was projected onto the similarity surface in the two-dimensional noise space after a spline interpolation, identical to what was done for the reliability. The interpolated ν_α and ν_β pair on the iso-reliability contour with the highest average overall similarity was chosen as the best model to estimate intra-subject variability ($\hat{\nu}_\alpha$ and $\hat{\nu}_\beta$).

For a new set of psychometric functions $\Psi^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta}$ with the optimal noise model, the simulation was repeated $N_{\text{simulations}}$ times, and the parameter distributions as well as $\bar{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta}$ and $\bar{s}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta}$ were calculated. In addition, the maximum attainable reliability $\hat{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta}$ (corresponding to no method variability) was computed based directly on the templates with intra-subject noise, but without simulating the psychophysical experiment.

To illustrate the intra-subject variability on a psychometric function, a population average model was computed by averaging the individual subject models. Using the intra-subject variability models with parameters $\hat{\nu}_\alpha$ and $\hat{\nu}_\beta$, 1000 templates were created. The estimate distributions originating from pure method variability as well as from intra-subject variability were compared with each other by plotting the percentage of estimates within a tolerance interval depending on the interval size (percentage within bounds, $PCTw/iB$), and the normalized area under these curves ($nAUC$) according to the methods proposed by Rinderknecht et al. (in preparation).

3 Results

The test-retest reliability coefficient of the behavioral study and its confidence interval was $r = 0.212$ [0.077, 0.394]. The simulated reliability $\bar{r}^T_{\nu_\alpha,\nu_\beta}$ for different ν_α and ν_β pairs as well as the matched iso-reliability contour at r are shown in **Figure 2**. In case of no intra-subject variability, the reliability would correspond to $\bar{r}^T_{\nu_\alpha=0,\nu_\beta=0} = 0.768$ [0.662, 0.859] for the psychophysical paradigm and sampling procedure used in this work (i.e., maximum attainable reliability using these methods for the present population of interest). The overall similarity $\bar{s}^T_{\nu_\alpha,\nu_\beta}$ (combined for threshold, slope, and lapse rate) is visualized in **Figure 3**, together with the same projected iso-reliability contour. The maximum overall similarity on the contour was found for the noise level pair $\hat{\nu}_\alpha = 0.363$ and $\hat{\nu}_\beta = 0.849$ ($\bar{s}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.764$), corresponding to the best intra-subject variability model estimate. The similarities of the distributions of the parameters of the psychometric functions are shown individually in **Figure 4**. The simulated reliability at this noise level pair was $\bar{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.207$ [0.076, 0.384]. The maximum attainable reliability without method

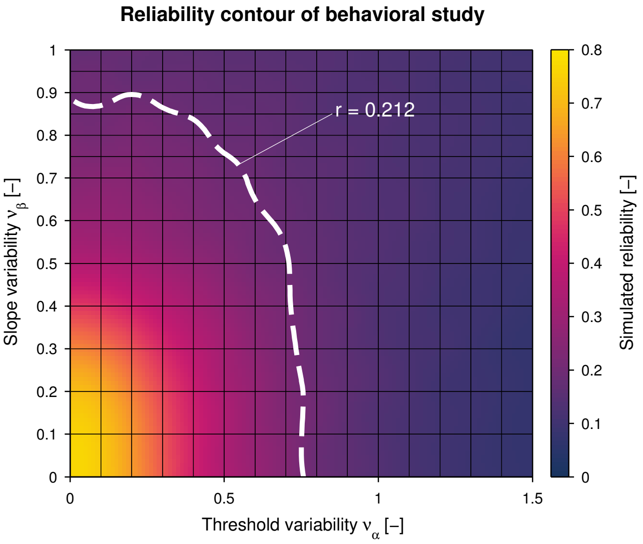

Figure 2. Simulated reliability and iso-reliability contour of behavioral study. For each pair of intra-subject threshold noise ν_α and slope noise ν_β, the simulated reliability averaged across $N_{\text{simulations}} = 1000$ simulations ($\bar{r}^T_{\nu_\alpha,\nu_\beta}$) is represented as a heat map. The dashed white line indicates the iso-reliability contour corresponding to the reliability obtained from the behavioral study ($r = 0.212$).

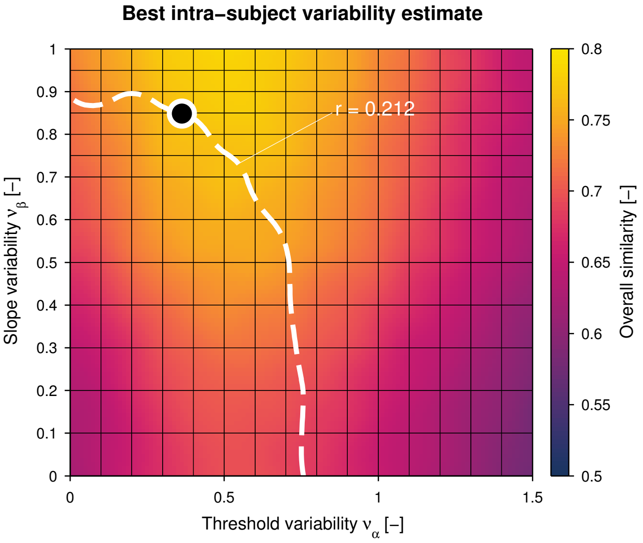

Figure 3. Best intra-subject variability estimate based on overall similarity. For each pair of intra-subject threshold noise ν_α and slope noise ν_β, the overall similarity (combined for threshold, slope, and lapse rate) averaged across $N_{\text{simulations}} = 1000$ simulations ($\bar{s}^T_{\nu_\alpha,\nu_\beta}$) is represented as a heat map. The dashed white iso-reliability contour is identical to **Figure 2**. The noise level pair on the contour with the highest overall similarity ($\bar{s}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.764$) is indicated with a black dot.

variability (i.e., assuming a perfect assessment) for the identified intra-subject variability model would be $\hat{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.552$ [0.403, 0.704]. For illustration purposes, the effect of intra-subject variability on the shape of the psychometric function is shown in **Figure 5** for the population average model $\psi(x; \alpha = 1.696, \beta = 1.708, \gamma = 0.500, \lambda = 0.036)$ and the noise level pair $\hat{\nu}_\alpha$, $\hat{\nu}_\beta$, together with the distributions of threshold and slope resulting from method and intra-subject variability. For the threshold estimation, the $nAUC$ was higher for the method variability compared to the intra-subject variability, whereas for the slope estimation, the opposite was the case. The maximum difference in estimation performance in terms of $PCTw/iB$ was 42.5% at a threshold tolerance of ±0.210°, and 38.1% at a slope tolerance of ±0.299.

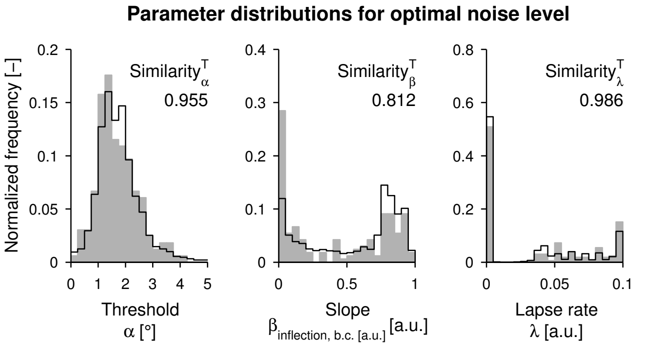

Figure 4. Histogram similarity for the optimal intra-subject variability. Histograms of the parameters of the psychometric functions of the behavioral data (gray fill, 33 × 5 data points) versus the simulated data (black outline, 33 × 5 simulated data points averaged over 1000 simulations) with optimal noise level at the pair $\hat{\nu}_\alpha$, $\hat{\nu}_\beta$.

4 Discussion

In this work we presented an approach to quantify intra-subject variability in psychophysical testing. This was achieved by introducing and adjusting a statistical noise model in computer simulations to match the test-retest reliability and histograms of the parameters of the estimated psychometric functions of a behavioral dataset. Using this approach we estimated the intra-subject variability of healthy subjects in a psychophysical assessment of proprioceptive perception at the wrist using a 2AFC paradigm, and compared the intra-subject variability with the inherent method variability of PEST.

The results showed that for a matched reliability, the similarity between the behavioral and simulated datasets was excellent for the optimal pair of intra-subject threshold and slope variability. Furthermore, the identified intra-subject variability of the threshold was larger compared to the method variability, whereas the opposite was the case for the slope.

4.1 Intra-subject and method variability

When trying to estimate the test-retest reliability based on the population model without intra-subject variability, the reliability coefficient would be largely overestimated. In the present sample population this would result in a considerable error of 262.3%. In contrast, when including intra-subject variability in the simulation, the reliability of the simulated experiment matched the reliability of the behavioral study with an absolute error of 0.005 ($\bar{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.207$ and $r = 0.212$, respectively), corresponding to a relative error of 2.4%. In theory, this error should be zero, however, since the estimates were based on a stochastic generation of responses, the simulated test-retest reliability varied across simulation runs. To improve the estimate of intra-subject variability, and therefore the match of reliability values, a high number of repetitions ($N_{\text{simulations}}$) were performed to obtain higher statistical power, and the grid of the simulated intra-subject variability levels in the two-dimensional reliability space was interpolated. This error could be further minimized by increasing the number of repetitions and the density of the simulation grid. Further indication for a good model estimation quality is provided by the fact that not only the simulated and behavioral reliability coefficient matched, but also matching errors for the CI were low (absolute [0.001, 0.010] and relative [1.3%, 2.5%] errors for the lower and upper bound). Moreover, cosine similarity between behavioral and simulated outcome measures was very high for all three parameters α, β, and λ (> 0.8), and thus demonstrates that the population's inter- and intra-subject variability models accurately represent the actual population.

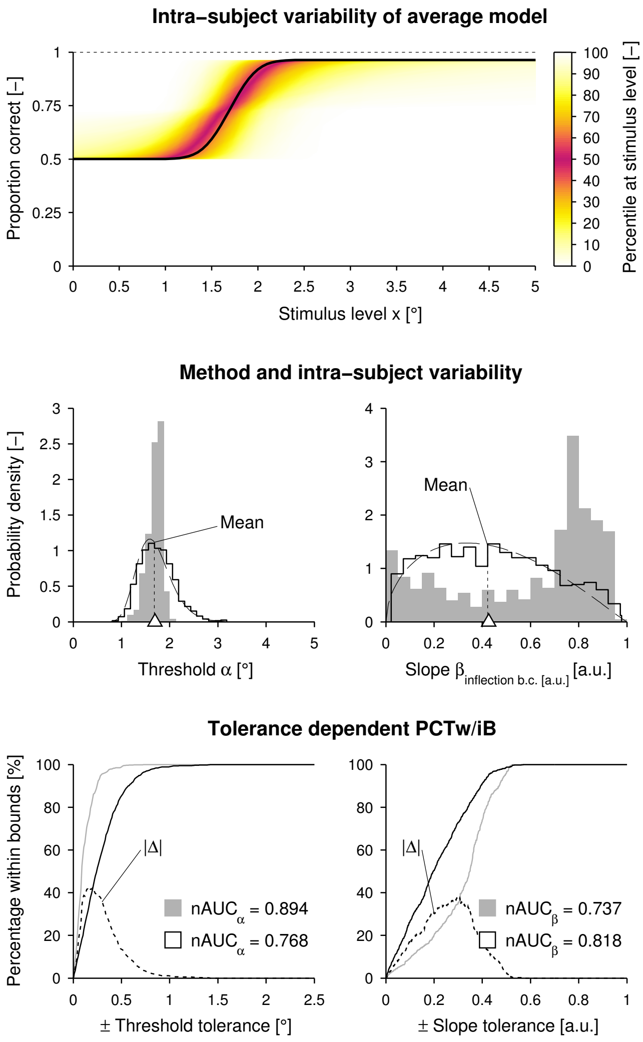

Figure 5. Illustration of intra-subject variability. (Top) For each stimulus level x the distribution of proportion correct is plotted as a heat map based on 1000 templates created with the population average model (bold black sigmoid) and the intra-subject variability models using $\hat{\nu}_\alpha$ and $\hat{\nu}_\beta$. **(Middle)** The dashed black distribution curves correspond to the parametric log–normal and beta intra-subject variability models. The histograms (black outline) as well as the dashed black lines for the means show the parameter distributions of the 1000 templates including intra-subject variability. The white triangles indicate the threshold and slope of the population average model (without noise). As a comparison, the inherent method variability (histogram with gray fill) for the same population average model without intra-subject variability is plotted. **(Bottom)** The percentage of estimates within a tolerance interval (percentage within bounds, $PCTw/iB$) around the parameters of the population average model is plotted against the size of the interval (gray: method variability, black: intra-subject variability), together with the absolute difference of percentage (|Δ|, dashed black line). For both method and intra-subject variability, the normalized area under the curve ($nAUC$) is calculated.

The presented method allows to discern between and compare intra-subject variability and method variability. When assuming invariant subjects (i.e., no intra-subject variability), the test-retest reliability for the threshold would be 39.2% higher compared to when the estimated intra-subject variability is included in the simulation, but a perfect method (i.e., no method variability) would be assumed. This is also reflected by the $nAUC$ for the threshold (a non-parametric metric to evaluate the variability of estimation errors), which is higher by 16.4% for the simulated case with method variability only. Based on these findings, if the assessment was to be improved, one could suggest to address factors influencing the intra-subject variability, before optimizing the psychophysical sampling procedure, as even with a perfect method, the reliability would ceil at $\hat{r}^T_{\hat{\nu}_\alpha,\hat{\nu}_\beta} = 0.552$ due to intra-subject variability. On the contrary, the slope estimates suffer from poor method performance and, according to the U-shaped estimate distribution (histogram with gray fill in **Figure 5 (Middle, right)**), outcome measures are predominantly severely under- or overestimated. As a consequence, the $nAUC$ for the simulated case with intra-subject variability only is 11% higher. Thus, if the slope estimation should be improved, it would be important to optimize the current sampling procedure or choose another sampling procedure (e.g., the Ψ method, designed to estimate both the threshold and the slope (Kontsevich and Tyler, 1999)).

4.2 Advantages and limitations of this method

The advantage of this method is that the test-retest reliability is affected by all terms of variability (inter- and intra-subject, and method variability). As a consequence, since the inter-subject variability can be approximated by taking the averaged psychometric functions for each subject and the method variability is given by the simulation, the intra-subject variability can be estimated. Furthermore, the reliability can be calculated even if only two measurements were done per subject, whereas, for example, calculating the standard deviation of two measurements for each subject is very likely a poor estimate of the real variability (besides being still confound with method variability). However, it should be noted that, depending on the intra-subject and method variability, the quality of the model of the population (and inter-subject variability) can be compromised if only two measurements are available per subject. Thus, in case of a poor population model, an overestimated inter-subject variability may be compensated by an underestimated intra-subject variability and vice versa when matching the reliability. Ideally, the available behavioral data would encompass a large sample size (for a good representation of the population) and a large number of measurements (for a good estimate of each subject's psychophysical function). An advantage of sampling templates from the computed distributions representing the population compared to using the averaged psychometric functions as templates, is that repeated randomly sampled templates should lead to more generalizable results than bootstrapping from a limited set of subjects. More importantly, it offers the possibility to sample more templates from the distribution, for example to predict how the reliability and its confidence interval changes with increasing sample size.

A limitation of the present simulations is that no intra-subject variability was modeled for the lapse rate. It would be possible model the lapse rate including intra-subject variability with a beta distribution as for the slope, but with an adapted support. However, for the sake of simplicity, this was omitted here. As a matter of fact, as the histogram similarity of the lapse rate parameter is almost 1, it shows that adding a lapse rate variability model may not be necessary. When identifying the best model of intra-subject variability, the noise level pair ν_α, ν_β, where overall similarity is the highest, may not lie on the iso-reliability contour corresponding to the reliability r of the behavioral data. One reason is that in the histogram, similarity inter- and intra-subject variability are confounded, and the similarity may vary depending on the selection of bin sizes. In contrast, using the reliability as a metric should provide a more robust and accurate estimate of the variability model, as it distinguishes between inter- and intra-subject variability despite taking both into account. Therefore, the overall similarity is used only as a second criterion to find the optimal model. One major limitation of this approach to estimate intra-subject variability, is that it only provides one variability model for the whole sample and not individual models for each subject. However, this is already a significant improvement over no variability model, and may be accurate enough for many applications.

5 Conclusions

Computer simulations offer a valuable and powerful tool to simulate and optimize psychophysical experiments. While they can be used to evaluate different procedures and their method variability, existing computer simulations are often not representative of real-world scenarios, as critical aspects such as the intra-subject variability are neglected. As a matter of fact, intra-subject variability cannot be directly quantified from behavioral data. This work introduces a new approach based on the combination of computer simulations and behavioral data to separate method variability from intra-subject variability and to estimate and model intra-subject variability in psychophysical experiments.

Given a realistic model of the population, different psychophysical procedures can be simulated and compared, and the procedures can be tuned to the specific application and target population. Quantifying the method and intra-subject variability allows putting them into perspective when developing assessments. Given the intra-subject variability, it allows simulating an experiment with an ideal psychophysical method (i.e., finding the theoretically maximally attainable performance of an assessment). These two aspects can inform the decision whether effort should be spent on improving the psychophysical procedure (i.e., reducing method variability) or if potential confounds affecting intra-subject variability should be addressed. The efficiency of attempts to reduce confounds (e.g., inattention (Rinderknecht et al., 2018)) could be quantified (using the presented method) based on a reduction of the intra-subject variability. Furthermore, based on the more complete model also containing intra-subject variability, it is also possible to examine the impact of a larger number of trials on reliability, or the converging behavior of the reliability's confidence interval bounds with a larger number of subjects, as well as retests, without having to conduct additional experiments. This presents a particular benefit for studies with populations where time for assessments is limited or expensive, as in the case of a clinical setting.

Disclosure/conflict-of-interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

MR, OL, and RG contributed to the conception of this work. MR developed the methodology, implemented the experiments and the computer simulations, performed the analysis, interpreted the results, and drafted the manuscript. MR, OL, and RG revised the manuscript and approved the final version.

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. 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.

References

Chapuis, D., De Grave, R. B., Lambercy, O., and Gassert, R. (2010). ReFlex, a haptic wrist interface for motor learning and rehabilitation. In *Haptics Symposium, 2010 IEEE* (Waltham, Massachusetts, USA: IEEE), 417–424

Cohen, M. R. and Maunsell, J. H. R. (2011). When attention wanders: how uncontrolled fluctuations in attention affect performance. *J Neurosci* 31, 15802–15806

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

Findlay, J. (1978). Estimates on probability functions: A more virulent PEST. *Attention, Perception, & Psychophysics* 23, 181–185

Fründ, I., Haenel, N. V., and Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *J Vis* 11

Gescheider, G. (1997). *Psychophysics: The Fundamentals* (New Jersey: Lawrence Erlbaum Associates)

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *The Journal of the Acoustical Society of America* 93, 2096–2105

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *The Journal of the Acoustical Society of America* 69, 1763

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics* 49, 227–229

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., and Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Research* 34, 885–912

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

Leek, M. R., Hanna, T. E., and Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *The Journal of the Acoustical Society of America* 90, 1385–1397

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

Macmillan, N. A. and Douglas Creelman, C. (2005). *Detection Theory: A User's Guide* (New Jersey: Lawrence Erlbaum Associates)

Madigan, R. and Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: evaluation and recommendations. *Perception & Psychophysics* 42, 240–249

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Attention, Perception, & Psychophysics* 28, 377–379

Prins, N. and Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data.

Rinderknecht, M. D., Lambercy, O., Raible, V., Liepert, J., and Gassert, R. (2017). Age-based model for metacarpophalangeal joint proprioception in elderly. *Clin Interv Aging* 12, 635–643

Rinderknecht, M. D., Popp, W. L., Lambercy, O., and Gassert, R. (2014). Experimental Validation of a Rapid, Adaptive Robotic Assessment of the MCP Joint Angle Difference Threshold. In *Haptics: Neuroscience, Devices, Modeling, and Applications*, eds. M. Auvray and C. Duriez (Berlin, Heidelberg: Springer Berlin Heidelberg), Lecture Notes in Computer Science. 3–10

Rinderknecht, M. D., Ranzani, R., Popp, W. L., Lambercy, O., and Gassert, R. (2018). Method for Improving Psychophysical Threshold Estimates by Detecting Sustained Inattention. *bioRxiv*

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

Simpson, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics* 45, 572–576

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception & Psychophysics* 63, 1348–1355

Streiner, D. L. and Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use* (USA: Oxford university press)

Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *The Journal of the Acoustical Society of America* 49, Suppl 2:505–Suppl 2:508

Taylor, M. M. and Douglas Creelman, C. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America* 41, 782

Watson, A. B. and Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics* 47, 87–91

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