Algorithm for improving psychophysical threshold estimates by detecting sustained inattention in experiments using PEST

Mike D. Rinderknecht¹ · Raffaele Ranzani¹ · Werner L. Popp^{1,2} · Olivier Lambercy¹ · Roger Gassert¹

Published online: 10 May 2018
© The Psychonomic Society, Inc. 2018

Abstract

Psychophysical procedures are applied in various fields to assess sensory thresholds. During experiments, sampled psychometric functions are usually assumed to be stationary. However, perception can be altered, for example by loss of attention to the presentation of stimuli, leading to biased data, which results in poor threshold estimates. The few existing approaches attempting to identify non-stationarities either detect only whether there was a change in perception, or are not suitable for experiments with a relatively small number of trials (e.g., < 300). We present a method to detect inattention periods on a trial-by-trial basis with the aim of improving threshold estimates in psychophysical experiments using the adaptive sampling procedure Parameter Estimation by Sequential Testing (PEST). The performance of the algorithm was evaluated in computer simulations modeling inattention, and tested in a behavioral experiment on proprioceptive difference threshold assessment in 20 stroke patients, a population where attention deficits are likely to be present. Simulations showed that estimation errors could be reduced by up to 77% for inattentive subjects, even in sequences with less than 100 trials. In the behavioral data, inattention was detected in 14% of assessments, and applying the proposed algorithm resulted in reduced test–retest variability in 73% of these corrected assessments pairs. The novel algorithm complements existing approaches and, besides being applicable post hoc, could also be used online to prevent collection of biased data. This could have important implications in assessment practice by shortening experiments and improving estimates, especially for clinical settings.

Keywords Psychophysics · Parameter Estimation by Sequential Testing (PEST) · Adaptive algorithm · Bias correction · Proprioception assessment

Introduction

Psychophysical procedures are used in many fields to assess sensory perception thresholds of stimuli in different modalities (e.g., vision, hearing, proprioception, or tactile sensation). During experiments, the assumption is generally made that the neurological and psychological mechanisms underlying perception are stationary. However, this may

not always be the case, as fatigue, attention, learning, or a change in decision criteria or response bias occurring during the experiment can alter perception (Cameron, Tai, & Carrasco, 2002; Carrasco, 2006; Carrasco, Ling, & Read, 2004; Cohen & Maunsell, 2011; Doll, Veltink, & Buitenweg, 2015; Hall, 1983; Leek, Hanna, & Marshall, 1991; Watson, 1980). Attentional deficits may be a common consequence of many types of brain damage, such as stroke, traumatic brain injuries, or brain tumors (Bruhn & Parsons, 1971; Rinne et al., 2013; Tuhim, 1993; van Zomeren & van den Burg, 1985). Furthermore, according to pedagogical literature, the attention span is also limited in neurologically intact humans (7–20 min) (Bligh, 1998; David & Dukette, 2009; Dent & Harden, 2013; Petty, 2004). Thus, depending on the assessment paradigm, inattention can be a significant confound. Subjects may be distracted during the assessment and forget about the intensity of

✉ Mike D. Rinderknecht
mike.rinderknecht@hest.ethz.ch

¹ Rehabilitation Engineering Laboratory, Department of Health Sciences and Technology, ETH Zurich, Lengghalde 5, 8092 Zurich, Switzerland

² Balgrist University Hospital, Spinal Cord Injury Center, Forchstrasse 340, 8008 Zurich, Switzerland

the stimulus after presentation, or completely miss them, resulting in biased or random answers and thus incorrect perception estimates.

Perception is often described by psychometric functions relating physical stimuli to the performance of a subject in a detection or discrimination task (Gescheider, 1997). It has been suggested that estimates of a threshold (i.e., stimulus level that produces a certain probability of some judgment on average) from adaptive stimulus placement procedures are robust against symmetric variations of the true threshold of the sampled psychometric function (Green, Richards, & Forrest, 1989). However, it is improbable that sustained periods of attention loss have a symmetric influence, but rather a detrimental effect on the mean and variance of the threshold estimates (Green, 1995). Furthermore, other parameters of the psychometric function such as the slope will also be affected (Leek et al., 1991). Finally, sustained inattention can lead to divergence of adaptive psychophysical procedures (refer to Treutwein (1995) and Leek (2001) for reviews), resulting in longer experiments, which in turn can be unfavorable for sustained attention. Thus, when being inattentive in the later course of the experiment where the stimulus levels should be converging thanks to the adaptive procedure, the psychometric function may be least stationary. In all cases, whether the estimate of the threshold is based on the final trials of an adaptive sequence (Taylor & Douglas Creelman, 1967), or on fitting a psychometric function to the proportions of correct responses for different stimulus levels (Treutwein & Strasburger, 1999), the resulting estimate may be unreliable and biased (Fründ, Haenel, & Wichmann, 2011), especially in two-alternative forced-choice (2AFC) tasks (Green, 1995). Thus, a general challenge remains: How can biases resulting from inattention during psychophysical experiments be corrected for?

While experienced experimenters may be able to potentially detect courses of performance being visibly influenced by inattention, based on sudden performance level decreases for a certain period, this way of analyzing the data is not reproducible. Therefore, more reproducible and automatable methods or algorithms would be desired to detect such periods. Physiological signals such as electrodermal activity (EDA) could potentially be used to detect inattention intervals, as arousal has been found to be a strong predictor for attention (Prokasy & Raskin, 1973). However, depending on the situation causing inattention to the psychophysical task, it may not be obvious to interpret a change in the EDA signal, as either the subject could be simultaneously inattentive to the task but aroused by a distractive situation, or as EDA could also respond to cognitive load due to the task at hand. Furthermore, the measurement of EDA requires additional equipment and may not be applicable in some settings.

Therefore, methods based solely on the stimulus sequence and responses recorded during the experiment would be of great advantage.

There have been only few attempts aiming at identifying sequences of questionable validity, as they may have been affected by inattention, and describing the overall stability of the threshold (Hall, 1983; Leek et al., 1991). However, these methods require computer simulations to compute normative tolerance limits and confidence intervals. As these statistical values are case specific (i.e., dependent on psychometric functions, sampling procedures, step sizes, etc.), new simulations would have to be performed for each specific experiment at hand, limiting the practicability of these approaches. This may motivate the use of empirically determined tolerance intervals (based on knowledge from prior experiments) (Hall, 1983). Together with the fact that they cannot identify specific trials affected by inattention within the sequences, nor correct for threshold estimate biases, these can only be used as preliminary screening tools. A different approach to improve parameter estimates of the psychometric function is to account for lapses (i.e., stimulus-independent errors, which can be caused by inattention) (Wichmann & Hill, 2001), or to improve sampling strategies of psychophysical procedures (Prins, 2013; Shen & Richards, 2012). While these approaches constitute viable solutions to identify questionable sequences as a whole, or to improve the goodness of fit and reduce variance of estimates in case of few isolated scattered lapses, these methods do not allow to detect local changes of attention on a trial-by-trial basis within a sequence.

A way to track threshold changes (i.e., varying threshold estimates throughout the experiment) consists in estimating the parameters of the psychometric function based on a subset of sequence data using a moving-window approach. This method has shown promising results in computer simulations using windows spanning 25 trials for modeled non-stationary thresholds following linear and exponential changes (Doll et al., 2015). Together with a classification or weighting algorithm, such a threshold tracking method could be used to identify and remove trials affected by attention loss. However, while this tracking method may be suitable for nociceptive detection thresholds (Doll et al., 2015), its applicability for assessments in other modalities may be limited for several reasons. First, the reliability of estimates may not be high enough even when using the complete sequence, as more than hundred trials may be required for acceptable confidence limits (McKee, Klein, & Teller, 1985), which is especially the case for the 2AFC paradigm (King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994) and psychometric functions with shallow slopes (Green, 1995). Second,

efficient adaptive procedures targeting different “sweet points” for the optimized estimation of parameters (e.g., Parameter Estimation by Sequential Testing (PEST) (Taylor & Douglas Creelman, 1967) or the maximum-likelihood procedure proposed by Shen and Richards (2012)) will not provide sufficient coverage of stimulus levels within each window given the limited window sizes.

Fründ et al. (2011) proposed two approaches to adjust for the underestimation of parameter estimate confidence intervals (CIs) from non-stationary psychometric functions by estimating the ratio of dependent trials (leading to non-stationarity) in blocks of specific number of trials and using this ratio to scale the CIs (correction of inference) or to reduce the number of trials in each block (correction of data). While proving successful for a large number of trials, these approaches show limited or no improvement for relatively small numbers of trials (e.g., < 300). Moreover, this work focused on the CIs and not on the actual parameter estimates, and only symmetric non-stationarities were modeled. In summary, existing methods lack the ability to precisely identify onset and end of inattention periods on a trial-by-trial basis within short sequences with a low number of trials to solve the underlying problem that a non-negligible amount of biased data from a non-stationary psychometric function may be included in the estimation.

We present a new empirical approach to identify sustained inattention periods for improving threshold estimates by rejecting trials affected by non-stationarity. The proposed algorithm is based on data from the sequence of an adaptive sampling procedure and a stability criterion of the evolution of threshold estimates. Computer simulations were used to evaluate the performance of the correction method for different levels of simulated inattention. Additionally, behavioral data were used to illustrate and support the applicability of the proposed method as well as to validate the simulation and inattention model design. This work was framed around a scenario of a proprioceptive difference threshold assessment using PEST (Taylor & Douglas Creelman, 1967) in combination with a two-interval 2AFC assessment paradigm (Macmillan & Douglas Creelman, 2005).

Background

PEST: Parameter estimation by sequential testing

The adaptive PEST algorithm (Taylor & Douglas Creelman, 1967) and the transformed and weighted up/down method (García-Pérez, 1998) for assessing perception thresholds offer many advantages over other psychophysical procedures: their simple implementation, freely selectable convergence point (i.e., desired target of percentage of correct

responses), and no required prior assumptions on the subject’s psychometric functions. The latter aspect is especially of advantage for a heterogeneous population, such as neurological patients. As PEST has already been previously used for the assessment of proprioceptive function (Metzger et al., 2014; Rinderknecht, Lambercy, Raible, Liepert, & Gassert, 2017), it was chosen as adaptive sampling procedure in this work. PEST is based on a Wald sequential likelihood-ratio test (Wald, 1947), defining at each trial whether the stimulus level should remain at the same level or be decreased, respectively increased, by a specific step size, depending on the proportion of correct responses at the specific stimulus level. The stimulus level is maintained if $N_{cor} \in (N_{tot} \times P_t \pm W)$, where N_{cor} and N_{tot} correspond to the correct and total number of trials at the current stimulus level, P_t the desired target performance, and W the Wald sequential likelihood-ratio test parameter. Direction reversals (switching from decreasing to increasing steps, or vice versa) allow PEST to converge towards the threshold to be estimated in an oscillatory manner. The next stimulus level to be presented is defined by a set of heuristic rules taking into account responses of past trials:

- The step size is halved at 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 size of 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 the preceding step size.

In our application using a 2AFC paradigm, 75% was chosen as the desired target of percentage of correct responses P_t for PEST. Thus, the stimulus level corresponding, on average, to 75% performance is defined as “threshold” in the present work. A Wald sequential likelihood-ratio test parameter W of 1 is recommended by Taylor and Douglas Creelman (1967) for a reasonable trade-off between more precise estimates and shorter assessments. PEST requires two starting parameters: the start level and the start step, which should be selected if possible using some prior knowledge about the distribution on thresholds to be assessed in an experiment. The PEST algorithm features different termination criteria: maximum total number of trials, maximum number of trials at the same stimulus level, and minimum step size. As soon as one of the termination criteria is fulfilled, the outcome measure (i.e., estimate of the threshold) of the assessment is defined by the level of the next stimulus which would be presented.

PEST has often been used in psychoacoustic experiments using auditory stimulus levels in dB (Hall, 1981; Leek, 2001; Taylor & Douglas Creelman, 1967). However, it has been suggested to implement the PEST algorithm

in a logarithmic domain for “non-logarithmic stimuli”, such as absolute angles in a proprioceptive assessment, to avoid inefficient behavior (e.g., due to zero crossings) (Rinderknecht, Popp, Lambercy, & Gassert, 2014). This adapted version making use of the logarithmic domain was used in the following computer simulations and behavioral experiment in stroke survivors on finger proprioception. As a two-interval 2AFC paradigm was used, the absolute angular difference between the two consecutive stimuli of one trial are referred to as stimulus level x , or in short “level”.

Fitting the psychometric function

Despite PEST being an adaptive algorithm able to directly provide an estimate of the threshold at convergence, some scenarios may not allow convergence of the algorithm, for example due to a premature termination of the experiment or too severe termination criteria. One robust solution to address this issue, consists in using all available data of the PEST sequence (i.e., pairs of stimulus level x and response) and estimating the threshold by fitting a psychometric function $\psi(x)$ (Hall, 1981). Note that defining the threshold x_T , where $\psi(x_T) = 0.75$, provides a comparable outcome measure—in our application the proprioceptive difference threshold—independently of whether only PEST or this hybrid approach combining PEST with a subsequent function fitting is used. This hybrid approach has been validated in a previous experimental study (Rinderknecht et al., 2014). Therefore, in both simulations and behavioral experiments of the present work, the thresholds were estimated based on the fitting procedure using a maximum likelihood criterion implemented in the Palamedes MATLAB routines (Prins & Kingdom, 2009).

The psychometric functions $\psi(x)$ are defined by the proportion of correct responses at different stimulus levels x :

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

where $F(x; \alpha, \beta)$ is a generic sigmoid function. In the present work, we used as sigmoidal function a cumulative normal function $F_{Gauss}(x; \mu, \sigma)$ with mean μ , corresponding to the inflection point of $\psi(x)$, and standard deviation σ , inversely proportional to the slope parameter β :

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

$\operatorname{erf}(x)$ being the standard definition of the error function:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad . \quad (3)$$

Given the 2AFC paradigm used in this proprioceptive assessment framework, the guessing rate γ was set to 0.5. When fitting simulated and behavioral data in the present work, the lapse rate λ was allowed to vary $\in [0, 0.1]$,

as it may reduce estimation bias and account for isolated scattered lapses (Swanson & Birch, 1992; Wichmann & Hill, 2001). While the results by Wichmann and Hill (2001) have been challenged by others showing that significant and systematic bias remains even when including a lapse rate estimate (Linares & López-Moliner, 2016; Prins, 2012), unpublished simulations from our group showed minor improvements in the threshold estimates, especially for larger lapse rates. We therefore decided to include the lapse rate parameter when fitting the psychometric function. Note that the interval was selected to be slightly larger than $[0, 0.06]$, as used in the work by Wichmann and Hill (2001), in order to account for a potentially higher probability of inattention in stroke patients (Rinne et al., 2013; Tuhim, 1993). If $\lambda \neq 0$, the inflection point μ does not correspond to the threshold x_T . The advantage of using as outcome measure $x_T = \psi^{-1}(0.75)$ based on a fixed performance level instead of the inflection point allows comparing thresholds independently of the used parametric versions of the psychometric functions. Furthermore, following the suggestions from Strasburger (2001) to obtain comparable values across different studies, the actual slope is defined as:

$$\beta_{\text{inflection}} = \frac{(1 - \gamma - \lambda)}{\sqrt{2\pi}} \frac{1}{\sigma} \quad , \quad (4)$$

being the first-order derivative of $\psi(x)$ at the inflexion point. At the same time, this slope corresponds to the maximal slope of the sigmoid function. Figure 1 illustrates the different functions and definitions of the various parameters.

Method to correct for sustained inattention

The aim of the proposed method to correct for sustained inattention consists in identifying any period in which the subject is inattentive, and excluding the biased data

Fig. 1 Definition and parameters of the psychometric function $\psi(x; \alpha, \beta, \gamma, \lambda)$ (bold black sigmoid), the cumulative normal function $F_{Gauss}(x; \mu, \sigma)$ (bold gray sigmoid), and its underlying normal probability density function (gray). The inflection points and slopes at these points are indicated by circles and lines, respectively, in the colors of the sigmoids

collected during this period when using the fitting process for threshold estimation. Thus, the time point at which the subject's performance starts to decrease due to inattentiveness and the potential time point at which the performance starts improving again (i.e., subject being attentive again) have to be estimated.

When visually inspecting recorded PEST stimulus level sequences, experienced exterminators would search for a fast increase of typically 3–5 steps, and a marked subsequent decrease after a certain period. This intuitive search process was implemented as the first part of the heuristic algorithm. Using visual inspection of samples from previously recorded data (different proprioceptive assessments in healthy subjects) we established a preliminary set of parameter values (e.g., number of increasing and decreasing steps), and the method's face validity was examined by two experienced researchers. To obtain more reliable results, the idea of combining the first approach with a continuous stability criterion based on the evolution of the threshold estimate emerged. The implementation of this second part was inspired from Bollinger bands (Bollinger, 2001), a volatility indicator in finance trading. Bollinger bands consist of a moving average and an upper and lower band defined by a multiple of the moving standard deviation. The algorithm's performance (e.g., accuracy of threshold estimates) was tested for a few sets of parameter values through computer simulations, where the psychometric functions to estimate and the modeled inattention periods were known. Note that the selection of values to be tested was based on a manual and non-systematic trial-and-error process. Here, we limit the presented results to the tested values resulting in the best performance. Values could be tuned in the future to each specific application using more automated methods to potentially obtain a better performance. The terminology “model/modeling” is used in this work for definitions in the context of computer simulations, and not for fitting a parametric psychometric function to experimental or simulated data in order to make the distinction clearer.

An example of inattention detection and trial exclusion in a simulated run for a psychometric function with a threshold $x_T = 1^\circ$ and a slope $\beta_{\text{inflection}} = 0.25/^\circ$ is presented in Fig. 2. Inattention was modeled by increasing the lapse rate $\lambda_{\text{inattention}}$ to 0.5 at trial 31 until trial 50. More details about how inattention is modeled are described in following section. This example clearly illustrates how the PEST sequence diverges rapidly after modeled inattention onset (Fig. 2, top).

Initially, the first trial, which is followed by four consecutive PEST level increases without reversal (denoted as $t_{\text{PEST},4\text{up}}$, t standing for trial), is identified (Fig. 2, top). If this trial is within the first eight trials, it is disregarded and replaced by the next occurrence. This allows the PEST algorithm to adapt the stimulus level in

Fig. 2 Example of a PEST sequence and corresponding adaptive boundary for inattention correction in a simulated run for a template $\psi_{1,0.25}$ (threshold $x_T = 1^\circ$ and slope $\beta_{\text{inflection}} = 0.25/^\circ$) and $\lambda_{\text{inattention}} = 0.5$. *Top*: Level of the PEST sequence (solid black line) and evolution of the threshold estimate $\hat{x}_{T,t}$ (dashed black line). The upward-pointing gray triangle designates $t_{\text{PEST},4\text{up}}$, the first trial being followed by four consecutive level increases without reversal, and the downward-pointing gray triangle designates $t_{\text{PEST},2\text{down}}$, the first (after detected inattention onset) trial followed by two consecutive level decreases without reversal. The period of trials which are excluded by the inattention correction method are indicated by the hatched region spanning from $t_{\text{excl},\text{start}}$ to $t_{\text{excl},\text{end}}$. The modeled inattention period is labeled and indicated by the gray bracket and gray dashed lines, and the threshold of the template by the black triangle. *Bottom*: Raw threshold estimate $\hat{x}_{T,t}$ (dashed line), exponentially weighted, smoothed threshold estimate average $\bar{\mu}_t$ (solid line), and threshold stability boundary based on $\bar{\sigma}_t$ (gray shaded band)

initial trials if the start level was chosen too low. Note that the length of eight trials strongly depends on the choice of the start level and start step, and tuning would require detailed knowledge of the threshold distribution of the subject population in question. However, in general, the start level should be chosen high enough to make the task easily understandable for the subject and thus render this additional rule superfluous.

As the first few level increases starting at trial $t_{\text{PEST},4\text{up}}$ could be a result of normal oscillations around the threshold induced by the adaptive sampling procedure, the onset of inattention may not always directly correspond to $t_{\text{PEST},4\text{up}}$ but may potentially lag behind. Thus, to improve the overall

robustness of the inattention onset detection, subsequently, a stability criterion of the evolution of the threshold estimate is calculated based on adaptive boundaries. By fitting the psychometric function to the data (i.e., proportion of correct responses at presented stimulus levels) available until trial t , the evolution of the threshold estimate $\hat{x}_{T,t}$ is calculated (Fig. 2, top). After the simulated inattention onset at trial 31, the performance dictated by the altered psychometric function changes for the presented level, and the following ratio of correct versus incorrect responses drops. This leads to the subsequent increases of the level and influences the threshold estimate $\hat{x}_{T,t}$ in a similar way. This example shows spiky estimation errors in $\hat{x}_{T,t}$, which can occur predominantly towards the beginning of the sequence when only a small number of data points are available during the fitting process. Therefore, it is necessary to use a smoothed version $\tilde{x}_{T,t}$. The threshold estimate is smoothed with a moving average filter (unweighted mean over a span of 7 trials, span length based on trial-and-error optimization using computer simulations). This filter type was chosen due to its simplicity for implementation compared to other filters (e.g., zero-phase digital Butterworth filter). To deal with the endpoint, the array of threshold estimates is padded (by replicating the last available value $\hat{x}_{T,t}$ for 3 trials) prior to the smoothing process and truncated to the original length after smoothing. When using the method offline, the last available value $\hat{x}_{T,t}$ corresponds to the last trial of the experiment $\hat{x}_{T,\text{end}}$. When using the method online, the smoothing of the four trials ($\tilde{x}_{T,t-4}$ to $\tilde{x}_{T,t-1}$) prior to the current trial t is recalculated, because the threshold evolution is padded with new values of trial t . The mean $\tilde{\mu}_t$ and standard deviation $\tilde{\sigma}_t$ of the smoothed threshold estimate evolution $\tilde{x}_{T,t}$ are calculated for each trial t according to the following equations:

$$\tilde{\mu}_t = \frac{1}{W_t} \sum_{i=1}^t w_i \tilde{x}_{T,i} \quad \text{and} \quad (5)$$

$$\tilde{\sigma}_t^2 = \frac{1}{W_t} \sum_{i=1}^t w_i (\tilde{x}_{T,i} - \tilde{\mu}_t)^2, \quad (6)$$

using exponential weights w_i for the trials 1 to t with a forgetting rate $\varepsilon = 0.5$:

$$w_i = (1 - \varepsilon)^{t-i}, \quad (7)$$

normalized with W_t being the sum of all weights w_i up to trial t :

$$W_t = \sum_{i=1}^t w_i. \quad (8)$$

By using this exponential weighting, more recent values are taken into account more strongly and $\tilde{\sigma}_t$ adapts its size depending on how stable the estimate $\tilde{x}_{T,t}$ is (Fig. 2, bottom). This plot visualizes how the width of the stability

boundary becomes wider the more the threshold estimate varies (e.g., shortly after inattention onset) and the narrower the more stable it is (e.g., towards the end of the sequence). This can be used to define a threshold stability based on the stability boundary criterion: $\tilde{\mu}_{t+1} > \tilde{\mu}_t + 2\tilde{\sigma}_t$. The first time $\tilde{\mu}_{t+1}$ exceeds this boundary after the first occurrence of trial $t_{\text{PEST},4\text{up}}$ is defined as trial t_{exclude} (with symbol $[$ denoting the start of exclusion) after which trial data is excluded for the calculation of the final threshold estimate. The range $\pm 2\tilde{\sigma}_t$ corresponds to about 95% of the data and has shown to be a reasonable choice in preliminary simulations. Note that using an exponential weighting and two times the standard deviation are common choices for the Bollinger bands (Bollinger, 2001). Thus, in the shown example, the rapid change in $\tilde{\mu}_t$ originating from inattention can be detected shortly after inattention onset ($t_{\text{exclude}} = 33$).

During the inattention period, the stability boundary becomes wider due to the rapid divergence from the real threshold. Once the subject starts paying attention again, there follows a fast decrease of the PEST level, which still prevents the stability boundary to narrow. Therefore, the stability boundary criterion is not suitable to detect the end of the inattention period, and the detection has to rely on the adaptive PEST sequence. Analogous to $t_{\text{PEST},4\text{up}}$, the first trial followed by two consecutive PEST level decreases (eight correct responses would lead to two level decreases, considering $P_t = 75\%$ and $W = 1$) after t_{exclude} is identified and denoted as $t_{\text{PEST},2\text{down}}$. This trial $t_{\text{PEST},2\text{down}}$ simultaneously corresponds to the last trial to be excluded (t_{exclude}], where symbol $]$ denotes the end of exclusion). All following trials (including the two level decreases) are reconsidered for the calculation of the final threshold estimate and all trials from t_{exclude} to t_{exclude} are excluded. In the present example, despite the last trial of the modeled inattention being at 50, the last trial of exclusion t_{exclude} is at 57, before the PEST sequence starts decreasing and converges towards the true threshold. Data from the plateau before the level decrease is always excluded for the threshold estimate, as the first few trials of the plateau may be still partly influenced by inattention.

Computer simulations

Computer simulations are a valuable tool to evaluate the ability of the proposed method to correct for sustained inattention, as when modeling a subject, its psychometric function (also referred to as template) and modeled inattention are completely defined. Furthermore, today's computational power offers the possibility to simulate an extensive number of experimental runs to calculate a precise distribution of estimation errors.

Computer simulations were run with the following aims: (i) evaluate the performance of inattention correction in templates featuring different levels of modeled inattention, and (ii) validate that the proposed method does not significantly affect threshold estimates in absence of inattention. Thus, in the latter, the same templates were used, with inattention correction but without modeled inattention. The same simulation without inattention model but without applying the inattention correction method served as a baseline for the validation. To obtain representative results from the computer simulations (see, e.g., Kollmeier, Gilkey, & Sieben, 1988), simulated templates and parameters of the sampling procedure were selected based on our prior experimental knowledge and experience.

Methods

Templates

Twelve different subject templates covering a difference threshold space ($x_T: \in \{1, 2, 4\}^\circ$) and a slope space ($\beta_{\text{inflection}}: \in \{0.0625, 0.25, 0.5, 1\}^\circ$) were simulated. These values are representative for proprioceptive perception in healthy adults (Brewer, Fagan, Klatzky, & Matsuoaka 2005; Cappello et al., 2015; Choi, Meeuwse, & Arnhold, 1995; Elangovan, Herrmann, & Konczak, 2014; Fry-Welch, Campbell, Foltz, & Macek, 2003; Lamercy, Juárez Robles, Kim, & Gassert, 2011; Rinderknecht et al., 2014; Tan, Srinivasan, Reed, & Durlach, 2007). The guess rate γ was 0.5 and the lapse rate λ was 0 for all templates. In the following text, the short notation $\psi_{i,j}^\top$ (symbol $\top$ denotes a template) is used for templates, with index i denoting the threshold space parameter value, and index j the slope space parameter value (e.g., $\psi_{1,0.25}^\top$ for $x_T = 1^\circ$ and $\beta_{\text{inflection}} = 0.25^\circ$).

Inattention model

Inattention can be modeled in different ways: by changing the threshold (Cameron et al., 2002; Doll et al., 2015; Hall, 1983; Leek et al., 1991), the slope (Cameron et al., 2002), the lapse rate (Green, 1995; Wichmann & Hill, 2001), or a combination thereof (Cameron et al., 2002; Fründ et al., 2011). Depending on the concrete implementation of the psychometric function, modifying function-specific parameters may have different effects on the psychometric function (e.g., in the implementation by Green (1995) adding a lapse rate > 0 decreases the actual slope of the psychometric function without changing the slope parameter).

As the slope parameter of the psychometric function describes the reliability of stimulus detection or discrimination (Strasburger, 2001), it seems appropriate to decrease the slope β of the psychometric function to model inattention.

However, this would lead to higher performance for stimulus levels below the inflection point, which is contradictory to the fact that performance is assumed to decrease if the subject is inattentive. Consequently, the threshold parameter α would need to be increased to resolve this issue. While it also seems plausible that loss of attention would result in a higher number of lapses, the lapse rate λ could also be increased. As a matter of fact, increasing λ in our implementation decreases the actual slope $\beta_{\text{inflection}}$ (see Eq. 4) and increases the actual threshold $x_T = \psi^{-1}(0.75)$. Thus, the performance of the resulting proprioceptive function is lower compared to the performance of the original psychometric function, for all stimuli $x > 0$. Hence, we hypothesize that changing the lapse rate is the most suitable approach, which is in addition simpler than changing multiple parameters simultaneously.

A set of different lapse rates were simulated ($\lambda_{\text{inattention}} \in \{0, 0.25, 0.5\}^\circ$). While $\lambda_{\text{inattention}} = 0.5$ corresponds to complete inattention (i.e., guessing) and $\lambda_{\text{inattention}} = 0.25$ leads to a psychometric function with asymptotic value of 0.75, $\lambda_{\text{inattention}} = 0$ models a perfectly attentive subject.

Procedure

The following parameters were used for the PEST procedure: start level of 5.5° , start step of 2° , desired performance level P_t of 75%, and Wald sequential likelihood-ratio test parameter W of 1. These parameters proved successful in prior experiments assessing proprioceptive difference thresholds of finger joints (Rinderknecht et al., 2014). Each simulated experiment lasted for 120 trials and no other PEST termination rule was used. Observations from previous experiments (e.g., Rinderknecht et al. (2014) and Metzger et al. (2014) with healthy subjects and stroke patients) and ongoing proprioceptive assessments using PEST conducted by our group described subjects losing attention after 5–10 min of performing the task (corresponding to about 20–40 trials). This time interval is similar to the widely used claim in teaching guides that student's attention span ranges from 7–20 min (Bligh, 1998; David & Duket, 2009; Dent & Harden, 2013; Petty, 2004). However, this claim is mostly based on subjective observations and is supported by few well-controlled studies (Wilson & Korn, 2007). Nevertheless, it would be reasonable that the attention span in our 2AFC task is reduced, as the task is very repetitive and active participation is limited because the subjects receive only passive stimulation. Therefore, the onset of the modeled inattention was empirically chosen to follow a normal probability distribution centered around trial $t = 30$ with a standard deviation of 10 trials: $\mathcal{N}(30, 10)$. The sustained inattention period was modeled for 20 consecutive trials. Since the performance of the inattention correction method will be evaluated as a function of trials $\in [1, 120]$, the

upper bound of 120 trials only presents a limit for analyzing the evolution of the performance and should be sufficiently high to show its asymptotic behavior. Furthermore, based on practical experience in our group, it can be noted that 120 trials are towards the limit of practical feasibility in proprioceptive assessments using the two-interval 2AFC paradigm, as (i) the presentation of passive movements is time-consuming, and (ii) the task becomes cognitively very demanding and tiring when presented stimuli are around the perception threshold. In a clinical setting, a maximum of 60 trials corresponding to 15 minutes might be more realistic, as assessments are required to be quick to administer (Gresham et al., 1996). For each inattention model, the experiment was simulated 1000 times per template $\psi_{i,j}^T$, leading to a total of 36000 simulated experimental runs ($1000 \times 12\psi_{i,j}^T \times 3\lambda_{\text{inattention}}$). All simulations were performed in MATLAB R2014a (MathWorks, Natick, MA, USA).

Data analysis

For all simulated templates $\psi_{i,j}^T$ and modeled inattention levels, the absolute error (AE = average absolute estimation error) as a measure of overall estimation accuracy, the constant error (CE = average estimation error) as a measure of bias, and the variable error (VE = standard deviation of the estimates around the average estimate) as a measure of inconsistency (Schmidt & Lee, 2011) were calculated to evaluate the estimation performance in both cases, without and with inattention correction. The estimation error is calculated as the estimated threshold minus the threshold of the template. All threshold estimates larger than 20° were considered as outliers, as such large values are unreasonable in this specific scenario of finger proprioception. However, it is important to not disregard such poor estimates, as they affect the general quality of a psychophysical assessment. Thus, they were included when evaluating the inattention correction method. To reduce their severe effect on error metric statistics, threshold estimates were saturated to a ceiling value of 20° . The average mean $\pm$ SD in the individual evaluation metrics and the absolute and percentage improvements were calculated across all templates $\psi_{i,j}^T$ for each inattention model without and with inattention correction. In order to calculate the overall average absolute value of the constant errors for the performance evaluation across the different templates $\psi_{i,j}^T$, the sign was removed from each CE of each template $\psi_{i,j}^T$ by taking the absolute value ($|CE|$). Positive values describing improvement represent either reduction of variance or bias, and negative values a deterioration of estimation quality. Furthermore, the exclusion frequency for each trial and the distributions of trials at which the correction method starts excluding the data (t_{exclude}) and trials at

which data is reconsidered after the detected inattention period (t_{exclude}) were calculated. Before computing mean $\pm$ SD, PEST sequences were synchronized post hoc to the same modeled inattention onset, as the inattention onset was randomly distributed across the simulated runs. Additionally, percentages of exclusion onset and exclusion end before, during, or after modeled inattention, or where no exclusion at all occurred, were computed. All analyses were conducted in MATLAB R2014a.

Results

The resulting psychometric functions (without and with inattention correction) after 60 and 120 trials are shown in Fig. 3 together with the template $\psi_{1,0.25}^T$. In both plots it can be seen that, in the condition where no inattention correction is applied, the proportions of correct responses for the same levels is lower (e.g., at 2.20°), and that there is additional data with low performance at above-threshold levels (e.g., at 4.40°). This results in a downwards shift of the psychometric function creating a rightwards shift (i.e., increase) of the threshold estimate. Furthermore, it should be noted that in this specific example the estimated psychometric function without inattention correction matches the general shape (among others the actual slope and the lapse rate) of the template $\psi_{1,0.25}^T$ better than with inattention correction. The large lapse rate of the fit with inattention correction may seem counterintuitive. It may be caused mainly because of the interaction between the high performance at low level 0.90° and the proportion of correct responses at 2.22° which is lower than the psychometric function of the template. Nevertheless, correction for inattention improves the threshold estimate, especially if the maximum number of trials of an assessment is short (e.g., 60 trials, Fig. 3, top). The longer the assessment, or the more trials follow after the inattention period, the more the two estimates approach each other (e.g., 120 trials, Fig. 3, bottom).

The estimation errors for the different inattention models (i.e., different $\lambda_{\text{inattention}}$) for the conditions with and without inattention correction are shown in Fig. 4. The plots show that the estimation performance was severely affected by large $\lambda_{\text{inattention}}$, reaching errors of the same order of magnitude as the thresholds to be estimated, when inattention was not corrected for. By applying the method to correct for sustained inattention, these estimation errors could be reduced. However, they were not eliminated completely, and remained larger than in the optimal case, where perfect attention ($\lambda_{\text{inattention}} = 0$) was modeled and no correction was performed. The largest estimation errors were found around trial 50, which corresponds to the average of the distribution of the end of the inattention period (center of the normal random distribution $\mathcal{N}(30, 10)$ of the modeled inattention onset plus 20 trials of sustained

Fig. 3 Example of fitted psychometric functions without and with inattention correction from a single simulated run for template $\psi_{1,0.25}^T$ (for the same run as in Fig. 2) (Left) and threshold estimate distributions (bin resolution of 0.5°) from 1000 simulated runs of the same template (Right), after 60 (Top) and 120 trials (Bottom), with $\lambda_{\text{inattention}} = 0.5$. Left: Psychometric function of the template $\psi_{1,0.25}^T$ (dashed black sigmoid), estimated psychometric function without inattention correction (solid gray sigmoid), and estimated psychometric function with inattention correction (solid black sigmoid) after 60 trials. The thresholds of the two latter psychometric functions are indicated by vertical dashed lines in the respective colors. The proportions of correct responses are indicated as circles (gray: without inattention correction, black outlines: with inattention correction) for the different presented levels. The diameter is proportional to the number of presentations. The threshold of the template is indicated by the black triangle. Right: Comparison of threshold estimate distributions (bin resolution of 0.5°) without (gray fill) and with (black outline) inattention correction. The threshold of the template $\psi_{1,0.25}^T$ is indicated by the black triangle. Reported are the absolute error (AE), absolute value of the constant error (CE), and variable error (VE) of the estimate distributions

inattention). After this peak, the estimate errors decreased exponentially with the amount of additional data collected after the inattention period, independently of whether inattention was corrected for or not. Large oscillations in estimation performance during the first 20 trials result from the low number of data points (mostly at high levels with high performance rates) available for fitting the psychometric function.

The frequency of threshold estimates ceiling at 20° could be reduced by applying the inattention correction, as

Fig. 4 Estimation errors (mean $\pm$ SD of the absolute error (AE), absolute value of the constant error (|CE|), and variable error (VE) across the different templates $\psi_{i,j}^T$) for the different inattention lapse rates $\lambda_{\text{inattention}}$. The colored lines and shaded bands indicate the estimation errors after inattention correction for different $\lambda_{\text{inattention}}$ (Top: $\lambda_{\text{inattention}} = 0.5$, Middle: $\lambda_{\text{inattention}} = 0.25$, Bottom: $\lambda_{\text{inattention}} = 0$). For the same inattention models, the estimation errors without correction are represented by solid black lines. As a reference, the estimation errors are visualized for no modeled inattention (i.e., perfectly attentive subject) with dashed black lines. Note that for $\lambda_{\text{inattention}} = 0$, the two latter are identical

presented in Fig. 5. For $\lambda_{\text{inattention}} = 0$ and $\lambda_{\text{inattention}} = 0.25$, it was below 0.1% independent of the number of trials, and below 0.4% on average across different templates $\psi_{i,j}^T$ for $\lambda_{\text{inattention}} = 0.5$, when inattention was corrected for. In the condition where the inattention correction method was not applied, the frequency of ceiling effects was consistently higher, especially for higher inattention lapse rates (e.g., reaching up to 5% on average for $\lambda_{\text{inattention}} = 0.5$ for a sequence length of 50 trials).

The relative and absolute improvements with inattention correction in AE, |CE|, and VE are visualized in Fig. 6 for different $\lambda_{\text{inattention}}$. For complete inattention ($\lambda_{\text{inattention}} = 0.5$), there was an estimation performance

Fig. 5 Percentage of ceiling threshold estimates (mean $\pm$ SD across the different templates $\psi_{i,j}^T$) for the different inattention lapse rates $\lambda_{\text{inattention}}$ (Left: $\lambda_{\text{inattention}} = 0$, Middle: $\lambda_{\text{inattention}} = 0.25$, Right: $\lambda_{\text{inattention}} = 0.5$). The colored lines and shaded bands indicate the percentage of ceiling effects with inattention correction for different $\lambda_{\text{inattention}}$, and the solid black lines without inattention correction

Fig. 6 Improvement of the threshold estimation quality when using the inattention correction method vs. when no correction for inattention is performed. Top: Relative improvement (mean $\pm$ SD across the different templates $\psi_{i,j}^T$) of the absolute error (AE), absolute value of the constant error (ICEI), and variable error (VE) for the different inattention lapse rates $\lambda_{\text{inattention}}$. Bottom: Absolute improvement of AE, ICEI, and VE for the different $\lambda_{\text{inattention}}$

improvement of up to 62%, 77%, and 66% in the error metrics when correcting for inattention (maximum absolute improvement: 1.5° for AE, 1.7° for ICEI, and 2.8° for VE). Estimation performance decreased for all error metrics in the case of absence of inattention ($\lambda_{\text{inattention}} = 0$). However, when inspecting the absolute improvements (Fig. 6, bottom), this performance decrease of up to 0.06° for AE, ICEI, and VE was marginal. The large oscillations and negative peaks in the relative improvement (mainly for ICEI) originate from the division by constant errors near zero and were not representative for all templates $\psi_{i,j}^T$.

A summary of the distribution of excluded trials (with post hoc synchronized inattention onsets), as well as exclusion onset (t_{exclude}) and exclusion end (t_{exclude}) is presented in Fig. 7. It can be seen that for $\lambda_{\text{inattention}} = 0.5$ and $\lambda_{\text{inattention}} = 0.25$, trials around the end of the inattention period were excluded based on the correction method in about 51% and 17% of the simulated runs, respectively. It is also noteworthy that the frequency at which trials were excluded increased linearly for trials during the inattention period and decreased even more rapidly around two trials after the end of the modeled inattention period. In the condition without inattention ($\lambda_{\text{inattention}} = 0$) the frequency of exclusion increases steadily with increasing trial number until around trial 70 after “theoretical inattention onset”, but does not reach more than 13%. The distribution of detected inattention onset and end (Fig. 7, middle) corresponds to the derivative of the distribution of excluded trials (Fig. 7, top) and highlights the aforementioned properties. This plot also shows that there was a small number of simulated runs (around 0.3% for each trial) where inattention is erroneously detected outside (i.e., after) the modeled inattention period.

The number of onsets (t_{exclude}) and exclusion ends (t_{exclude}) before, during, or after the modeled inattention period, or where no exclusion occurred, are summarized in Fig. 7 (bottom). The inattention onset during the inattention period increased with higher inattention, and was detected in 45% of the simulated runs for $\lambda_{\text{inattention}} = 0.5$ and 16% for $\lambda_{\text{inattention}} = 0.25$ during the inattention period. For all inattention models, the inattention onset was falsely detected before the actual onset of inattention in less than 8% of all simulation runs. The frequency of cases where no exclusion occurred decreased with higher inattention lapse rates, and was 64% for $\lambda_{\text{inattention}} = 0$, 61% for $\lambda_{\text{inattention}} = 0.25$, and 32% for $\lambda_{\text{inattention}} = 0.5$.

Discussion

Through the computer simulations, we aimed to (i) investigate whether and by how much the proposed correction method for sustained inattention improves the threshold estimates in the presence of sustained inattention, and (ii) to validate that the inattention correction does not significantly bias the

Fig. 7 Distributions of data exclusion based on the inattention correction method with post hoc synchronized inattention period. *Top*: Percentage with which each trial is excluded. Reported are mean $\pm$ SD across the different templates $\psi_{i,j}^T$ for the different inattention lapse rates $\lambda_{\text{inattention}}$. The modeled inattention period is labeled and indicated by the gray bracket and gray dashed lines. *Middle*: Distributions of exclusion onset (t_{exclude}) and exclusion end (t_{exclude}). Note that the distribution of the exclusion end is visualized on the inverted axis to make start and end distributions better distinguishable on the same plot. *Bottom*: Grouped percentage of the exclusion onset and end occurrence (*before*, *during*, or *after* inattention period, or *no exclusion* if the correction method did not exclude any trials). For consistency, the exclusion end is again visualized on the inverted axis. Reported are mean $\pm$ SD

threshold estimates in absence of sustained inattention. The main results demonstrated that estimation errors could be reduced by up to 77% in case of complete inattention, and that in case of absence of inattention the errors were only marginally different from the condition where no inattention correction was applied. Furthermore, when correcting for inattention, there were very few estimates $> 20^\circ$ (outside of the reasonable range of thresholds for this scenario of finger proprioception), compared to when inattention was not corrected for, in which case it occurred in up to 9% of the simulated runs for some templates.

The excluded trials were predominantly within the modeled period of inattention, and the end of the simulated inattention period (t_{exclude}) could be well identified. The short lag of t_{exclude} (resulting in overestimation of the inattention period) can be explained by the Wald sequential likelihood-ratio test, as a minimum ratio of correct versus incorrect responses are required for the PEST level to be decreased. Since part of the trials at the stimulus level towards the end of the modeled inattention are still influenced by inattention (i.e., lower performance), a certain number of additional correct trials may be required after the end of the modeled inattention period to reach the condition for the PEST level to change. Since psychophysical assessments are based on probabilities and are not a fully deterministic process, it is possible that there is a small number of PEST sequences showing temporary divergences similar to sustained inattention, but which do not originate from the inattention model. However, as demonstrated by the results of the simulations, erroneous detection of inattention (in up to 8% of the runs) is not detrimental to the estimation quality. This can be supported by the fact that, in the case of absence of inattention, the estimation performance is only marginally changed, considering all erroneous detections before, during, and after the modeled inattention period (estimation quality decreased by less than 0.1°, which is at least one order of magnitude smaller than the thresholds, see Fig. 7 (bottom) for $\lambda_{\text{inattention}} = 0$). It should be noted that the number of inattention onset detections after the end of the modeled inattention period depends on the number of simulated trials (in this case 120 trials), as there remains a small probability of false detection which is summed up over the trials. Therefore, the reported frequencies in Fig. 7 should be interpreted with care. It may seem counterintuitive that in almost 32% of the simulated runs with $\lambda_{\text{inattention}} = 0.5$ and 61% with $\lambda_{\text{inattention}} = 0.25$ no trials were excluded despite the presence of modeled inattention. Yet, it should be considered that if the subject is distracted (i.e., forced to guess), this corresponds to a performance rate of 50%. As a consequence, the levels of a PEST sequence do not increase as rapidly as when all responses were incorrect, or as it would be the case in a yes–no detection or same–different discrimination paradigm. Therefore, the level increase may very well take more than eight trials (and even more for $\lambda_{\text{inattention}} = 0.25$), and thus may not reach four level increases within the 20 trials with modeled inattention. Furthermore, since the simulation of responses with and without modeled inattention is based on a stochastic process, there can be cases where the PEST level is decreased within the increasing step series, due to “lucky” guesses. The proposed method may not identify such events as inattention but consider them as normal oscillatory fluctuation towards the threshold. Another reason could be that PEST did not converge

sufficiently in the preceding trials and therefore the width of the threshold estimate stability boundary did not decrease enough to detect a drift in the estimate. Regardless, the high proportion of runs where no inattention was detected for $\lambda_{\text{inattention}} = 0.25$ does not present a strong limitation, as the estimation quality of uncorrected sequences suffers far less than for $\lambda_{\text{inattention}} = 0.5$.

The proposed method to correct for bias due to sustained inattention has several limitations. One major limitation is that sustained but short periods of inattention or intermittent inattention leading to less than four consecutive level increases will not be captured. The rule could be modified to $t_{\text{PEST},3\text{up}}$. However, preliminary simulations showed that the trade-off between an increased false positive rate and false-negative rate using $t_{\text{PEST},4\text{up}}$ lead to an overall better estimation performance. On the other hand, combining the present method with a free lapse rate during the fitting process, allows addressing not only longer, but also shorter inattention periods or isolated lapses, while keeping the bias marginal. In the current implementation of the inattention correction method, if the PEST start level is chosen too low (i.e., below the threshold to be estimated), early level increases will occur, which could be interpreted as inattention. Although this has been addressed by not allowing $t_{\text{PEST},4\text{up}} \leq 8$, this may not suffice in every case. Yet, for PEST to be efficient, experimenters should have a rough idea of the order of magnitude of the threshold to be estimated, in order to choose an appropriate start level and start step. A more robust solution may involve modifying the stability boundary to be also a function of trial t .

It is possible that in some cases the estimation quality of parameters besides the threshold (e.g., slope or lapse rate) of the psychometric function may decrease when using the inattention correction method (e.g., in the example shown in Fig. 3). However, PEST was only designed to converge towards the threshold and not to provide reliable slope estimates (Taylor, 1971). Therefore, for experiments with a small number of trials as in the present application, slope estimates are poorly estimated, even in absence of inattention. There exist other sampling procedures specifically designed to estimate the slope or multiple parameters (e.g., King-Smith & Rose, 1997; Kontsevich & Tyler, 1999; Levitt, 1971).

Using the computer simulations, only a finite set of templates $\psi_{i,j}^T$ were tested and evaluated. However, they cover a range similar to the performance of healthy individuals in assessments of joint proprioception thresholds using various paradigms (Brewer et al., 2005; Cappello et al., 2015; Choi et al., 1995; Elangovan et al., 2014; Fry-Welch et al., 2003; Lamercy et al., 2011; Rinderknecht et al., 2014; Tan et al., 2007). Moreover, the variability in relative improvement across templates with different thresholds and slopes

is one order of magnitude lower compared to the average relative improvement. Thus, the selected range of templates is representative enough to describe the performance of the inattention correction method.

The simulations were limited to complete ($\lambda_{\text{inattention}} = 0.5$), partial ($\lambda_{\text{inattention}} = 0.25$) and no inattention ($\lambda_{\text{inattention}} = 0$). For low values of λ (< 0.25), the absolute improvement is rather small: The absolute error for lapse rates smaller or equal to 0.25 are at maximum around 0.25° for thresholds in the range of 1° to 4° if the assessment is as short as 50 trials, and is reduced to $< 0.01^\circ$ for 120 trials. As the applying the proposed method for inattention correction to PEST sequences where there is no presence of inattention does not deteriorate the estimation quality, applying the method to simulated experimental runs with partial inattention ($\lambda \in (0, 0.5)$) should not lead to a estimation performance inferior to the case with no inattention. This is also supported by the results based on the simulated condition with partial inattention. Thus, while including a lapse rate estimate (Swanson & Birch, 1992; Wichmann & Hill, 2001) may be sufficient to reduce the estimation bias by improving the fit of the psychometric function for low values of λ , using the inattention detection method can be beneficial, not only populations known to suffer from inattention, but also in populations where cases with severe inattention are not common but may still happen sporadically.

In spite of the limitations, the proposed method goes beyond approaches identifying sequences of questionable validity such as the existing approaches by Hall (1983) and Leek et al. (1991). The results of the computer simulations demonstrated that, in contrast to other methods (Doll et al., 2015; Fründ et al., 2011), our method is capable of identifying specific inattention periods on a trial-by-trial basis, even within sequences of a small number of trials in a 2AFC experiment. Thus, it allows correcting the data by removing these biased trials, leading to an improvement of estimates. The benefit of applying this method on all recorded sequences of an experiment of this type outweighs the negligible biases it may introduce in case of no inattention. Improving threshold estimates, even if only possible for individual subjects, is especially important if these outcome measures are subsequently used in a clinical setting to determine optimal treatments. Furthermore, it offers a standardized and reproducible way to detect inattention and does not require experienced experimenters to subjectively classify data.

The heuristic rules and parameter values used here worked well for the purpose of the simulations and our specific behavioral data, and the performance of the inattention detection algorithm is not claimed to be optimal with these values. The parameter values could be tuned through

an extensive automated multidimensional optimization process objectively evaluating the performance in computer simulations. Despite being specifically tailored to this application using 2AFC and PEST for proprioceptive assessments, this new approach could provide some inspiration for similar algorithms to be used in experiments using other psychophysical paradigms or sampling procedures.

Behavioral data

As inattention is likely to be a strong confound in stroke patients (Rinne et al., 2013; Tuhim, 1993) and will negatively affect assessment quality, the proposed method to correct for sustained inattention was tested on a set of previously recorded data from a proprioceptive assessment in this population. The experiment consisted of two sessions (test and retest, inter-session time span between 1 and 4 days). The joint angle difference threshold at the metacarpophalangeal joint of the index finger of both hands was assessed in both sessions using a one-degree-of-freedom robotic device (Rinderknecht et al., 2014).

As the “real” thresholds and whether or not the subjects suffered from sustained inattention are unknown (in contrast to the computer simulations), the threshold estimation errors could not be calculated. However, if a significant inattention period occurred during either the test or the retest, one can expect the discrepancy $|\Delta|$ between the two assessments to increase. We hypothesize that by applying the inattention correction, the difference between test and retest outcomes will be reduced in those cases where “inattention” was detected, as the biasing effect of a present inattention period would be reduced.

Methods

Subjects

Data were collected from 20 stroke patients (> 2 weeks after their first clinical stroke, ten right hemisphere stroke, ten left hemisphere stroke, 12 male, eight female, age mean $\pm$ SD: 66.8 ± 8.3 years). Besides the inability to detect large passive finger movements, exclusion criteria comprised severe hand edema, high muscle tone, or pain, as it would hinder the correct use of the assessment device. Further exclusion criteria were severe cognitive impairment, aphasia and neglect. Eighteen subjects were right handed and two were ambidextrous, according to the Edinburgh Handedness Inventory (Oldfield, 1971). In accordance with the Declaration of Helsinki, all subject gave written informed consent prior to their participation. The study was approved by the institutional ethics committee of the University of Konstanz.

Procedure

A two-interval 2AFC paradigm was used in combination with PEST with a logarithmic domain adaptation (Rinderknecht et al., 2014). In each trial, two consecutive passive flexion movements were presented by the robotic device, both starting from a neutral position (i.e., 0° , hand aligned with forearm). Each flexion movement lasted 1 s, independent of the movement amplitude. The finger remained at the tested angle for 1.5 s, before moving back to neutral position at the same speed. After the presentation of the two different angles, the subject was asked to indicate which of the two angular movements was larger by providing their response directly with their other hand on a touchscreen. PEST defined the angular difference (referred to as level) between the two intervals. Each pair of presented angles was symmetrically arranged around 20° , and the order of larger versus smaller angle was randomized. By limiting the movements to flexion, the stimulus range was constrained to $[0^\circ, 40^\circ]$. The same PEST parameters and termination conditions as in the computer simulations were used (start level of 5.5° , start step of 2° , P_t of 75%, W of 1). Termination conditions were (i) a maximum number of 60 trials, (ii) 20 consecutive trials at the same level, or (iii) a minimum step of 0.1° . The tested limb was occluded from vision during the entire experiment. In each of the two sessions, proprioception of both hands was assessed in randomized order.

Data analysis

For each test and retest pair the absolute difference $|\Delta|$ between the two difference threshold estimates was calculated for both conditions, without and with inattention correction. These differences $|\Delta|$ were pooled across subjects and assessed hands. The level of impairment, as well as whether the contra- or ipsilesional hand was tested, was irrelevant for the purpose of evaluating the inattention correction method. Furthermore, no inferences were made about the performance of proprioceptive discriminability of the stroke patients, and assessing and pooling both hands allowed a larger dataset with subjects being likely to suffer from inattention. All analyses were conducted in MATLAB R2014a.

Results

Inattention was detected in 11 out of 80 assessment sequences. In those 11 sequences, 18 ± 8 trials were rejected on average, and the exclusion period started on average at trial 33 ± 9 (onset t_{exclude}). In six of those sequences the inattention period ended before the end of the assessment, whereas in the other five sequences

Fig. 8 Behavioral test-retest data from the proprioceptive assessment conducted for both hands in 20 stroke patients. *Left:* Absolute test-retest differences $|\Delta|$ for no inattention correction vs. when the algorithm was applied to detect and correct for sustained inattention. Test-retest differences from sequences which were not modified by the inattention correction method are indicated by circles. Where the algorithm had an effect on the threshold estimates, the $|\Delta|$ is indicated by triangles pointing downwards (decreased $|\Delta|$ with inattention correction) or upwards (increased $|\Delta|$ with inattention correction) connected by black vertical lines to the dashed diagonal identity line to better emphasize the change. In 8 of 11 inattention cases the test-retest discrepancy could be reduced by applying the inattention correction method. *Right:* Comparison of $|\Delta|$ -distributions (bin resolution of 0.5°) without (gray fill) and with (black outline) inattention correction. The overall decrease of variability when applying the inattention correction method is visible as the shift of the distribution towards $|\Delta| = 0^\circ$

the end of the assessment was reached before. Thus, in the latter cases only data of trials up to the inattention onset were used for the threshold estimate when applying the inattention correction method. Figure 8 illustrates the absolute differences $|\Delta|$ between test and retest for both conditions without and with inattention correction plotted against each other (left) and their distributions (right). In eight sequence pairs $|\Delta|$ was decreased (7/11 with an absolute change $\geq 0.1^\circ$, corresponding to the minimum step of the PEST algorithm), while in three sequence pairs $|\Delta|$ increased (2/11 with an absolute change $\geq 0.1^\circ$).

Discussion

The aim of applying the inattention correction algorithm on a real dataset from a behavioral experiment was to explore whether the proposed method would detect “inattention” in non-simulated data and whether it could decrease the test-retest discrepancy $|\Delta|$ in those cases, by excluding potentially biased data. Indeed, the algorithm detected “inattention” in 14% of the assessments. Correcting the data by removing trials during inattention periods provided more similar test-retest assessment results in 73% of the corrected assessment test-retest pairs. In the other few cases, the difference between test and retest was only marginally augmented (on average by 0.53°).

It is noteworthy that the two largest discrepancies were not reduced with the inattention correction method. Detailed

analysis of the PEST sequences revealed that in one case the PEST sequence increased by three steps right from the beginning before the PEST level decreased at trial 22. Due to the early increase, the PEST algorithm could not fully converge (defined by minimum step termination criterion) within the remaining trials of the assessment. What looked like suboptimally chosen start parameters of the sampling procedure, may have resulted from initial inattention or unfamiliarity with the task (e.g., the task may not have been fully understood), affecting performance, as the retest of the same hand showed immediate convergent behavior towards a threshold lower than the start level. Hence, it may be advisable to either increase the start level even more or conduct a few familiarization trials before the actual assessment.

In the other case of large $|\Delta|$, “inattention” was detected in the first of the two test-retest assessments and the difference threshold was reduced by about 15% by the algorithm. In the retest, no “inattention” was detected with the algorithm. Yet, experienced experimenters speculate that there could be an inattention period when visually inspecting the PEST sequence. Indeed, there are six consecutive PEST level increases. However, it was not captured by the inattention detection algorithm due to the stability boundary. When manually removing the visually identified “inattention” period, $|\Delta|$ decreased from around 10.6° to 1.7° . Further tuning or a decrease of the stability boundary width would allow detecting such a case, but to the detriment of false-negative detections.

Coincidentally, the average detected inattention onset and exclusion period length in this dataset with stroke patients were very similar to the values used for the inattention models in the computer simulations. Consequently, this might suggest that the way inattention was modeled in the computer simulations (onset distribution and inattention period) corresponds to a realistic scenario, despite being a simplified model (e.g., constant inattention level during the entire inattention period). Furthermore, considering that the task in the present proprioceptive assessments is highly repetitive and that stroke patients may suffer from increased fatigue (de Groot, Phillips, & Eskes, 2003; Duncan, Wu, & Mead, 2012; Staub & Bogousslavsky, 2001) and inattentiveness (Rinne et al., 2013; Tuhim, 1993), it seems reasonable that some subjects start losing attention to the task after 5–10 min (around trial 30), also considering literature pointing at attention spans of 7–20 min in students following lectures (Bligh, 1998; David & Dukette, 2009; Dent & Harden, 2013; Petty, 2004). Contributing factors to inattention during the proprioceptive assessment may also be (i) increased familiarity with the task and (ii) increased task difficulty (as the adaptive sampling procedure converges towards the threshold) leaving the subject with the potentially uncomfortable impression that he/she is guessing and thus uncertain about the correctness of the answer (Kaernbach, 2001;

Treutwein, 1995). Nevertheless, the way to model inattention (e.g., altering the psychometric function) should be validated in experimental studies when considering other applications with different psychophysical paradigms and stimuli.

Considering restrictions for administration time of clinical assessments in stroke patients (Gresham et al., 1996), or in clinical populations in general, only assessments with a small number of trials can be integrated into a clinical setting, for which reason the current proprioceptive assessment was limited to 60 trials. Based on the behavioral results with stroke patients, the average end of the inattention period was around trial 50 ± 11 , which is close to the end of the assessment. In the computer simulations it could be shown that, when the inattention correction is not applied, the estimation quality is poorest if the assessment is stopped right after the end of the inattention period, as the ratio of biased versus unbiased data is the largest. Thus, using the proposed method to address inattention confounds might be even more essential in such scenarios where the assessments have to be short.

Conclusions

Loss of attention presents challenges in threshold testing which have not been sufficiently addressed. For instance, inattention can affect perception during sensory assessments and bias psychophysical threshold estimates. We proposed a method capable of efficiently identifying periods of sustained severe inattention based on the recorded stimulus-response pairs of a PEST sequence together with the evolution of the estimated threshold. This algorithm allows excluding specific trials with biased data due to inattention and improving the threshold estimates, as evaluated in computer simulations and tested on a set of behavioral data, which has important implications for the psychometric properties of an assessment, such as reliability and validity. The method can also be used for sequences with relative small numbers of trials and in experiments using a 2AFC paradigm. Furthermore, the inattention correction method complements existing methods addressing lapses and goodness of fit (e.g., including a lapse rate when fitting the psychometric function), such that not only isolated lapses but also longer inattention periods can be addressed to improve the quality of the outcome measure. The proposed method can be used both offline to correct previously recorded experimental data post hoc as well as online (with some limitations) to generate an alert of inattention, and pause or stop the experiment to avoid further collection of biased data. When resuming the experiment, the sampling procedure could continue at the stimulus level at which the inattention onset was detected.

This may allow shortening experimental procedures, which could have a significant impact on the improvement of assessment practice, especially in studies with patient populations potentially suffering from attention deficits and where experimental time is costly, while providing more reliable estimates. In addition, it could be shown that the benefit of correcting for biases is even higher when assessments are short, as in these cases a sustained inattention period has a larger detrimental effect on the estimation quality. While the statistics based on the results of the computer simulations have shown that there is an overall improvement of estimation quality at the group level when applying inattention correction, it can also positively impact diagnosis, prognosis, or treatment planning for the individual patient if inattention confounds can be removed from the assessment's outcome measures.

Acknowledgments The authors would like to thank J. Liepert and V. Raible from the Kliniken Schmieder Allensbach for their collaboration and acquisition of the behavioral data, as well as K. Leuenberger for fruitful discussions regarding the development of the algorithm. This research was supported by the ETH Zurich Foundation in collaboration with Hocoma AG, by the Janggen-Pöhn Foundation, and the Schmieder Foundation for Science and Research.

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

Compliance with ethical standards

Conflict of interests 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.

References

- Bligh, D. A. (1998). *What's the Use of Lectures?* Intellect books.
- Bollinger, J. (2001). *Bollinger on Bollinger bands*. USA: McGraw Hill Professional.
- Brewer, B. R., Fagan, M., Klatzky, R. L., & Matsuoka, Y. (2005). Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(1), 1–11.
- Bruhn, P., & Parsons, O. A. (1971). Continuous reaction time in brain damage. *Cortex*, 7(3), 278–291.
- Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42(8), 949–967.
- Cappello, L., Elangovan, N., Contu, S., Khosravani, S., Konczak, J., & Masia, L. (2015). Robot-aided assessment of wrist proprioception. *Frontiers in Human Neuroscience*, 9, 198.
- Carrasco, M. (2006). Covert attention increases contrast sensitivity: psychophysical, neurophysiological and neuroimaging studies. *Progress in Brain Research*, 154, 33–70.

- Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, 7(3), 308–313.
- Choi, S. O., Meeuwse, H. J., & Arnhold, R. W. (1995). On the psychophysics of arm-positioning movements. *Perceptual and Motor Skills*, 80(3 Pt 2), 1163–1169.
- Cohen, M. R., & Maunsell, J. H. R. (2011). When attention wanders: how uncontrolled fluctuations in attention affect performance. *Journal of Neuroscience*, 31(44), 15802–15806.
- David, C., & Dukette, D. (2009). *The essential 20: Twenty components of an excellent health care team*. Pittsburgh: Dorrance Publishing.
- de Groot, M. H., Phillips, S. J., & Eskes, G. A. (2003). Fatigue associated with stroke and other neurologic conditions: implications for stroke rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 84(11), 1714–1720.
- Dent, J., & Harden, R. M. (2013). *A practical guide for medical teachers*. Elsevier Health Sciences.
- Doll, R. J., Veltink, P. H., & Buitenweg, J. R. (2015). Observation of time-dependent psychophysical functions and accounting for threshold drifts. *Attention, Perception, & Psychophysics*, 77(4), 1440–1447.
- Duncan, F., Wu, S., & Mead, G. E. (2012). Frequency and natural history of fatigue after stroke: A systematic review of longitudinal studies. *Journal of Psychosomatic Research*, 73(1), 18–27.
- Elangovan, N., Herrmann, A., & Konczak, J. (2014). Assessing proprioceptive function: evaluating joint position matching methods against psychophysical thresholds. *Physical Therapy*, 94(4), 553–561.
- Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, 11(6), 16.
- Fry-Welch, D., Campbell, J., Foltz, B., & Macek, R. (2003). Age-Related Changes in upper extremity kinesthesia. *Physical & Occupational Therapy in Geriatrics*, 20(3–4), 137–154.
- García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, 38(12), 1861–1881.
- Gescheider, G. (1997). *Psychophysics: The fundamentals*. New Jersey: Lawrence Erlbaum Associates.
- Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *The Journal of the Acoustical Society of America*, 97(6), 3749–3760.
- Green, D. M., Richards, V. M., & Forrest, T. G. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *The Journal of the Acoustical Society of America*, 86(2), 629–636.
- Gresham, G., Duncan, P., Stason, W., Adams, H., Adelman, A., Alexander, D., ..., Trombly, C. (1996). Post-stroke rehabilitation: assessment, referral, and patient management. Quick Reference Guide for Clinicians, Number 16. *Journal of Pharmacoepidemiology*, 5(2), 35–63.
- Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *The Journal of the Acoustical Society of America*, 69, 1763.
- Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *The Journal of the Acoustical Society of America*, 73(2), 663–667.
- Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63(8), 1377–1388.
- King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, 37(12), 1595–1604.
- King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Research*, 34(7), 885–912.
- Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America*, 83(5), 1852–1862.
- Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16), 2729–2737.
- Lamercy, O., Juárez Robles, A., Kim, Y., & Gassert, R. (2011). Design of a robotic device for assessment and rehabilitation of hand sensory function. In *2011 IEEE International Conference on Rehabilitation Robotics (ICORR)*, (pp. 1–6). Zurich.
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, 63(8), 1279–1292.
- Leek, M. R., Hanna, T. E., & Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *The Journal of the Acoustical Society of America*, 90(3), 1385–1397.
- Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B), 467–477.
- Linares, D., & López-Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *R Journal*, 8(1), 122–131.
- Macmillan, N. A., & Douglas Creelman, C. (2005). *Detection Theory: A User's Guide*. New Jersey: Lawrence Erlbaum Associates.
- McKee, S. P., Klein, S. A., & Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, 37(4), 286–298.
- Metzger, J.-C., Lamercy, O., Califfi, A., Dinacci, D., Petrillo, C., Rossi, P., ..., Gassert, R. (2014). Assessment-driven selection and adaptation of exercise difficulty in robot-assisted therapy: a pilot study with a hand rehabilitation robot. *Journal of NeuroEngineering and Rehabilitation*, 11(1), 154.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97–113.
- Petty, G. (2004). *Teaching today: A practical guide*. Cheltenham: Nelson Thornes.
- Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, 12(6), 25.
- Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of vision*, 13(7), 3.
- Prins, N., & Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data.
- Prokasy, W. F., & Raskin, D. C. (1973). *Electrodermal activity in psychological research*. New York: Academic Press.
- Rinderknecht, M. D., Popp, W. L., Lamercy, O., & Gassert, R. (2014). Experimental Validation of a Rapid, Adaptive Robotic Assessment of the MCP Joint Angle Difference Threshold. In Auvray, M., & Duriez, C. (Eds.) *Haptics: Neuroscience, Devices, Modeling, and Applications, Lecture Notes in Computer Science*, (pp. 3–10). Berlin: Springer Berlin Heidelberg.
- Rinderknecht, M. D., Lamercy, O., Raible, V., Liepert, J., & Gassert, R. (2017). Age-based model for metacarpophalangeal joint proprioception in elderly. *Clin Interv Aging*, 12, 635–643.
- Rinne, P., Hassan, M., Goniatakis, D., Chohan, K., Sharma, P., Langdon, D., ..., Bentley, P. (2013). Triple dissociation of attention networks in stroke according to lesion location. *Neurology*, 81(9), 812–820.
- Schmidt, R. A., & Lee, T. (2011). *Motor control and learning*, (5th ed.). Champaign: Human Kinetics.
- Shen, Y., & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: thresholds, slopes, and lapses of attention. *The Journal of the Acoustical Society of America*, 132(2), 957–967.
- Staub, F., & Bogousslavsky, J. (2001). Fatigue after stroke: a major but neglected issue. *Cerebrovascular Diseases*, 12(2), 75–81.

- Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception & Psychophysics*, 63(8), 1348–1355.
- Swanson, W. H., & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, 51(5), 409–422.
- Tan, H. Z., Srinivasan, M. A., Reed, C. M., & Durlach, N. I. (2007). Discrimination and identification of finger joint-angle position using active motion. *ACM Transactions on Applied Perception (TAP)*, 4(2), 10.
- Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *The Journal of the Acoustical Society of America*, 49(2), Suppl 2:505–Suppl:2:508.
- Taylor, M. M., & Douglas Creelman, C. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, 41, 782.
- Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, 35(17), 2503–2522.
- Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, 61(1), 87–106.
- Tuhrim, S. (1993). Medical therapy of ischemic stroke. In Gordon, W. A. (Ed.) *Advances in Stroke Rehabilitation, chapter Medical therapy of ischemic stroke*, (pp. 3–15). London: Andover Medical Publishers.
- van Zomeren, A. H., & van den Burg, W. (1985). Residual complaints of patients two years after severe head injury. *Journal of Neurol Neurosurg Psychiatry*, 48(1), 21–28.
- Wald, A. (1947). *SEquential analysis*. New York: Wiley.
- Watson, C. S. (1980). Time course of auditory perceptual learning. *Annals of Otology, Rhinology, and Laryngology Supplement*, 89(5 Pt 2), 96–102.
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8), 1293–1313.
- Wilson, K., & Korn, J. H. (2007). Attention during lectures: Beyond ten minutes. *Teaching of Psychology*, 34(2), 85–89.