Behavior Research Methods
https://doi.org/10.3758/s13428-019-01301-6

The multivariate adaptive design for efficient estimation of the time course of perceptual adaptation

Ping Chen[1] · Steve Engel[2] · Chun Wang[3]

Abstract

In experiments on behavioral adaptation, hundreds or even thousands of trials per subject are often required in order to accurately recover the many psychometric functions that characterize adaptation's time course. More efficient methods for measuring perceptual changes over time would be beneficial to such efforts. In this article, we propose two methods to adaptively select the optimal stimuli sequentially in an experiment on adaptation: These are the minimum entropy (ME) method and the match probability (MP) method. The ME method minimizes the uncertainty about the joint posterior distribution of the function parameters at each trial and is mathematically equivalent to Zhao, Lesmes, and Lu's (2019) method, which efficiently measures time courses of perceptual change by maximizing information gain. The MP method selects the next stimulus that makes the value of the psychometric function closest to .5—that is, where the probability of choosing either one of the two options for each stimulus is closest to .5. We extended Zhao et al.'s (2019) work by evaluating the ME method in a new domain (contrast adaptation) with two simulation studies that compared it to MP and two other methods (i.e., traditional staircase and random methods), and also explored the optimal block length. ME outperformed the other three methods in general, and using fewer longer blocks generally produced better parameter recovery than using more shorter blocks.

Keywords Adaptive design · Minimum entropy · Perceptual adaptation · Time course · Tilt aftereffect

The properties of most biological systems, including human vision, change over time. Many such changes are functionally significant and take place on a reasonably rapid timescale—seconds to minutes—that enable them to be studied in the laboratory. Visual adaptation consists of a number of processes acting over this timescale to help the visual system optimize its performance within a given environment (e.g., Clifford & Rhodes, 2005). Examples include dark adaptation (e.g., Pugh, Nikonov, & Lamb, 1999), the development of color afterimages (e.g., Zaidi, Ennis, Cao, & Lee, 2012), and the well-studied phenomenon of contrast adaptation (e.g., Clifford et al., 2007). In this latter form of adaptation, visual neurons reduce their responsiveness following exposure to their preferred stimuli, which are typically patterns of high and low luminance, characterized by their ratio, or contrast. Behavioral measures of perceptual sensitivity and bias show corresponding changes over time, as documented in a long history of psychophysical experiments.

Whereas many experiments simply measure effects or aftereffects of adaptation, others focus on measuring the time course of the adaptive changes (e.g., Mei, Dong, & Bao, 2017; Patterson, Wissig, & Kohn, 2013; Pavan, Marotti, & Campana, 2012). For example, it can be important to characterize the conditions under which adaptation is faster or slower, or whether certain individuals adapt more slowly or rapidly. The main challenge to studying the time course of visual adaptation is that experiments generally take a long time; repeated observations are required over a large portion of the time course, which itself can run many minutes. Hence, efficient methods to measure changes over time may be particularly valuable in this case.

In a typical psychophysical experiment on adaptation, one measures perception as a function of small changes in a test stimulus (a psychometric function), and observes how parameters of this function change over time. Stimuli are presented and responses are made in discrete events called "trials." In between trials, observers are exposed to an environment intended to cause changes in perception and the corresponding

✉ Chun Wang
wang4066@uw.edu

[1] Beijing Normal University, Beijing, China

[2] University of Minnesota, Minneapolis, MN, USA

[3] University of Washington, Seattle, WA, USA

Published online: 01 November 2019

psychometric function. In contrast adaptation, this adapting environment is typically a high contrast pattern of stripes (called a grating) and the test stimulus is another grating whose properties vary under experimental control. The main property of the test that is varied corresponds to the x axis of the psychometric function, and it usually controls task difficulty—for example the intensity or contrast of the grating. Fifty to a hundred of trials are often required to estimate the desired parameters of the psychometric function with reasonable precision. Multiplying this by the number of time points required to trace out the time course of adaptation can cause experiments to take thousands of trials per subject. Here, we study two newly developed methods for efficiently measuring the time course of changes in perception.

In the 1970s, a Bayesian adaptive psychometric method called QUEST (Watson & Pelli, 1983) was developed to efficiently estimate psychometric functions, by placing each trial at the current most probable estimate of key parameters. A simpler, earlier technique is the staircase procedure that does not assume a functional form for the psychometric function (Cornsweet, 1962; Levitt, 1971). A one-up, one-down staircase reduces the stimulus level when the subject's response is positive and increases the stimulus level when the response is negative. Although these earlier approaches are capable of recovering parameters of a single psychometric function, the method was recently extended to allow simultaneous efficient estimation of many functions: The quick contrast sensitivity function (CSF) method (Lesmes, Lu, Baek, & Albright, 2010) simultaneously recovers parameters for a family of psychometric curves that vary with the spatial frequency (bar width) of the grating. The family of curves is called the contrast sensitivity function, which is a useful measure of visual performance in normal and disease states, and it has a characteristic functional form (shape). The quick CSF method gains its efficiency by estimating parameters of this functional form rather than estimating parameters of individual psychometric curves directly.

In this study, we propose two similar approaches to efficiently estimate the family of psychometric functions that specify the time course of adaptive changes. First, we use a minimum entropy method to adaptively select optimal stimuli to quickly and accurately recover the time course of adaptation for each subject. This method is inherited from the psychometrics literature on multidimensional computerized adaptive testing (MCAT), which in turn was heavily influenced by the advances in statistics and information science (Wang & Chang, 2011). Second, we adopt another commonly used multivariate adaptive method in CAT, namely the match probability method, due to its simplicity and efficiency. Specifically, this method selects the next stimulus such that the value of psychometric function (i.e., probability of yes/no judgment or correct/wrong response) is closest to .5. Both these methods are called multivariate adaptive methods because multiple target parameters of interest are updated during the course of experiment, and all different parameters need to be recovered precisely.

Given that the time course of visual adaption, like the contrast sensitivity function, has a typical form that is characterized by a few parameters (asymptote, height, and rate), the primary goal of our methods is to estimate these key parameters. Take the minimum entropy method as an example, this method minimizes uncertainty (entropy) about the joint posterior distribution of the key parameters at each trial; this allows it to rapidly shrink the uncertainty with respect to the entire joint distribution of the unknown parameters. We use a Bayesian approach that takes into account all the observed information from preceding trials to update the joint posterior distribution trial by trial, which increases efficiency. At the end of the experiment, an expected a posterior (EAP) estimate is computed to get the point estimates of the key parameters, resulting in an estimated visual adaption curve.

The idea of using adaptive stimulus selection and a functional form to recover the time course of changes in perception is not new. The approach first appeared in an abstract (Zhao, Lesmes, & Lu, 2017) and more recently in a complete article from the same group (Zhao, Lesmes, & Lu, 2019). Their method selects stimuli that maximize the information gain in the next trial, and has been applied to characterize perceptual sensitivity changes that result from large changes in light level, a process called dark adaptation. Mathematically, maximizing the information gain is equivalent to minimizing the entropy of the joint posterior distribution, which makes our proposed minimum entropy method formally very close to this past work.

Here we build upon the work of Zhao, Lesmes, and Lu (2019), to carefully evaluate the performance of the minimum entropy approach and compare it to the match probability method and other methods. We do this in two simulation studies, which also shed light on how optimal designs vary as a function of the specific shape of the adaptation time course that is being estimated. In particular, we extend the minimum entropy method from the domain of light adaptation to the domain of contrast adaptation, which has a differently shaped psychometric function than that explored in the previous work. We provide a different and simplified computational framework in which the optimal stimuli are determined. In addition, we consider the easy-to-implement match probability method as a fast alternative and evaluate its performance in the domain of contrast adaptation.

The rest of the article is organized as follows. We first introduce the psychometric functions used in the study, followed by the Bayesian parameter estimation method (i.e., EAP) and the adaptive selection of stimulus. Then we present two comprehensive simulation studies with detailed designs and results. A discussion is given in the end.

Method

Psychometric models

The work here will involve simulated experiments on a form of contrast adaptation called the tilt aftereffect: Viewing one grating causes another grating to appear tilted away from it. In experiments on the tilt aftereffect, observers are generally asked to indicate whether a test grating appears tilted to the left or to the right of some standard, and a psychometric function is formed by computing the percentage of responses in one direction as a function of the orientation of the test. The location on the curve that passes through the 50% point is where the test grating appears the same as the standard, and is referred to as the "point of subjective equality." In between test trials, an adapting grating is shown, and the point of subjective equality evolves over time. The experiments below simulate effects of a vertical adapting grating on a 45-deg test grating, conditions that match experiments run in one of the author's labs.

The basic task is a yes/no (or two-alternative) judgment, given at different time points. At each time point, the probability of choosing "yes" follows the logistic form as follows,

$$P(Y=1)=\gamma+(1-\gamma-\lambda)\frac{1}{1+e^{-\beta(x-\alpha(t))}} \tag{1}$$

where

$$\alpha(t)=a-be^{-ct} \tag{2}$$

In Eq. 1, "x" denotes the feature of the stimulus that is manipulated during the experiment. In the present context, "x" represents the orientation of a patch of bars. For a fixed t, Eq. 1 results in an S-shape logistic curve, where β denotes the slope, γ and λ are the "guessing rates" that shift the asymptotes of the function for small and large values of x. γ can be considered as the lower asymptote whereas $1-\lambda$ is the upper asymptote. $\alpha(t)$ is the threshold, also known as the "point of subjective equality," which varies as a function of time t.

The key goal of the experiment is to estimate $\alpha(t)$ as a function of time points based on subjects' binary responses on repeated trials. The trials are grouped into different blocks, each block lasts 1 or 2 min, and within a block, the value of t varies from 1 to 60 or 120 with a step size of 3 s (i.e., one trial will be done at a fixed rate every 3 s). It is assumed that there are J repeated blocks and T discrete time points (i.e., the number of trials) within a block. Figure 1 provides an illustration of the experimental design.

The target function in Eq. 2 takes the exponential form, where a is the asymptote, b controls its height, and c is the rate of growth or decay. For this study, we will only consider the blocks of adaptation where b is positive, but the same methods could easily be extended to the scenario where "b" is negative.

In experimental design, we can only vary x, and for the present example we will consider effects of adaptation on a diagonal (45-deg) grating, causing x to vary between 41 and 55. For simplicity, we assume $\gamma=\lambda$, and assume b is positive for now. Given the observed responses (binary), the goal is to find appropriate x for each value of t (t = 1, 2, . . . , T) to maximize the estimation precision of person parameters $\boldsymbol{\theta}=(a, b, c, \beta, \lambda)$. We later fixed β and λ to be known because otherwise, the number of trials will be much larger to properly recover a five-dimensional $\boldsymbol{\theta}$.

Update of $a(t)$: Expected a posteriori (EAP)

The EAP is a Bayesian estimator that finds the posterior *mean* as the point estimate of the target parameter. As compared to maximum likelihood estimator (MLE), EAP takes advantage of the prior information on the parameters of interest, and hence it usually yields smaller standard errors than MLE. As compared to maximum a posteriori (MAP) that finds the posterior *mode*, EAP is relatively more stable (Wang, 2015). Moreover, because EAP uses the information from the entire posterior distribution rather than just its mode, EAP is sometimes more efficient than MAP. In behavior research, the EAP, although not explicitly called this name, has been used by Zhao et al. (2019).

To compute EAP, for a given subject i, the likelihood function for (a_i, b_i, c_i) can be expressed as

$$L_i(\boldsymbol{y})=\prod_{j=1}^{J}\prod_{t=1}^{T}P(\alpha_i(t),\beta,\lambda)^{y_{i,j(t)}}(1-P(\alpha_i(t),\beta,\lambda))^{1-y_{i,j(t)}} \tag{3}$$

where $P(\alpha_i(t), \beta, \lambda)$ takes the form in Eq. 1; $\boldsymbol{y}$ denotes the response vector of subject i to the $J\times T$ stimuli, and $y_{i,\,j(t)}=1$ if subject i chooses "yes" at time point t within the jth block and $y_{i,\,j(t)}=0$ otherwise. Then the EAP estimates of (a_i, b_i, c_i) is

$$\hat{a}_i^{EAP}=\frac{\int a\left[\iint L_i(\boldsymbol{y})\pi(a,b,c)dbdc\right]da}{\iiint L_i(\boldsymbol{y})\pi(a,b,c)dadbdc}, \tag{4}$$

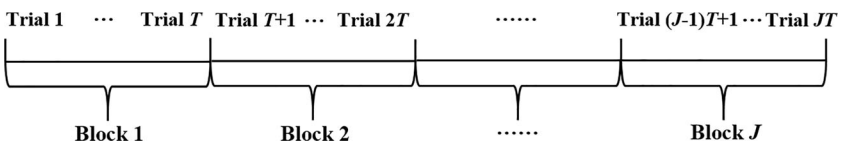

Fig. 1 An illustration of the experimental design.

$$\hat{b}_i^{EAP} = \frac{\int b\left[\iint L_i(\boldsymbol{y})\pi(a,b,c)dadc\right]db}{\iiint L_i(\boldsymbol{y})\pi(a,b,c)dadbdc}, \tag{5}$$

$$\hat{c}_i^{EAP} = \frac{\int c\left[\iint L_i(\boldsymbol{y})\pi(a,b,c)dadb\right]dc}{\iiint L_i(\boldsymbol{y})\pi(a,b,c)dadbdc}. \tag{6}$$

The multiple integrations in Eqs. 4–6 could be done numerically via Monte Carlo integration (e.g., Chen, Wang, Xin, & Chang, 2017; Tuerlinckx, Rijmen, Verbeke, & De Boeck, 2006; Wang, 2015). Take (4) as an example, sample M data points $(a^{(m)}, b^{(m)}, c^{(m)})$ $(m = 1, 2, \ldots, M)$ from their joint prior distribution $\pi(a, b, c)$, then $\hat{a}_i^{EAP} \approx \frac{\sum_{m=1}^{M} a^{(m)} L_i\left(\boldsymbol{y}|a^{(m)},b^{(m)},c^{(m)}\right)}{\sum_{m=1}^{M} L_i\left(\boldsymbol{y}|a^{(m)},b^{(m)},c^{(m)}\right)}$; the standard error of $\hat{a}_i^{EAP}$ can be computed as $SE\left(\hat{a}_i^{EAP}\right) = \left(\frac{\int\left(a-\hat{a}_i^{EAP}\right)^2\left[\iint L_i(\boldsymbol{y})\pi(a,b,c)dbdc\right]da}{\iiint L_i(\boldsymbol{y})\pi(a,b,c)dadbdc}\right)^{1/2} \approx \left(\frac{\sum_{m=1}^{M}\left(a^{(m)}-\hat{a}_i^{EAP}\right)^2 L_i\left(\boldsymbol{y}|a^{(m)},b^{(m)},c^{(m)}\right)}{\sum_{m=1}^{M} L_i\left(\boldsymbol{y}|a^{(m)},b^{(m)},c^{(m)}\right)}\right)^{1/2}$. Efficient matrix programming is implemented in MATLAB, which makes the computation extremely fast. Details on how to sample data points $(a^{(m)}, b^{(m)}, c^{(m)})$ from $\pi(a, b, c)$ will be introduced in the Simulations Studies section.

In addition to the point estimate of $\alpha(t)$ [i.e., $\hat{\alpha}(t)$], we must also report the standard error of $\hat{\alpha}(t)$ and the 95% confidence band of $\hat{\alpha}(t)$ to quantify the uncertainty around the point estimates. Because $\hat{\alpha}(t) = \hat{a}^{EAP} - \hat{b}^{EAP} e^{-\hat{c}^{EAP} t}$ (the superscript *EAP* is dropped hereafter) and it is assumed a, b and c are independent, according to the delta method,

$$\mathrm{var}\left(\hat{\alpha}(t)\right) = \begin{pmatrix} 1 & -e^{-\hat{c}t} & t\hat{b}e^{-\hat{c}t} \end{pmatrix} \begin{pmatrix} var\left(\hat{a}\right) & 0 & 0 \\ 0 & var\left(\hat{b}\right) & 0 \\ 0 & 0 & var\left(\hat{c}\right) \end{pmatrix}$$

$$\begin{pmatrix} 1 & -e^{-\hat{c}t} & t\hat{b}e^{-\hat{c}t} \end{pmatrix}^T = var\left(\hat{a}\right) + \left(e^{-\hat{c}t}\right)^2 var\left(\hat{b}\right) + \left(t\hat{b}e^{-\hat{c}t}\right)^2 var\left(\hat{c}\right).$$

On this basis, the 95% confidence interval of $\hat{\alpha}(t)$ can be expressed as:

$$\left(\hat{\alpha}(t) - 1.96 \times \sqrt{\mathrm{var}\left(\hat{\alpha}(t)\right)}, \hat{\alpha}(t) + 1.96 \times \sqrt{\mathrm{var}\left(\hat{\alpha}(t)\right)}\right).$$

Adaptive designs

The main purpose of adaptive design is to select stimuli that can provide maximal information to quickly recover each subject's $\alpha(t)$, or in other words, (a, b, c). Several adaptive procedures in psychophysical research have been proposed, and in a review by Leek (2001), he grouped the methods into three general categories defined by their systems for placing trials along a stimulus array as well as the estimation procedure. They are (1) parameter estimation by sequential testing (PEST), which is an algorithm for threshold searching that changes both step sizes and direction across a set of trials (Hall, 1981), (2) maximum likelihood procedures, and (3) staircase procedures. Most recently, Lesmes et al. (2015) proposed Bayesian adaptive estimation of CSF based on the idea of minimizing the expected entropy (e.g., Cobo-Lewis, 1996; Kujala & Lukka, 2006). Similar to Lesmes et al. (2015), the same Bayesian method will be adopted here, but the target function is not CSF but rather the time-series of perceptual adaptation function in Eq. 2.

Not that in addition to psychophysical research, adaptive design has also been used for cognitive modeling and psychometrics research. In the former context, Cavagnaro et al. (2010) studied a mutual information based approach to model discrimination in cognitive science (Cavagnaro, Myung, Pitt, & Kujala, 2010; Cavagnaro, Pitt, & Myung, 2011; Myung, Cavagnaro, & Pitt, 2013). In the latter context, the continuous entropy method (Wang & Chang, 2011) was studied in multidimensional adaptive achievement tests.

Minimum entropy (ME) method Shannon entropy measures the uncertainty inherent in the distribution of a random variable (Shannon, 1948). When the random variable follows a continuous distribution, Shannon entropy becomes continuous entropy. In Bayesian framework, each target parameter is considered as a random variable, hence the uncertainty of the parameter (or parameter vector) is quantified by the entropy of its posterior density.

During adaptive stimulus selection, to maximize the estimation precision of the target parameters, the next administered stimulus should be the one that can minimize the uncertainty of the parameter estimates. This is the principle idea of the ME method, which intends to select the next stimulus to minimize the expected entropy (and hence uncertainty) of the posterior density of the target parameters. The ME method is specifically described as follows. Denote the posterior distribution of (a, b, c) after $(k–1)$ trials as $p\left(a, b, c|\boldsymbol{y}^{k-1}\right) = \frac{L\left(\boldsymbol{y}^{k-1}|a,b,c\right)\pi(a,b,c)}{\iiint L\left(\boldsymbol{y}^{k-1}|a,b,c\right)\pi(a,b,c)dadbdc}$, then the corresponding posterior continuous entropy becomes

$$H\left(p\left(a,b,c|\boldsymbol{y}^{k-1}\right)\right) = \int p\left(a,b,c|\boldsymbol{y}^{k-1}\right)\log\left(\frac{1}{p\left(a,b,c|\boldsymbol{y}^{k-1}\right)}\right)d(a,b,c) \tag{7}$$

where a three-dimensional integration is taken over the distribution of (a, b, c).

Then, the predicted (or expected) posterior continuous entropy after administering the kth stimulus is,

$$E_k\left[H\left(p\left(a,b,c|\boldsymbol{y}^{k-1},y^k\right)\right)\right] \tag{8}$$
$$= \sum_{y=0}^{1}\left[H\left(p\left(a,b,c|\boldsymbol{y}^{k-1},y^k=y\right)\right)\right]\left[P\left(y^k=y|\boldsymbol{y}^{k-1}\right)\right]$$
$$= \sum_{y=0}^{1}\left[\int p\left(a,b,c|\boldsymbol{y}^{k-1},y^k=y\right)\log\left(\frac{1}{p\left(a,b,c|\boldsymbol{y}^{k-1},y^k=y\right)}\right)d(a,b,c)\right]$$
$$\times\left[\int P\left(y^k=y|a,b,c\right)p\left(a,b,c|\boldsymbol{y}^{k-1}\right)d\left(a,b,c\right)\right]$$

Hence, the next stimulus to be selected is the one with orientation that can minimize Eq. 8 (i.e., $E_k[H(p(a, b, c|\boldsymbol{y}^{k-1}, y^k))]$). In brief, the adaptive design intends to effectively stimulate the next trial that is most informative for the subject, and it helps avoid large regions of the stimulus space that are less likely to be useful to the experiment. The detailed steps and simplified strategies for the calculation of the ME method are provided in Appendix A.

Match probability (MP) method The MP method selects the next stimulus (i.e., the *k*th stimulus) that makes the probability of yes/no judgement (i.e., Eq. 1) closest to .5. More formally, this method selects the stimulus with orientation (x) that minimizes the following formula:

$$\left|\left\{\gamma+(1-\gamma-\lambda)\frac{1}{1+e^{-\beta(x-\alpha(t))}}\right\}-0.5\right| \tag{9}$$

where $\alpha(t)$ are evaluated at $\left(\hat{a},\hat{b},\hat{c}\right)$ in practice, and $\hat{a}$, $\hat{b}$, and $\hat{c}$ refer to the current estimates, which are obtained on the basis of the subject's responses on the previous k–1 stimuli. Thus, the performance of MP will be affected by the closeness of $\hat{\alpha}(t)$ to $\alpha(t)$, especially in the initial stage of the experiment. Because there is no response history yet when selecting the first stimulus, the MP method randomly selects the first stimulus from the set of stimuli with medium-sized orientations (e.g., $\boldsymbol{x}$' = 47:0.25:49 if the entire orientation space is $\boldsymbol{x}$ = 41:0.25:55)[1] as in the ME method.

Note that under the assumption of $\gamma = \lambda$, the MP method is equivalent to selecting the stimulus with the orientation (x) closest to $\alpha(t)$ (i.e., formula [9] reduces to $|x-\alpha(t)|$), which is very similar to the "match-*b*" (b stands for item difficulty) item selection criterion in CAT (Chang & Ying, 1999; Cheng, 2008). The MP method also bears close resemblance to the QUEST method in which the next trial is placed at the current most probable estimate of parameters (Watson & Pelli, 1983).

One-up, one-down staircase (UD) method We compare the Bayesian approach and MP method to a relatively simple adaptive method, the one-up, one-down staircase (UD), which proceeds with two steps: (1) in the first block, T stimuli are sequentially selected for each subject via the simple random sampling with replacement (i.e., random method) from the range of $\boldsymbol{x}$ = 41:0.25:55. (2) For the subsequent repeated blocks, the selection of the stimulus assigned to the current subject at time point t solely depends upon his/her response to the stimulus presented to him/her at the same time point within the previous block. That is, if subject i chooses "yes" at time point t within the jth block (i.e., $y_{i,\ j(t)}=1$), the stimulus assigned to him/her at time point t within the (j+1)th block will be more "difficult"[2] (e.g., $x_{i,\ j(t+1)}=x_{i,\ j(t)}-0.25$, if $x_{i,\ j(t)}>41$, and $x_{i,\ j(t+1)}=x_{i,\ j(t)}$ otherwise); otherwise, the stimulus will be "easier" (e.g., $x_{i,\ j(t+1)}=x_{i,\ j(t)}+0.25$ if $x_{i,\ j(t)}<55$, and $x_{i,\ j(t+1)}=x_{i,\ j(t)}$ otherwise).

Please note that the adaptation scheme in the UD method only depends on the subject's response to the stimulus presented at exactly the same time point in a previous block, thus it can be considered as a local adaptive scheme. In contrast, the ME and MP methods are global adaptive approaches because they select the optimal stimulus from the entire stimulus space each time. Therefore, the ME and MP methods are theoretically better than the UD method. For the two global adaptive methods, despite of the simplicity of MP, ME has many advantages, for instance: (1) ME is more general than MP because it is not only suitable for dichotomous responses but also suitable for polytomous responses; (2) unlike MP, ME does not rely on the interim estimates to select the next stimulus, so it is not directly affected by interim estimation precision; and (3) ME can make full use of prior information [i.e., $\pi(a, b, c)$] if informative priors are available. The performance of the above methods will be carefully evaluated through simulation studies.

The random method (denoted as RM) will be used as a baseline reference method. Random selection is performed by simple random sampling with replacement of stimulus from the range of $\boldsymbol{x}$ = 41:0.25:55 throughout the experiment. It is germane to note that the first block of stimuli from the UD method are indeed selected from the random method.

Simulation studies

Simulation Study 1

The primary objective of this study is to compare the three adaptive stimulus selection methods (i.e., ME, MP and UD methods) in the context of the time-course of perceptual adaptation under different conditions. The random method served as a baseline for comparison throughout the simulations. Similar to the fixed-length termination rule in CAT, we fixed the total number of trials to be 200 (i.e., Z = 200) here. Then the two important factors, the number of trials within each block (T) and the number of repeated blocks (J),

[1] The expression $\boldsymbol{x} = x_1{:}\ x_2{:}\ x_3$ throughout this article refers to an array starting at x_1, with a step size of x_2 and a final value of x_3.

[2] Note that smaller x value means the stimulus is more difficult because it implies smaller value of probability $P(Y=1)$.

were manipulated to form two experimental conditions: (1) 40 trials per block and 5 blocks (i.e., $T = 40$ and $J = 5$); (2) 20 trials per block and 10 blocks (i.e., $T = 20$ and $J = 10$). Moreover, when comparing the same stimulus selection methods under both conditions, the simulation details and procedures were identical except for the manipulated factors to ensure fairest comparison. The entire simulation process was implemented using 64-bit MATLAB 2013a (The Mathworks, Inc., 2013), and all codes ran on a ThinkPad laptop equipped with i7-7500U CPU (2.70GHz and 2.90 GHz duo processors), 16 GB RAM and 64-bit operating system. Note that the source codes for this study are available at "https://sites.uw.edu/pmetrics/publications-and-source-code" for interested readers' reference.

Generation of stimulus pool and subjects

The orientation of stimulus was varied from 41 to 55 with a step size of 0.25 (i.e., $\boldsymbol{x}$ = 41:0.25:55), resulting in 57 stimuli in total ($R = 57$). For each subject, the same stimuli can be re-selected an unlimited number of times during the experiment. The subjects' parameters $\boldsymbol{\theta} = (a, b, c)$ were sampled from the entire space, such that $\boldsymbol{a}$ = 46:0.5:50, $\boldsymbol{b}$ = 1:0.25:3, and $\boldsymbol{c}$ = 0.01:0.01:0.1, thereby resulting in 810 (9×9×10) possible combinations of (a, b, c). For each possible combination of (a, b, c), only one subject was simulated for simplicity. Accordingly, a total of 810 subjects ($N = 810$) were generated in this study.

Experimental procedures

For each condition, the experimental procedures proceed with four main steps. First, a stimulus was selected from the stimulus pool using the method's stimulus selection approach. Second, the response of the current subject to the selected stimulus was simulated by directly comparing the probability in Eq. 1 with a random number between 0 and 1. To make a strictly fair comparison among the stimulus selection methods, we pre-generated a response matrix of size N-by-Z (810-by-200 here) and extracted the corresponding response for the specific subject and trial when needed. Thirdly, the EAP method was employed to update estimates of the person parameters $\boldsymbol{\theta} = (a, b, c)$ sequentially based on his/her response history. Finally, the experiment was ended when the pre-fixed number of trials had been reached.

Remarks

Because the 810 combinations of (a, b, c) mentioned above cover the entire regions of the three-dimensional space, they were also considered as the M ($M = 810$) data points $(a^{(m)}, b^{(m)}, c^{(m)})$ $(m = 1, 2, \ldots, M)$ sampled from prior distribution $\pi(a, b, c)$, which would then be used for implementing EAP estimation and ME selection method. Additionally, we adopted uniform prior[3] for $\pi(a^{(m)}, b^{(m)}, c^{(m)})$ in Eq. (A2) (see Appendix A). As we alluded to above, β, γ and λ in Eq. 1 were fixed throughout the experiment, that is, $\beta = 2$ and $\gamma = \lambda = 0.05$.

Evaluation criteria

For each condition, three types of criteria were used to evaluate the performance of each stimulus selection method, which, respectively reflect (1) person parameter recovery, (2) stimulus pool usage, and (3) computation efficiency.

Person parameter recovery The evaluation indicators in this category include bias, relative bias, mean squared error (*MSE*), mean standard error of estimate (Mean_*SE*), and mean area formed by the two curves of $\alpha(t)$ (Mean_Area). With respect to the last criterion, one curve is characterized by the estimated value $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}, \hat{c})$ and the other by the true value $\boldsymbol{\theta} = (a, b, c)$ in a two-dimensional space with the time of tth *trial* ($val(t)$) on the x-axis and $\alpha(t)$ on the y-axis. They are computed as follows:

$$Bias_d = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{id} - \theta_{id}\right)$$

$$Relative\_Bias_d = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\hat{\theta}_{id} - \theta_{id}\right)}{\theta_{id}}$$

$$MSE_d = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{id} - \theta_{id}\right)^2$$

$$Mean\_SE_d = \frac{1}{N}\sum_{i=1}^{N} SE_{\hat{\theta}_{id}}$$

$$Mean\_Area = \frac{1}{N}\sum_{i=1}^{N}\left\{\int_{t_l}^{t_u}\left|\left(\hat{a}_i - \hat{b}_i e^{-\hat{c}_i t}\right) - \left(a_i - b_i e^{-c_i t}\right)\right| dt\right\}$$

where $\hat{\theta}_{id}$ and θ_{id} represent the final estimated and true values of the ith subject on the dth dimension, respectively. $SE_{\hat{\theta}_{id}}$ indicates the standard error for estimate $\hat{\theta}_{id}$. $t_l = \min(val(t))$ and $t_u = \max(\text{val}(t))$ $(t = 1, 2, \ldots, T)$[4].

As for the Mean_Area index, it is the area between the true and estimated curves. Figure 2 provides an illustration of the index under two scenarios: (1) The two curves do not intersect, and (2) the two curves have only one intersection.

Moreover, a trend line that depicts the change in the mean error of $\alpha(t)$ over the time points was also plotted, where the mean error of $\alpha(t)$ (i.e., Mean_Error$_t$) is given by

[3] This prior is uniform over the range of values selected for a, b, and c in the experiment; that is, $a \sim U(46, 50)$, $b \sim U(1, 3)$, and $c \sim U(0.01, 0.1)$.

[4] $\hat{\theta}_{id}$ here represents the final parameter estimate after all trials, rather than the trial-by-trial parameter estimate.

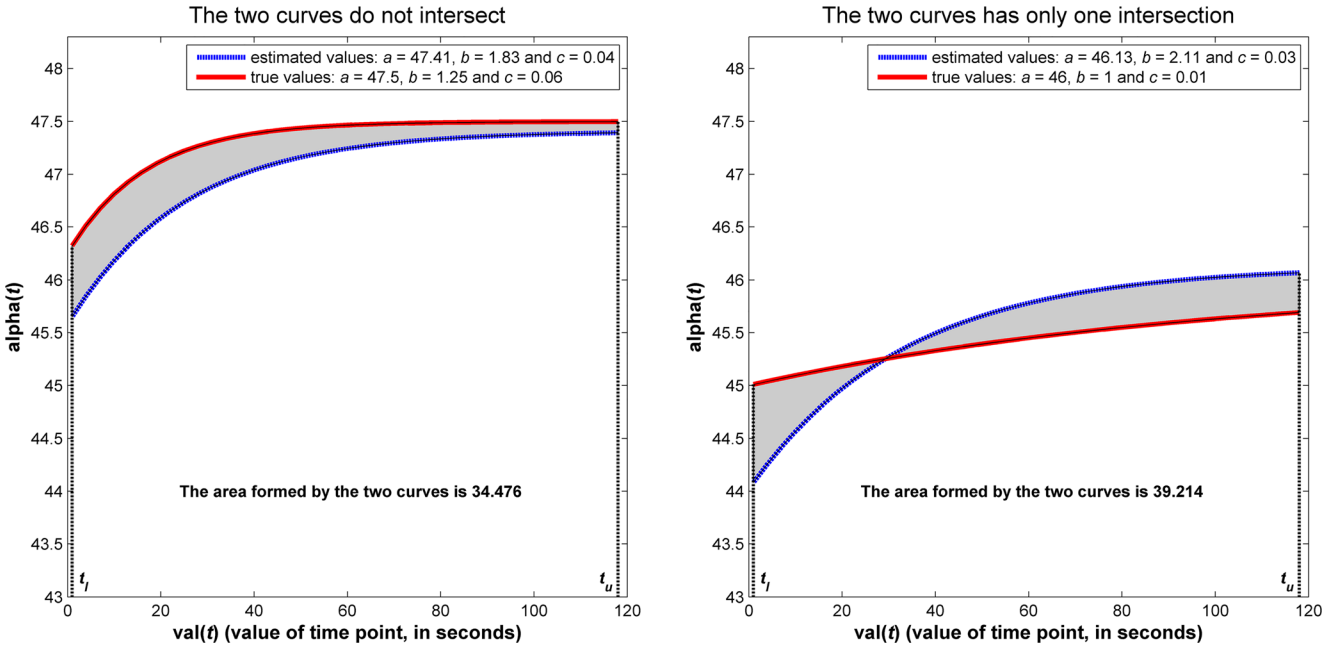

Fig. 2 Graphical illustration of the calculation of the Mean_Area index (the shaded areas are computed).

$$Mean\_Error_t = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_i(t) - \alpha_i(t)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{a}_i - \hat{b}_i e^{-\hat{c}_i t}\right) - \left(a_i - b_i e^{-c_i t}\right)\right)$$

Stimulus pool usage The number of times each stimulus gets selected (i.e., stimulus frequency) was used to reflect the uniformity of stimulus pool usage resulted from different stimulus selection methods.

Computation efficiency It is evaluated by computing the average time for selecting each stimulus (Mean_Time)

$$Mean\_Time = \frac{1}{N}\frac{1}{J}\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} Time\_Used_{ijt},$$

where Time_Used_{ijt} denotes the time used for selecting the stimulus at time point t within the jth block for the ith subject, and it is measured in seconds. For all evaluation indicators, values closer to zero indicate better the performance of the stimulus selection method.

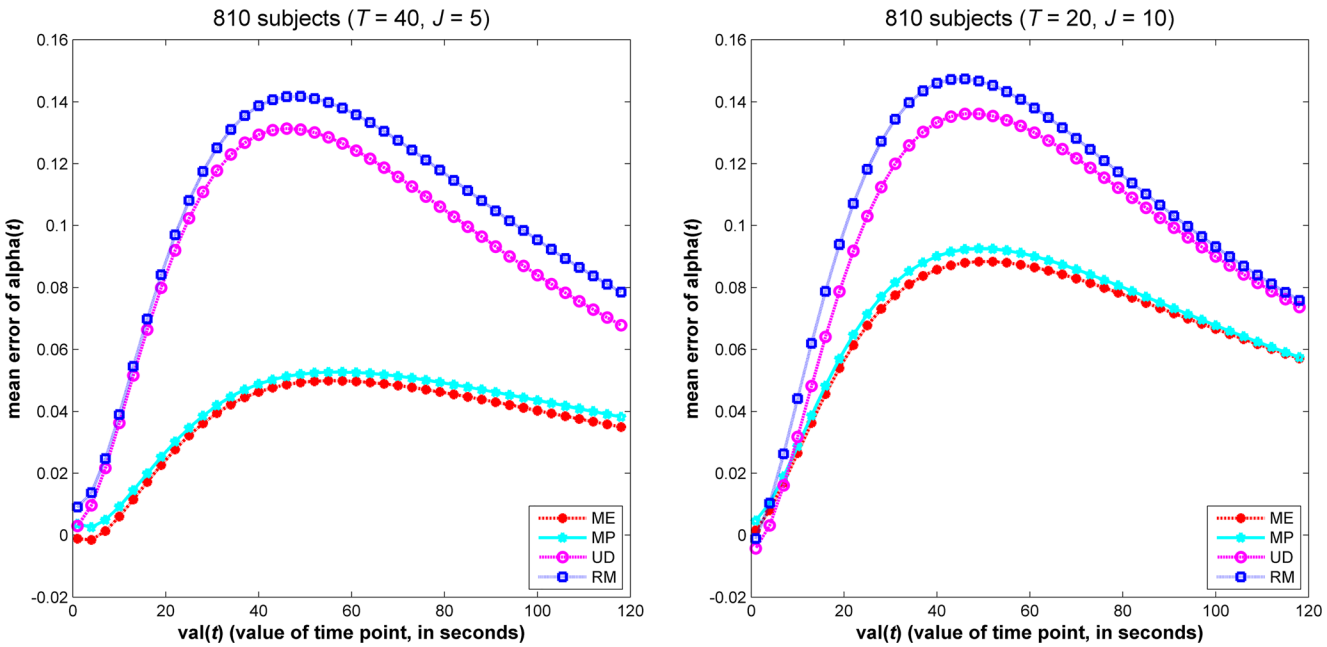

Fig. 3 Mean error of α(t) averaged across all subjects.

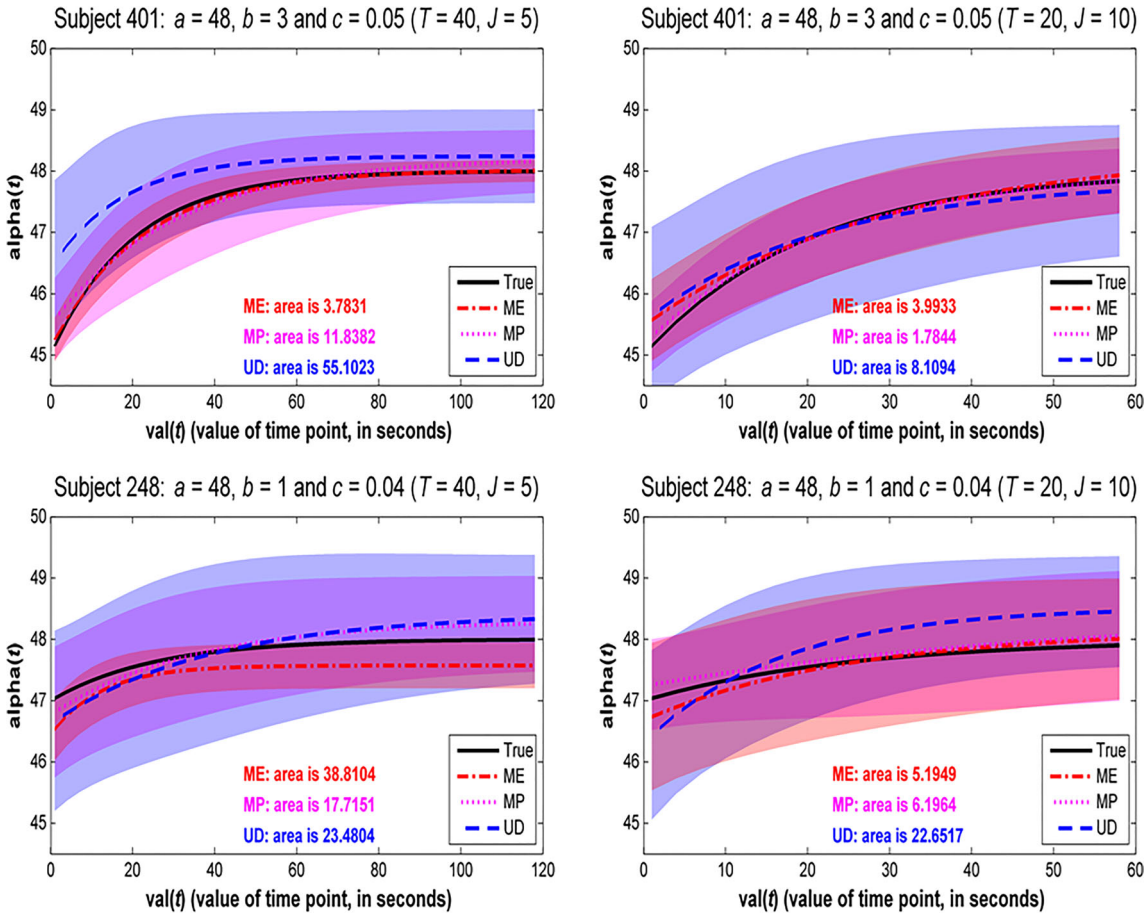

Fig. 4 Estimates and 95% confidence intervals of $\hat{\alpha}(t)$ for subjects #401 and #248.

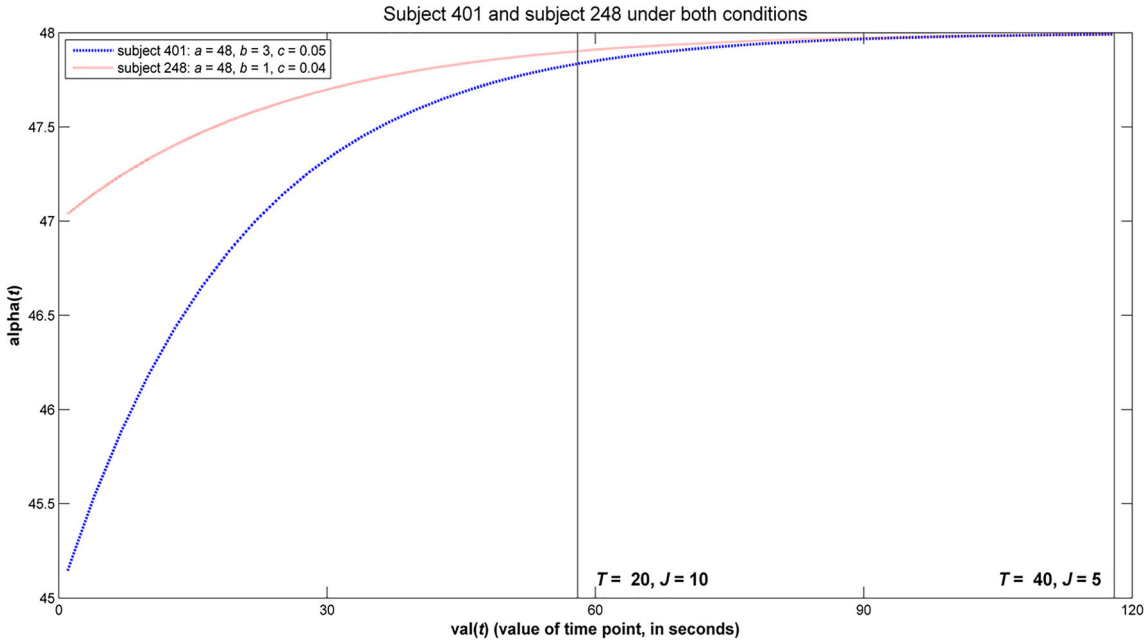

Fig. 5 True values of $\alpha(t)$ for subject #401 and subject #248 under two conditions.

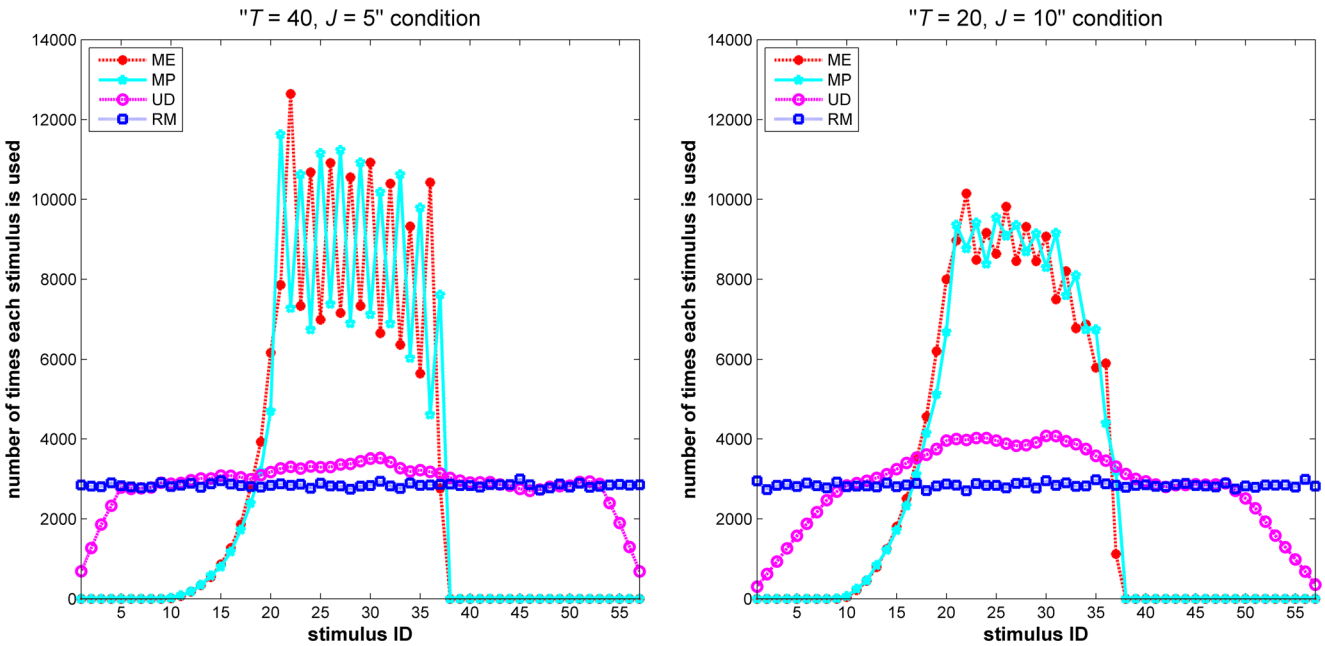

Fig. 6 Numbers of times each stimulus was selected.

Results

Results regarding the parameter recovery

Table 1 provides the Bias, Relative Bias, *MSE*, Mean_*SE*, and Mean_Area results for different stimulus selection methods under different conditions. Note that the first four indicators quantify the recovery of each parameter separately, whereas the last one is indicator of overall recovery. The following points can be observed: (1) The bias values obtained by all methods under both conditions were very close to 0, and the relative bias values of the *b* and *c* parameters were more

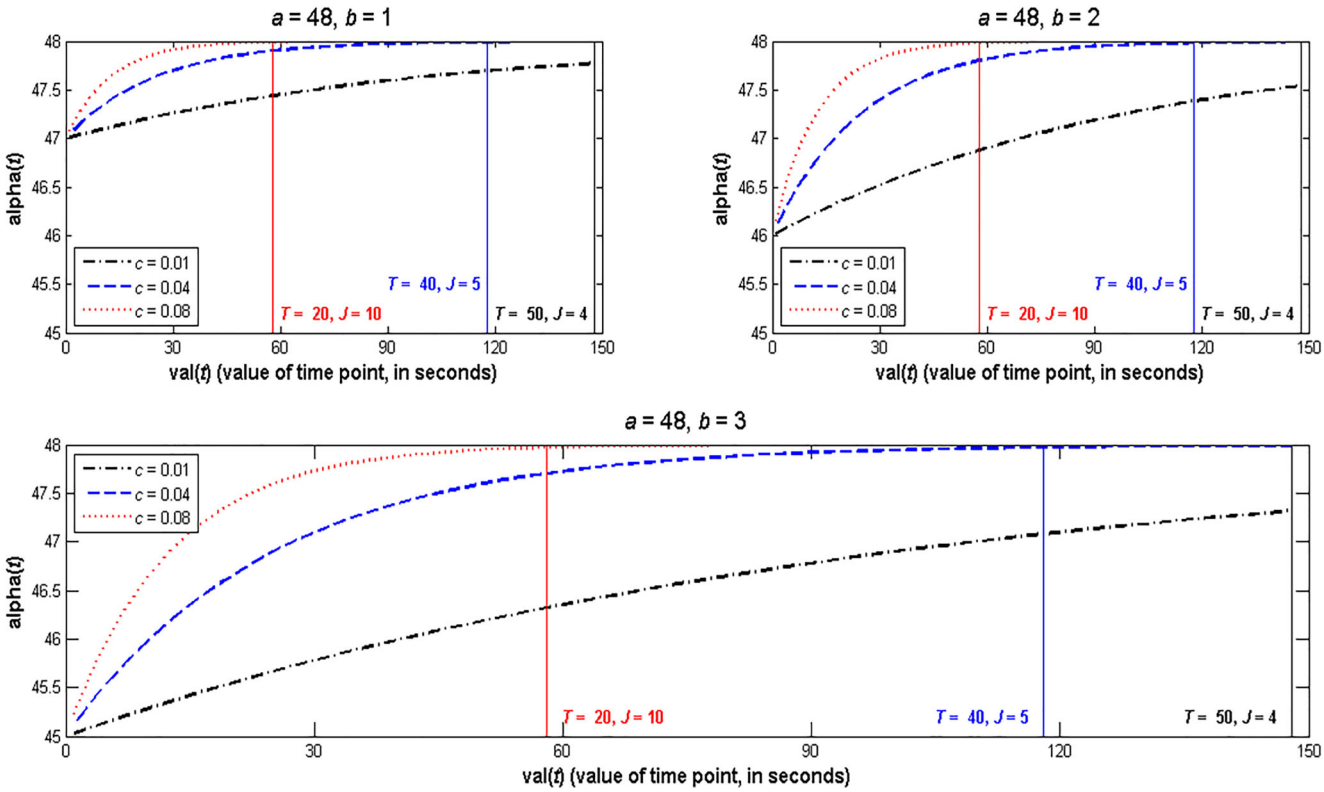

Fig. 7 Graphical illustration of the experimental design for Study 2.

deviated from 0 than those of the *a* parameter. (2) In terms of all remaining indicators, the two global adaptive methods (i.e., ME and MP) consistently performed the best as expected, and the UD method performed slightly better than the random method. (3) As to the two global adaptive methods, ME worked slightly better than MP under the longer block condition, whereas the performance of the two methods are almost indistinguishable under the shorter block condition. This is mainly because the stimuli selected by the two methods under the latter condition were very similar, resulting in a .986 correlation of stimulus frequency between the two methods. Under the former condition, due to the relatively larger bias between $\hat{\alpha}(t)$ to $\alpha(t)$, the stimuli selected by MP tended to be slightly different from those selected by ME, resulting in a .845 correlation of stimulus frequency between the two methods. (4) All the Mean_Area values resulted from the longer block condition were smaller, implying that using fewer long blocks produced better overall recovery than using more short blocks. (5) Bias, *MSE*, and Mean_*SE* are all scale-sensitive, and therefore they cannot be compared directly across the three parameters.

Furthermore, in addition to evaluating the recovery per parameter, we are also interested in the recovery of $\alpha(t)$. Figure 3, which plots the mean error of $\alpha(t)$ averaged across all subjects as a function of time point, provides additional evidence that ME was superior to the other three methods from a whole perspective.

Figure 4 shows the estimates and 95% confidence intervals of $\hat{\alpha}(t)$ of two representative subjects (i.e., #401 and #248) for different methods under both conditions. Note that the results for the random method are not presented in this figure, due to space limits. As is shown, even for the same subject, the three methods performed differently under the two manipulated conditions. In particular, for subject #401, the ME method had satisfactory confidence band and smallest area value when there are five blocks of 40 stimuli each, whereas it had larger area value than the MP method and a relatively wider confidence band in the second condition. Moreover, the three methods showed different performance for different subjects under the same condition. For example, in the first condition, the ME method worked best for subject #401, but performed the worst for the other subject. Overall, the ME method generated the narrowest confidence band, and in most cases, the true curve of $\alpha(t)$ lay within the band. However, some level of variabilities indeed emerges from Fig. 4.

One theme cutting across Fig. 4 is that the confidence band gets narrower as val(t) increases. This is because $\mathrm{var}(\hat{\alpha}(t)) = w_1\mathrm{var}(\hat{a}) + w_2\mathrm{var}(\hat{b}) + w_3\mathrm{var}(\hat{c})$, where $w_1 = 1$, $w_2 = \left(e^{-\hat{c}t}\right)^2$ and $w_3 = \left(t\hat{b}e^{-\hat{c}t}\right)^2$, and as val($t$) goes to infinity, both the weights for $\mathrm{var}(\hat{b})$ and $\mathrm{var}(\hat{c})$ (i.e., w_2 and w_3) approach zero. Moreover, the variance of $\hat{c}$ (i.e., $\mathrm{var}(\hat{c})$) was small in the current situation, $\mathrm{var}(\hat{a})$ would predominantly determine the total variance size [$\mathrm{var}(\hat{\alpha}(t))$] when the val(t) reached 120 or 60. According to this reasoning, the increasingly narrower confidence band in the upper-left subgraph of Fig. 4 indicates that the $\mathrm{var}(\hat{a})$ of subject #401 from the ME method is relatively small. In fact, the ME estimates (standard errors) of $\hat{a}$, $\hat{b}$, and $\hat{c}$ for subject #401 in the "$T = 40$, $J = 5$" condition were 48.017 (**0.091**), 2.892 (0.164), and 0.045 (0.007), and their counterparts in the "$T = 20$, $J = 10$" condition were 48.318 (**0.261**), 2.842 (0.226), and 0.034 (0.008), respectively.

Figure 5 helps us further understand the impact of different experiment conditions on the recovery of $\alpha(t)$. Remember that $a = 48$ denotes the asymptote of $\alpha(t)$, and the dash-dotted line representing subject #401 approaches the asymptote when $T = 40$. By comparison, the dash-dotted line has certain distance from the asymptote when $T = 20$, thereby the *a* parameter may not be accurately recovered in spite of more repeated blocks. This phenomenon is also true for most of subjects with other combinations of (a, b, c) (including subject #248),[5] which is the reason why all *MSE* and Mean_*SE* results on *a* parameter from the "$T = 40$, $J = 5$" condition were smaller than those from the "$T = 20$, $J = 10$" condition (see Table 1).

Results regarding the stimulus pool usage

Figure 6 presents the numbers of times each stimulus was used. As is shown, the ME and MP methods produced very similar results (i.e., the most unbalanced stimulus pool usage), the random method generated the most balanced use, and the UD method was in-between. In addition, it was found that 29/28 stimuli (about half of the total number of stimuli) in the stimulus pool were never used by the ME/MP method under both conditions, whereas all of the stimuli were used in the other two methods. Taking the ME method as an example, the 29 unused stimuli were all distributed at both ends; they were Stimuli 1 through 9 ($\boldsymbol{x}$ = 41:0.25:43) and Stimuli 38 through 57 ($\boldsymbol{x}$ = 50.25:0.25:55); however, Stimuli 20 through 36 with medium-sized orientations ($\boldsymbol{x}$ = 45.75:0.25:49.75) for the "$T = 40$, $J = 5$" condition, and Stimuli 19 through 36 ($\boldsymbol{x}$ = 45.50:0.25:49.75) for the "$T = 20$, $J = 10$" condition were used more than 5,000 times.

Another interesting observation from Fig. 6 is that both the ME and MP methods showed a jagged up–down pattern, in particular under the "$T = 40$, $J = 5$" condition. Given the similar performance of the ME and MP methods, the explanation for stimulus usage pattern is for the MP method as an example. As we alluded to earlier, the MP method boils down to selecting the stimulus with the

[5] The ME estimates (standard errors) of $\hat{a}$, $\hat{b}$ and $\hat{c}$ in the "$T = 40$, $J = 5$" condition were 47.575 (**0.189**), 1.127 (0.194) and 0.081 (0.025), and their counterparts in the "$T = 20$, $J = 10$" condition were 48.145 (**0.461**), 1.465 (0.420) and 0.041 (0.024).

orientation (x) closest to $\alpha(t)$. If using the true values of (a, b, c) to compute $\alpha(t)$ and find the closest x, then across all 810 simulees, the frequencies for stimuli 20 through 35 are 5,662, 12,248, 5,662, 12,248, 5,662, 12,339, 5,759, 12,298, 5,686, 12,194, 5,512, 11,921, 5,102, 11,410, 4,255, and 10,290, respectively, which matches closely to the pattern shown in Fig. 6. Although $\hat{\alpha}(t)$ s instead of $\alpha(t)$s are actually used in the MP method, the same stimulus usage pattern emerges. In fact, the stimulus usage pattern depends on the simulee population distribution (we used a uniform distribution in the simulation study), the block design, and the step size of stimulus' orientation.

Results regarding the computation efficiency

As to all stimulus selection methods, the average times for selecting each stimulus (i.e., Mean_Time) were acceptable for instantaneous stimulus selection. For example, the more computationally intensive method (i.e., ME) took only 0.003s and 0.004s under "$T = 40$, $J = 5$" condition and "$T = 20$, $J = 10$" condition, respectively.

Simulation Study 2

In Study 1, four stimulus selection methods (i.e., ME, MP, UD and RM) were compared on the basis of the entire group of subjects (usually termed an unconditional study in adaptive design research) under two experimental conditions, corresponding to study designs. However, the two designs may not be optimal for every individual. For instance, individuals with lower c parameters (e.g., 0.01 and 0.02) may need even longer blocks. Thus, the primary objective of these simulations was two-fold: First, the performance of the three methods (i.e., ME, UD, and RM) were evaluated conditioning on a set of unique combinations of the parameters (a, b, c) that may be sensitive to the block design. Because the MP method

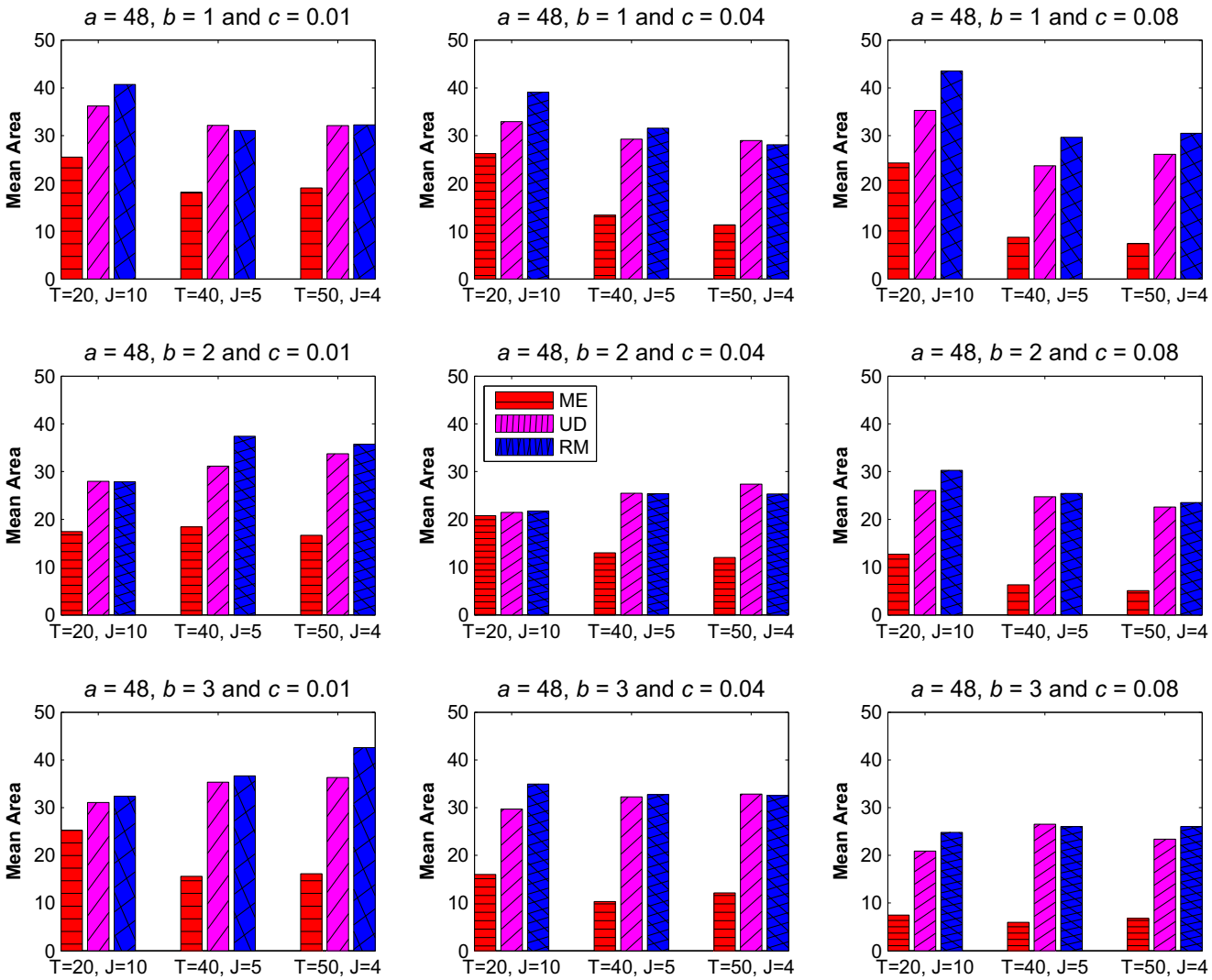

Fig. 8 Mean_Area results of different stimulus selection methods under different conditions for all representative points.

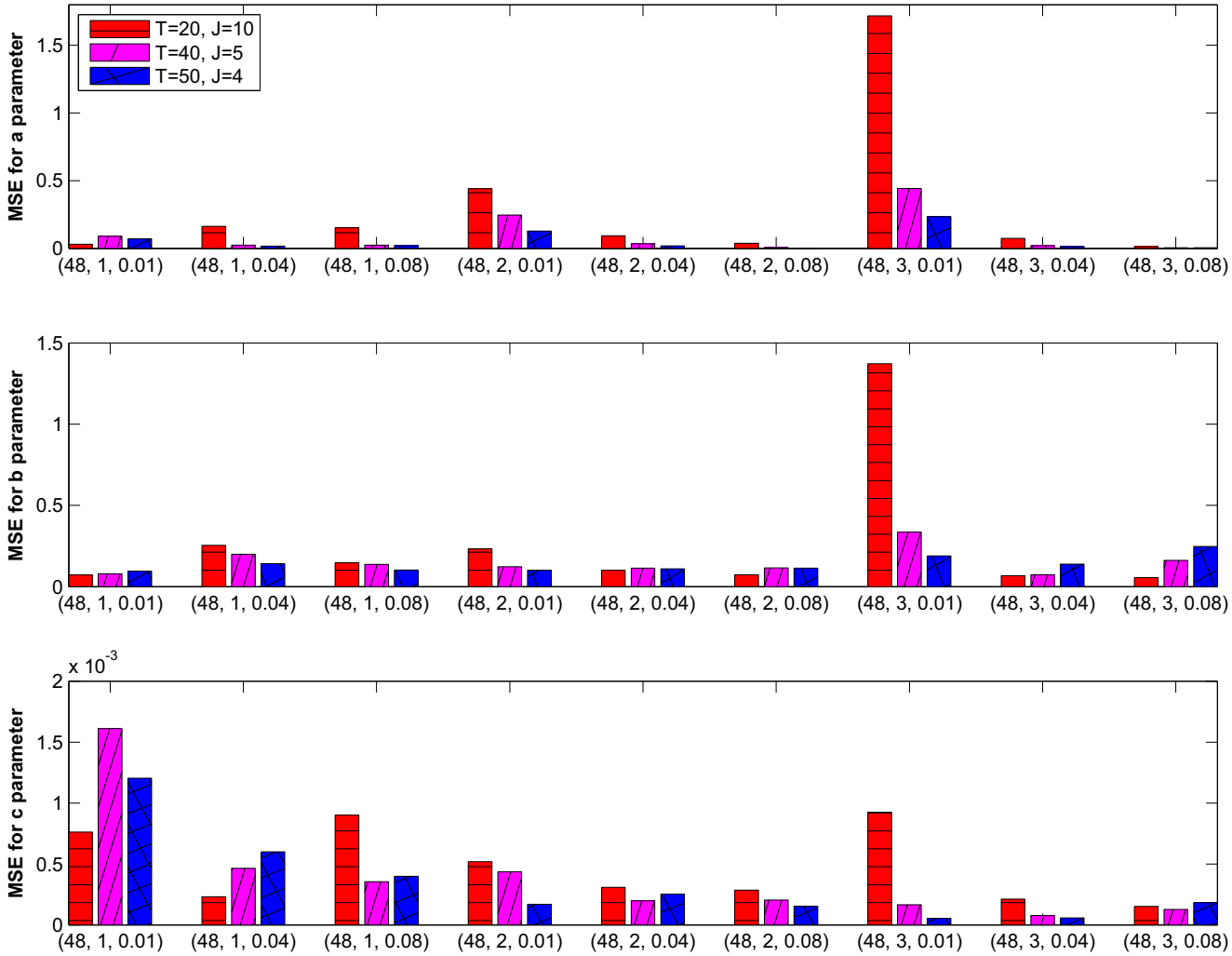

Fig. 9 Conditional *MSE* results of the ME method for all representative points under different conditions.

performed slightly worse than the ME method from simulation study, it is no longer considered in Study 2. This makes the work a conditional study in the terminology of adaptive design research. Second, more experimental conditions with varying combinations of number of trials within each block (T) and number of blocks (J) were considered to explore what the optimal block length was. Specifically, this study differs from Study 1 in the following aspects:

1. *Generation of subjects*. Nine representative combinations (or points) were obtained first by fixing a parameter at a medium level (i.e., $a = 48$) and, respectively varying b and c parameters at three levels of low, medium, and high (i.e., $b = 1$, 2, and 3; $c = 0.01$, 0.04, and 0.08). Then, a sample of 900 subjects was generated, with 100 subjects at each particular point. Note that generating 100 subjects at the same point could be considered as 100 replications at the point.
2. *Experimental conditions.* In addition to the two experimental conditions discussed earlier, another condition of 50 trials per block and four blocks ($T = 50$, $J = 4$) was added to this study. The experimental design of this study is graphically illustrated in Fig. 7.
3. *Evaluation criteria.* The evaluation criteria presented in Study 1 were adapted to this conditional study by calculating them at each of the nine particular points.

Figure 8 shows the Mean_Area results of the different methods under three conditions at nine particular points. The ME method was again consistently superior to the other two methods. Specifically, the ME method produced smaller estimation errors than the other two methods in all scenarios. On the other hand, by comparing the results of all methods at all points across different conditions (i.e., conditions "$T = 20$, $J = 10$" through "$T = 50$, $J = 4$"), one can notice that the "$T = 50$, $J = 4$" condition indeed improved the overall recovery for some particular points and/or methods in terms of Mean_Area. For instance, for the ME method, increasing the number of trials to 50 within each block worked best among the three conditions

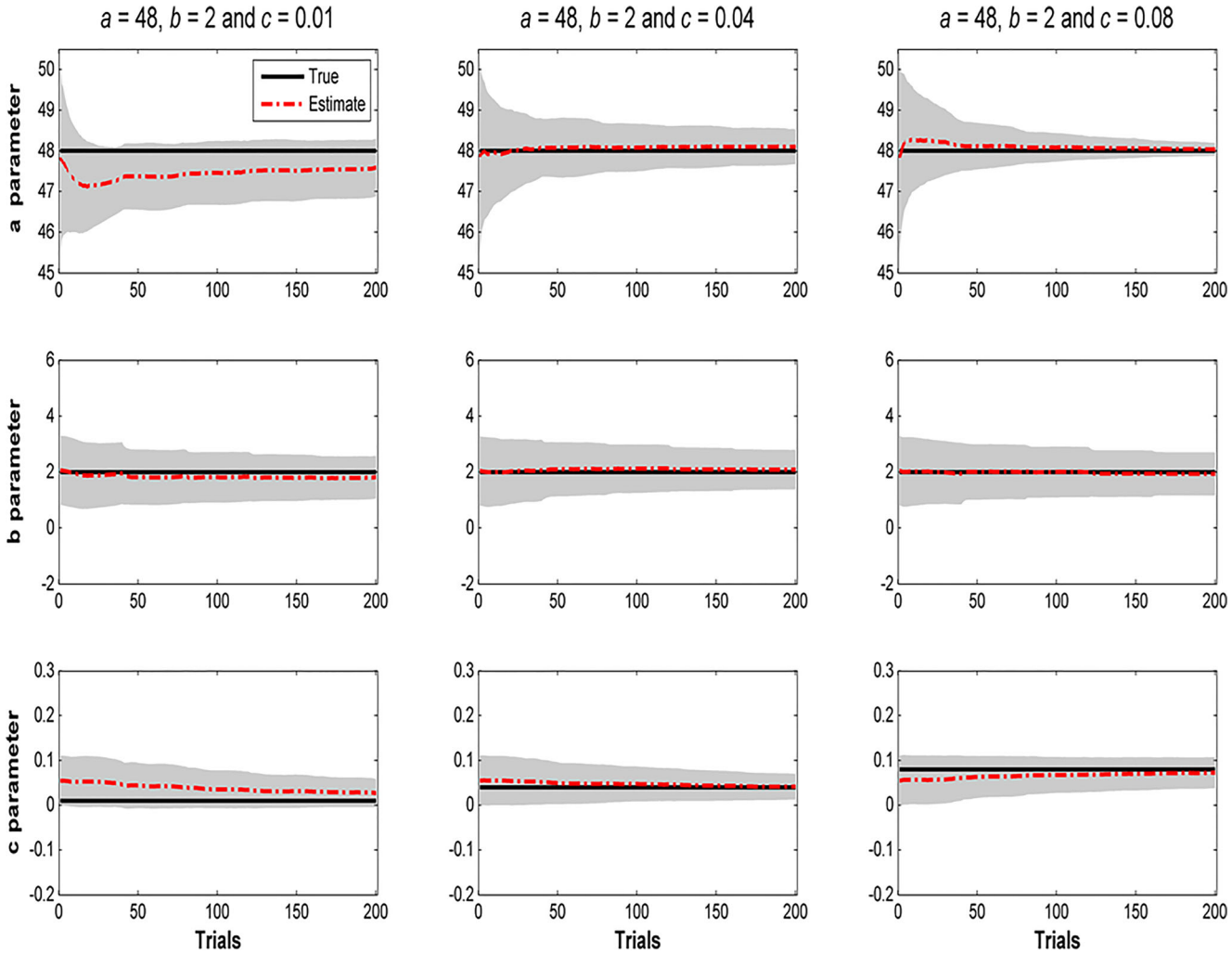

Fig. 10 Trial-by-trial parameter estimates and their 95% confidence intervals, averaged across 100 replications for three selected subjects under the "$T = 40$, $J = 5$" condition.

for points 2 through 6;[6] and for points (48, 2, 0.08) and (48, 1, 0.04), all methods had the best performance in the "$T = 50$, $J = 4$" condition. To sum up, on the basis of the results of all nine particular points, it is strongly recommended to use "$T = 40$, $J = 5$" or "$T = 50$, $J = 4$" as the experimental condition if the total number of trials was fixed at 200.

Figure 9 summarizes the conditional *MSE* values of the best-performing method (i.e., the ME method) for the nine hypothetical subjects under three conditions. Several interesting observations from the figure merit illustration: (1) Except for the first subject (48, 1, 0.01), the *MSE* values of the *a* parameter for the other eight subjects monotonically decreased as the number of trials within each block (T) increased from 20 to 50, and the degree of such decrement became smaller when the *c* parameter increased, regardless of the value of the *b* parameter. In other words, for medium- and large-sized *b* parameters (b = 2 and 3) combined with a small *c* parameter ($c = 0.01$), increasing the value of T had the greatest benefit in improving the estimation accuracy of *a* parameter. This result could be explained in conjunction with Fig. 7: First, longer blocks implied that the $\alpha(t)$ values grow to their asymptotes; hence, the *a* parameter could be recovered more accurately. Second, the three lines in each of the subgraphs in Fig. 7 indicated that the larger the *c* parameter, the smaller the difference in the distance of $\alpha(t)$ from the asymptote across the three conditions. (2) As compared to the *a* parameter, with a wider range of values, the *b* parameter possessed larger *MSE* values for subjects with medium- and large-sized *c* parameters ($c = 0.04$ and 0.08), especially under the latter two conditions. This indicated that the *a* parameter was easier to recover than the *b* parameter under these conditions. (3) The *MSE* values of the *c* parameter for all subjects and conditions were very small, mainly because the *c* parameter itself has relatively small values; thus, the differences among them could be ignored.

[6] The five points refer to (48, 1, 0.04), (48, 1, 0.08), (48, 2, 0.01), (48, 2, 0.04), and (48, 2, 0.08).

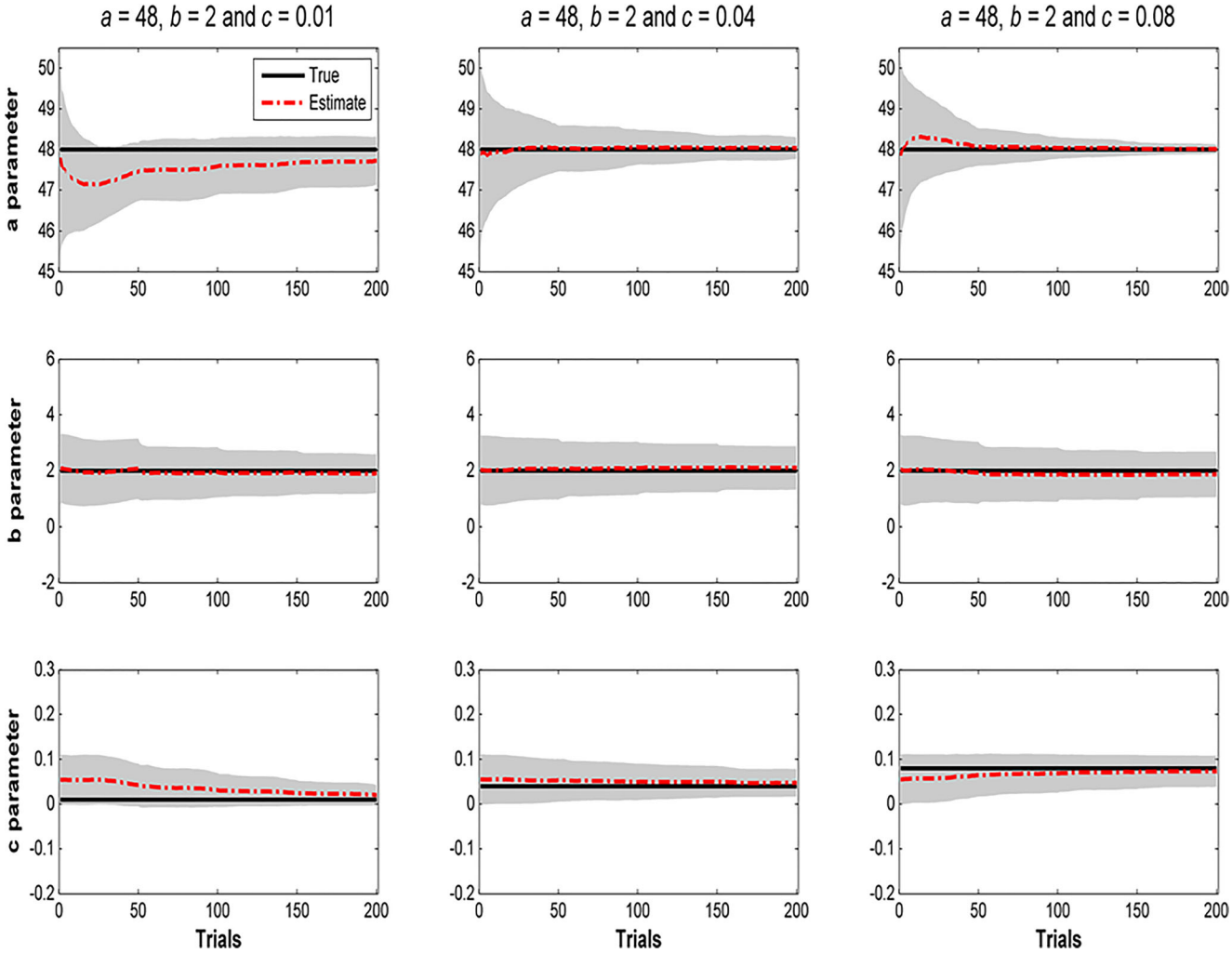

Fig. 11 Trial-by-trial parameter estimates and their 95% confidence intervals, averaged across 100 replications for three selected subjects under the "T = 50, J = 4" condition.

Following the study of Zhao, Lesmes, and Lu (2019), we also plotted the trial-by-trial parameter estimates of a, b, and c parameters and their 95% confidence intervals for three selected subjects with medium-sized a and b parameters [i.e., subject (48, 2, 0.01), subject (48, 2, 0.04), and subject (48, 2, 0.08)] in Figs. 10 and 11. These two figures depict only the results for the ME method under the "T = 40, J = 5" condition and the "T = 50, J = 4" condition, respectively. As expected, the parameter estimates (dash–dotted lines) were closer to the true parameter values (solid lines) as the number of trials increased. Moreover, the decreasing width of the confidence band (grey shaded areas) implied increased estimation accuracy of person parameters as more data came in.

Relating to Fig. 10, we also plotted the trial-by-trial parameter estimates and their 95% confidence intervals for a single replication in Fig. 12. As expected, the confidence intervals became narrower when more trials were added, albeit with slightly more fluctuation than in Fig. 10. Moreover, the interim point estimate of each parameter moved closer to its true value when more trials were added, with only a few exceptions, such as the a parameter for the first selected subject and the b parameter for all three selected subjects. However, the slight departures from the true values toward later stages of the experiment were so small that they are unlikely to affect the recovery of $\alpha(t)$. The observation also implies that a variable-length approach may be preferred, to stop the experiment when the estimation precision is adequate. Hence, not only will the experiment be more efficient, but also the experiment will be terminated before the estimation worsens.

In addition, because 100 replications were conducted at each true point of (a, b, c), the standard deviation (SD) of the 100 estimates of each parameter was also calculated for all points and conditions. The results show that the ME method had smaller SD values of a and b parameters than the other two methods at most of the points under all conditions (the difference in the SD values on c parameter among different methods was negligible), indicating that the ME method was

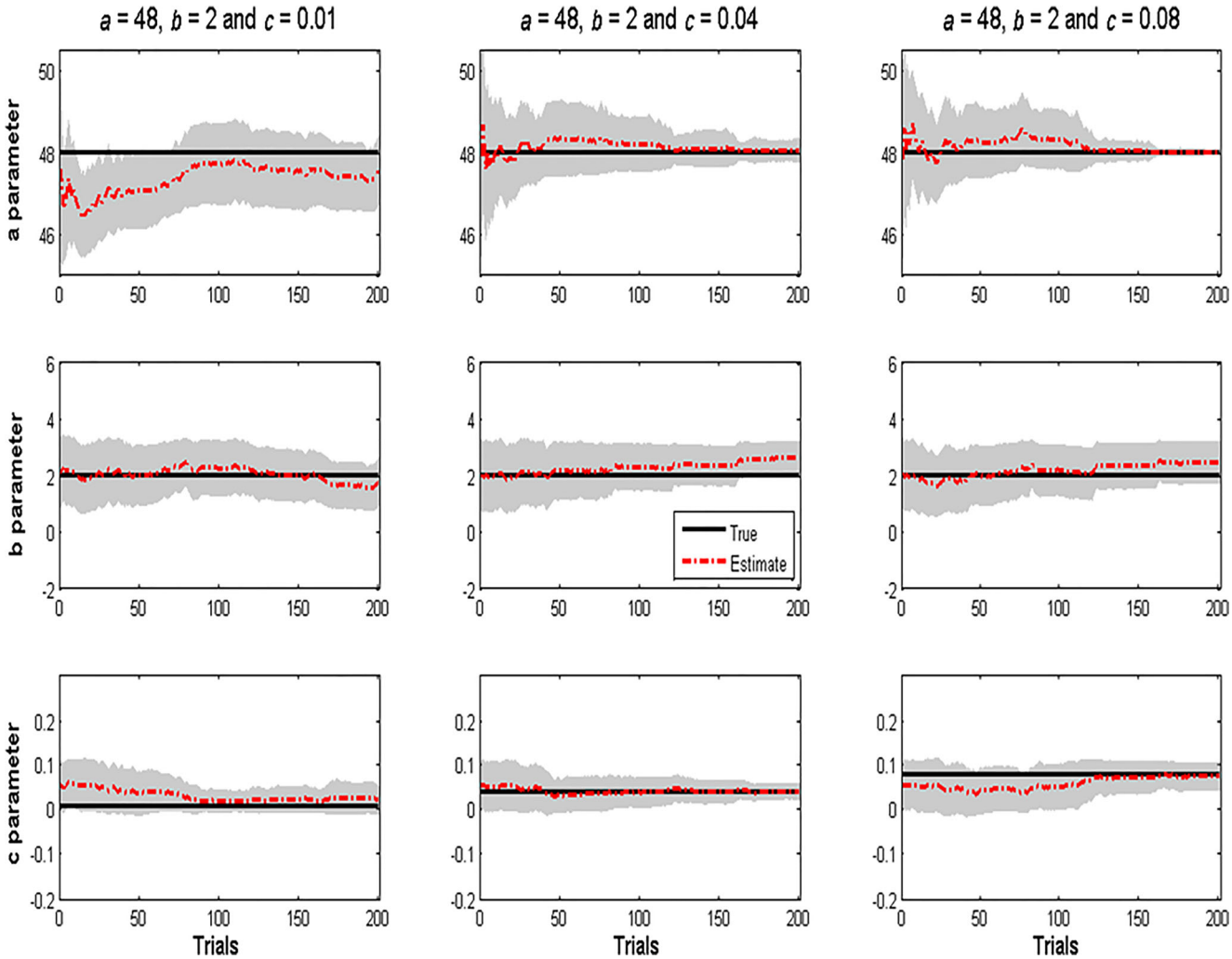

Fig. 12 Trial-by-trial parameter estimates and their 95% confidence intervals of a single replication for three selected subjects under the "$T = 40, J = 5$" condition.

more robust (against randomness introduced by probabilistic responses) than the other two methods as a whole. Due to limited space, the complete *SD* results are not presented here, but they are available from the authors upon request.

Discussion

Our main aim was to evaluate two multivariate adaptive methods (i.e., the ME and MP methods) for efficiently estimating the time-course of perceptual adaptation and to thoroughly evaluate their performance through simulation studies. Detailed calculation steps and simplification strategies for the ME method (see Appendix A for details) are also provided for interested researchers. In simulation studies, the ME method was compared with the MP method and two other methods (i.e., UD and random methods) under varying experimental conditions (or block designs).

The two studies showed that the ME method not only as a whole performs the best in terms of overall recovery (see Table 1), but also consistently works the best in terms of recovery at specific points after reducing the random errors[7] (see Fig. 8). Moreover, ME exhibits satisfactory computation efficiency in selecting the stimulus, making it an overall superior stimulus selection method for recovering the time-course function. These results agree with prior work of Zhao, Lesmes and Lu (2019) who first showed the superiority of adaptive methods assuming a functional form of the time course of perceptual change. We extended their results to a new domain, simulations of contrast adaptation, and compared the new methods to both traditional staircases and random stimulus selection.

[7] For subject #248 ($a = 48$, $b = 1$, $c = 0.04$), the ME method had the largest Mean_Area among all methods under the "40 trials per block and 5 blocks" condition (see the lower-left subgraph of Fig. 4). However, after reducing the random errors by replicating 100 times at the same point, ME performed the best under the same condition (see the middle three bars of Fig. 8).

Table 1 Results of person parameter recovery for different stimulus selection methods under different conditions

Index	Method	$T = 40, J = 5$			$T = 20, J = 10$		
		a	*b*	*c*	*a*	*b*	*c*
Bias	ME	– 0.002	– 0.002	– 4.826e-5	0.004	0.004	– 2.467e-4
	MP	0.001	– 0.004	– 1.859e-5	0.004	0.001	– 3.794e-4
	UD	– 0.002	– 0.004	7.277e-4	0.004	0.009	– 5.999e-4
	RM	0.006	– 0.003	7.765e-5	0.005	0.008	3.455e-4
Relative_ Bias	ME	– 1.966e-5	0.032	0.177	1.411e-4	0.037	0.251
	MP	4.430e-5	0.032	0.199	1.389e-4	0.036	0.250
	UD	1.288e-5	0.087	0.466	1.567e-4	0.083	0.426
	RM	1.911e-4	0.089	0.468	1.778e-4	0.092	0.469
MSE	ME	**0.031**	0.107	**2.916e-4**	0.115	0.118	**3.665e-4**
	MP	0.037	**0.103**	3.326e-4	**0.106**	**0.113**	3.783e-4
	UD	0.111	0.288	6.101e-4	0.173	0.250	6.393e-4
	RM	0.129	0.291	6.335e-4	0.177	0.280	6.708e-4
Mean_*SE*	ME	**0.144**	**0.326**	**0.0156**	**0.280**	**0.326**	**0.0177**
	MP	0.150	0.329	0.0157	0.285	0.328	0.0180
	UD	0.317	0.535	0.0246	0.393	0.491	0.0242
	RM	0.340	0.543	0.0251	0.415	0.527	0.0251
Mean_Area	ME	**9.691**			14.917[a]		
	MP	10.096			**14.796[a]**		
	UD	25.595			25.729[a]		
	RM	26.819			28.617[a]		

The best results among the three methods are in **boldface**. ME = minimum entropy method, MP = match probability method (i.e., select next stimulus that makes P($Y = 1$) closest to 0.5), UD = one-up, one-down staircase method, RM = random method. *MSE* = mean squared error, Mean_*SE* = mean standard error of estimate, Mean_Area = mean area formed by two curves. [a] To make the results under both conditions comparable, the Mean_Area values for the "$T = 20, J = 10$" condition are calculated on the basis of interval [min(1:3:120), max(1:3:120)] (i.e., [1, 118]) as in the "$T = 40, J = 5$" condition.

In this research, we assumed there are 57 stimuli in the stimulus pool, which are generated by varying the orientation of stimulus from 41 to 55 with a step size of 0.25. If we decreased the step size from 0.25 to 0.125 (which would result in 113 stimuli in total), there would be some decline in the correlation of stimulus frequency between ME and MP methods, especially in the "$T = 40, J = 5$" condition. This is because when the stimulus becomes more fine-grained, the stimuli selected by the two methods will be more different. As to the three block designs discussed in Study 2, moreover, the block designs with relatively more trials per block and fewer blocks (i.e., 40 trials per block and 5 blocks, 50 trials per block and 4 blocks) produce better parameter recovery, and therefore they should be preferred when designing the experiment. In addition, among the three key parameters of the psychometric function considered, the asymptote parameter (i.e., *a* parameter) is the easiest to recover in general.

The present research could be extended in multiple directions. First of all, the continuous entropy method (i.e., ME) and the match probability method (i.e., MP), two of the commonly used adaptive selection methods in MCAT scenario, are migrated here to efficiently measure the time-course of adaptive changes in perception. Naturally, an interesting follow-up study is to investigate whether other proven-to-be-effective adaptive selection methods in MCAT can also be applied to the perception adaptation experiments. These methods include, for instance, D-optimality (Segall, 1996) and A-optimality (van der Linden, 1999) based on Fisher information, the Kullback–Leibler criterion (Veldkamp & van der Linden, 2002), and mutual information (Mulder & van der Linden, 2010). Moreover, even though our simulation study focuses on designs with a fixed number of trials, another parallel simulation study could be conducted to evaluate savings of trials using adaptive termination procedures. For example, the experiment could terminate when the precision of (a, b, c), quantified by the generalized standard error (e.g., Wang, Chang, & Boughton, 2013), is below a certain cutoff. Our preliminary result shows that, for the 40 trials-by-5 blocks design, the ME method could save up to about 50% of trials, as compared to the random selection method.

Second, the joint prior distribution of the key parameters [i.e., $\pi(a, b, c)$] is required in the implementation of both the ME method and EAP estimation. Therefore, selecting appropriate prior $\pi(a, b, c)$ is important for the success of the

method. In this article, we adopt the non-information prior (i.e., uniform prior) and assume the prior is uniform over the range of values selected for these parameters in the experiment—that is, $a \sim U(46, 50)$, $b \sim U(1, 3)$ and $c \sim U(0.01, 0.1)$. Thus, another line of research worth considering is to determine $\pi(a, b, c)$ in other appropriate ways—that is, how to provide informative priors, if possible.

Third, we use Monte Carlo integration to approximate the three-dimensional integrals in ME and EAP. However, for higher dimensional scenarios, the population distribution [i.e., $\pi(a, b, c)$] from which the target parameters are drawn may be particularly complicated, making it impossible to sample directly from $\pi(a, b, c)$. In such cases, one alternative approach could be rejection sampling, where $\pi(a, b, c)$ is considered as the target distribution and a proposal distribution $q(a, b, c)$ that is more easily sampled from is selected such that $\pi(a, b, c) \le c \times q(a, b, c)$ where c is a constant.

Fourth, in our simulations, we only consider three block designs under the premise that the total number of trials is fixed at 200. To explore the optimal block length more accurately, more block designs (e.g., "T = 100, J = 2" and "T = 200, J = 1") should be discussed in future studies. According to our findings—using fewer longer blocks produce better parameter recovery than using more shorter blocks—does this mean that the "100 trials per block and 2 blocks" design will perform better than the designs with 40 (50) trials per block and 5 (4) blocks? Moreover, just like the variable length MCAT in which the test stops when the measurement precision is deemed adequate, stopping rules could also be combined with the ME method to adaptively determine the optimal block design in the future.

Additionally, in this research we fix some parameters (i.e., $\beta = 2$ and $\gamma = \lambda = 0.05$) instead of estimating them, and assume they are the same for all time points. If these parameters are allowed to be freely estimated, then new challenges may emerge and the current ME method may need further modifications.

Last but not the least, this article only considers the scenario in which the b parameter value is positive (i.e., $\alpha(t)$ in Eq. 2 is a "growth" curve). In addition to the "growth" curve, however, in the real experiment we will simultaneously measure the "decay" curve. Thus, the experiment case in which the b value is negative should also be studied in the future so that the complete "growth and decay" curves can be characterized. Fortunately, the methodology presented in this article can be used to characterize the "decay" curve in a straightforward manner.

Acknowledgements This project was supported by grant numbers IES R305D160010, NSF SES-1659328, NIH R01HD079439, and NSFC 31300862.

Open Practice Statement The simulated data and source code for all experiments are available at https://sites.uw.edu/pmetrics/publications-and-source-code/.

Appendix A: Detailed calculation steps and simplified calculation strategies for ME method

The integral terms in Eq. 8 can also be calculated numerically by using Monte Carlo integration. More specifically, the posterior predictive probability

$$P(y^k = y|\mathbf{y}^{k-1}) = \frac{\int P(y^k = y|a,b,c)L(\mathbf{y}^{k-1}|a,b,c)\pi(a,b,c)d(a,b,c)}{\int L(\mathbf{y}^{k-1}|a,b,c)\pi(a,b,c)d(a,b,c)} \approx \frac{\sum_{m=1}^{M} P\left(y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)L\left(\mathbf{y}^{k-1}|a^{(m)},b^{(m)},c^{(m)}\right)}{\sum_{m=1}^{M} L\left(\mathbf{y}^{k-1}|a^{(m)},b^{(m)},c^{(m)}\right)}, \# \tag{A1}$$

and the posterior continuous entropy after k trials

$$\begin{aligned}
&H\left(p(a,b,c|\mathbf{y}^{k-1},y^k = y)\right)\\
&= \int \frac{L(\mathbf{y}^{k-1},y^k = y|a,b,\ \ c)}{\int L(\mathbf{y}^{k-1},y^k = y|a,b,c)\pi(a,b,c)d(a,b,c)} \log\left(\frac{\int L(\mathbf{y}^{k-1},y^k = y|a,b,c)\pi(a,b,c)d(a,b,c)}{L(\mathbf{y}^{k-1},y^k = y|a,b,c)\pi(a,b,c)}\right)\pi(a,b,c)d(a,b,c)\\
&\approx \frac{1}{M}\sum_{m=1}^{M}\left(\frac{L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)}{\frac{1}{M}\sum_{m=1}^{M} L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)} \log\left(\frac{\frac{1}{M}\sum_{m=1}^{M} L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)}{L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)\pi\left(a^{(m)},b^{(m)},c^{(m)}\right)}\right)\right)\\
&= \sum_{m=1}^{M}\left(\frac{L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)}{\sum_{m=1}^{M} L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)} \log\left(\frac{\frac{1}{M}\sum_{m=1}^{M} L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)}{L\left(\mathbf{y}^{k-1},y^k = y|a^{(m)},b^{(m)},c^{(m)}\right)\pi\left(a^{(m)},b^{(m)},c^{(m)}\right)}\right)\right)
\end{aligned} \tag{A2}$$

where

$$L\left(\boldsymbol{y}^{k-1}, y^k = y, a^{(m)}, b^{(m)}, c^{(m)}\right) = L\left(\boldsymbol{y}^{k-1}|a^{(m)}|b^{(m)}|c^{(m)}\right)P\left(y^k = y|a^{(m)}|b^{(m)}|c^{(m)}\right), \#(A3)$$

and $\pi(a^{(m)}, b^{(m)}, c^{(m)})$ is the prior probability of data point $(a^{(m)}, b^{(m)}, c^{(m)})$.

Several details deserve further elaboration when implementing the ME method: (1) the first stimulus can be randomly selected from the set of stimuli with medium-sized orientations (e.g., x' = 47:0.25:49 if the entire orientation space is x = 41:0.25:55). (2) The computation of Eq. 8 could be much faster if the interim steps are stored. Specifically, for both $L(\boldsymbol{y}^{k-1}|\, a^{(m)}|\, b^{(m)}|\, c^{(m)})$ and $P(y^k = y|\, a^{(m)}|\, b^{(m)}|\, c^{(m)})$ (m = 1, 2, . . . , M) because $L(\boldsymbol{y}^{k-1}|\, a^{(m)}|\, b^{(m)}|\, c^{(m)}) = L(\boldsymbol{y}^{k-2}|\, a^{(m)}|\, b^{(m)}|\, c^{(m)})P(y^{k-1}|\, a^{(m)}|\, b^{(m)}|\, c^{(m)})$, the currently calculated likelihood value based on $\boldsymbol{y}^{k-2}$can be saved and used for the calculation of the next likelihood, hence greatly reducing the computation cost. A similar strategy can be applied to calculate $P(y^k = y|\, a^{(m)}|\, b^{(m)}|\, c^{(m)})$ because this probability evaluated at time point t within the jth block is equivalent to that evaluated at the same time point within the (j+1)th block. (3) Matrix operations in MATLAB can be used to efficiently calculate the above formula.

References

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*, 887–905.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review*, *18*, 204–210.

Chang, H.-H., & Ying, Z. L. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Chen, P., Wang, C., Xin, T., & Chang, H.-H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 81–117.

Cheng, Y. (2008). *Computerized adaptive testing—New developments and applications* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Clifford, C. W. G., & Rhodes, G. (2005). Fitting the mind to the world: Adaptation and after-effects in high-level vision. New York, NY: Oxford University Press.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, *47*, 3125–3131.

Cobo-Lewis, A. (1996). An adaptive method for estimating multiple parameters of a psychometric function. *Journal of Mathematical Psychology*, *40*, 353–354.

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology*, *75*, 485–491.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, *69*, 1763–1769.

Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, *50*, 369–389.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*, 1279–1292.

Lesmes, L. A., Lu, Z.-L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision*, *10*(3), 17:1–21. doi:https://doi.org/10.1167/10.3.17

Lesmes, L. A., Lu, Z.-L., Baek, J., Tran, N., Dosher, B. A., & Albright, T. D. (2015). Developing Bayesian adaptive methods for estimating sensitivity thresholds (d') in yes–no and forced-choice tasks. *Frontiers in Psychology*, *6*, 1070. doi:https://doi.org/10.3389/fpsyg.2015.01070

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–477.

Mei, G. X., Dong, X., & Bao, M. (2017). The timescale of adaptation at early and mid-level stages of visual processing. *Journal of Vision*, *17*(1), 1:1–7. doi:https://doi.org/10.1167/17.1.1

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback–Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 77–101). New York, NY: Springer.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*, 53–67.

Patterson, C. A., Wissig, S. C., & Kohn, A. (2013). Distinct effects of brief and prolonged adaptation on orientation tuning in primary visual cortex. *Journal of Neuroscience*, *33*, 532–543.

Pavan, A., Marotti, R. B., & Campana, G. (2012). The temporal course of recovery from brief (sub-second) adaptations to spatial contrast. *Vision Research*, *62*, 116–124.

Pugh, E. N., Nikonov, S., & Lamb, T. D. (1999). Molecular mechanisms of vertebrate photoreceptor light adaptation. *Current Opinion in Neurobiology*, *9*, 410–418.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225–255.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test context. *Psychometrika*, *67*, 575–588.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*, 428–449.

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, *76*, 363–384.

Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99–122.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113–120. doi:https://doi.org/10.3758/BF03202828

Zaidi, Q., Ennis, R., Cao, D. C., & Lee, B. (2012). Neural locus of color afterimages. *Current Biology*, *22*, 220–224.

Zhao, Y. K., Lesmes, L. A., & Lu, Z.-L. (2017). The quick change detection method: Bayesian adaptive assessment of the time course of