BRILL

Seeing and Perceiving 24 (2011) 91–124

SEEING AND PERCEIVING

brill.nl/sp

Statistical Significance Testing with Mahalanobis Distance for Thresholds Estimated from Constant Stimuli Method

Takehiro Nagai [1,2,*], **Takahiro Hoshino** [3] **and Keiji Uchikawa** [1]

[1] Department of Information Processing, Tokyo Institute of Technology, 4259- G2-1 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan

[2] Current address: Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka Tenpaku, Toyohashi, Aichi 441-8580, Japan

[3] Graduate School of Economics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Received 5 August 2010; accepted 8 February 2011

Abstract

The *t*-test and the analysis of variance are commonly used as statistical significance testing methods. However, they cannot assess the significance of differences between thresholds within individual observers estimated from the constant stimuli method; these thresholds are not defined as averages of samples, but they are rather defined as functions of parameters of psychometric functions fitted to participants' responses. Moreover, the statistics necessary for these statistical testing methods cannot be derived. In this paper, we propose a new statistical testing method to assess the statistical significance of differences between thresholds estimated from the constant stimuli method. The new method can assess not only threshold differences but also main effects and interactions in multifactor experiments, exploiting the asymptotic normality of maximum likelihood estimators and the characteristics of multivariate normal distributions. This proposed method could be used in similar cases to the analysis of variance for thresholds estimated from the adjustment method and the staircase method. Finally, we present some data on simulations in which we tested assumptions, power and type I error of the proposed method.

Keywords

Psychophysics, constant stimuli method, statistical hypothesis testing, Monte Carlo simulation

1. Introduction

In many psychophysical studies, thresholds such as detection thresholds or discrimination thresholds are estimated from various responses of the participants. After the threshold estimations, an intriguing question is whether there are significant differ-

* To whom correspondence should be addressed. E-mail: nagai@tut.jp

DOI:10.1163/187847511X568180

ences between the thresholds, that is, whether there are some significant effects of experimental condition differences on thresholds. Statistical significance testing is a natural approach for answering this question, because participants' responses may be very variable and the thresholds estimated from these responses are considered as probability variables.

The t-test and the analysis of variance (ANOVA) are popular statistical testing methods used to assess such threshold differences. Both these methods test differences between multiple population means using their samples. In both these methods, it is necessary to calculate statistical values, such as t-values, F-values, and degrees of freedom from means and variances of samples. In psychophysical experiments, the means of response values in the adjustment method and the means of 'reversal' points where the experimental series turn in the staircase method are frequently defined as thresholds. We can apply the t-test and the ANOVA to assess differences between thresholds estimated from the adjustment method by regarding each of those response values as an independent sample from a population. When a conclusive threshold is defined as a mean of multiple thresholds that is repeatedly measured in different sessions for each experimental condition, the t-test and the ANOVA can also be applied in the staircase method by considering each threshold as a single sample.

The constant stimuli method can also be used to estimate thresholds accurately and is frequently used in psychophysical experiments. The definition of a threshold in the constant stimuli method is mostly different from that in the adjustment method and the staircase method. In the constant stimuli method, a psychometric function, such as a cumulative normal distribution function or a logistic function, is fitted to the percentage of 'yes' responses in a yes–no task (or correct responses in a forced-choice task) as a function of the variable of interest, and then the variable values corresponding to some percentages of 'yes' responses (typically 50% and such are used) are defined as thresholds. We cannot apply the t-test and the ANOVA for thresholds estimated from the constant stimuli method, because they are not means of samples. If the thresholds are estimated for many participants from the constant stimuli method, we might apply the t-test and the ANOVA for those thresholds by regarding each individual threshold as a sample. Many studies have tested threshold differences in this way (e.g., Kingdom and Kasrai, 2006; Nagy *et al.*, 2005; te Pas and Koenderink, 2004). However, this method has some disadvantages; for example, threshold differences for an individual participant cannot be assessed. Moreover, this method does not evaluate statistics in estimating each participant's thresholds, and if data are sparse, then the imprecision of the individual estimates can make this method inefficient.

There are also some other statistical testing methods. Yssaad-Fesselier and Knoblauch (2006) (hereafter, we abbreviate this method to the YFK method) have proposed a testing method for threshold differences. This method introduces psychometric function parameters corresponding to threshold differences between multiple psychometric functions. Although the detailed methods for testing differences

between more than three thresholds are not explicitly stated in their paper, their method is very flexible and can be applied to different kinds of statistical testing about psychometric functions. Matlab scripts employing the rationale to analyze data from the constant stimuli method have also been developed (Prins and Kingdom, 2009). The bootstrap method (Efron, 1982; Wichmann and Hill, 2001b) may also be applied as a testing method for these threshold difference issues by resampling thresholds from two different psychometric functions under null hypothesis and evaluating bootstrap distributions of the threshold differences. For example, Wichmann and Hill (2001b) suggested that the bootstrap confidence interval can be employed as an index for testing threshold differences, although, to our knowledge, there are no publications explicitly indicating this bootstrap approach for threshold differences. In other words, although these approaches may be used for testing threshold differences estimated from the constant stimuli method, there are no publications that specifically exposit statistical testing methods for such threshold differences from the constant stimuli method.

In this paper, we propose a new statistical significance testing method for thresholds estimated from the constant stimuli method. In this proposed method, it is assumed that an estimated threshold is normally distributed according to the asymptotic normality of maximum likelihood estimators. We apply characteristics of multivariate normal distributions to assess threshold differences, since threshold differences are normally distributed under this assumption. A unique aspect of the proposed method is to use the characteristics of a multivariate normal distribution in which a Mahalanobis squared distance is distributed as chi-square. This method is superior to the t-test and the ANOVA in some aspects; the threshold differences in each participant can be independently assessed, and all the responses of participants are used for testing, leading to efficient testing results. Although we do not propose any novel mathematical theories but rather combine existing statistical theories to test the significance of threshold differences, we propose this method because it might help to statistically analyze the data in practical experiments with the constant stimuli method. In addition, the proposed method can be applied to the data derived from the staircase method if the data can be analyzed by fitting a psychometric function in a manner similar to the constant stimuli method.

2. Statistical Significance Testing in a One-Factor Experiment

In this section, we describe the method to assess differences of multiple thresholds estimated from the constant stimuli method in a one-factor experiment. This method corresponds to the one-way ANOVA for thresholds defined as sample means.

2.1. Threshold in the Constant Stimuli Method

Figure 1 shows an example of synthetic data derived from the constant stimuli method. In the constant stimuli method, the observer's response to a stimulus with a given parameter value is a binary response: 'yes' or 'no' in a yes–no task, or

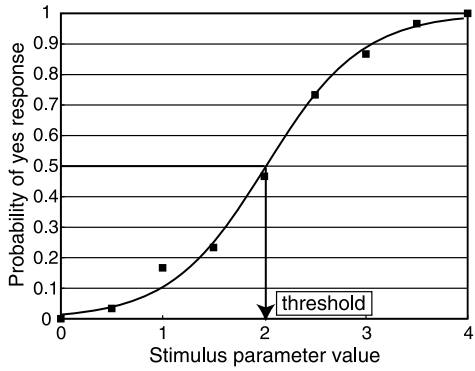

Figure 1. Example of synthetic data derived from the constant stimuli method. A logistic model is fitted to the data with the maximum likelihood method. The stimulus parameter value corresponding to a certain probability of a 'yes' (or correct) response, such as 50%, is defined as a stimulus value.

correct or incorrect in a forced-choice task. The probability of a 'yes' (or correct) response for each stimulus parameter value can be estimated from the collective response of the observers. In general, this probability of a 'yes' response monotonically increases with the stimulus parameter in many cases, forming a psychometric function. This can be observed in Fig. 1.

As described above, a threshold is defined as the stimulus parameter value that corresponds to a certain probability of a 'yes' response. To derive this stimulus parameter value, a psychometric function is usually fitted using a sigmoid statistical model such as a logistic model, a probit model, or a Weibull model. The sigmoid model used in Fig. 1 is a logistic model. The logistic model is expressed as

$$f(x) = \frac{1}{1 + \exp((\theta_0 - x)/\theta_1)}, \tag{1}$$

where x is stimulus parameter value, and θ_0 and θ_1 are parameters (independent variables) of the model. By modeling the psychometric function, a stimulus parameter value corresponding to an arbitrary probability of a 'yes' response can be easily calculated with the parameters of the model. For example, in a logistic model, the stimulus parameter value corresponding to 50% of a 'yes' response (x_{50}) is simply θ_0. Thus, thresholds can be expressed in terms of the parameters of the fitted model based on data collected using the constant stimuli method.

In the following subsections, we describe the proposed model for the statistical testing of the difference between multiple thresholds estimated from different psychometric functions and the statistics needed for the testing.

2.2. Null Hypothesis and Alternative Hypothesis

Here, $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n)$ denotes n thresholds estimated from the constant stimuli method from one observer for different experimental conditions, and $\mathbf{y} = (y_1, y_2, y_3, \ldots, y_n)$ denotes the true values of $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n)$ (an estimated value of an arbitrary value z is denoted by $\hat{z}$ in this paper). For these thresholds,

we want to assess whether there are any differences between the true values of the thresholds. Then, the null hypothesis and the alternative hypothesis are

$$\begin{aligned} H_0\text{: } y_1 = y_2 = y_3 = \cdots = y_n \quad \text{and} \\ H_1\text{: not } H_0. \end{aligned} \tag{2}$$

We introduce a new H_0 that is equivalent to equation (2) but easier to assess than equation (2). Using $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, \ldots, \tau_{n-1})$, whose elements are defined by

$$\tau_i = y_1 - y_{i+1}, \quad i = 1, 2, 3, \ldots, n-1, \tag{3}$$

the null hypothesis shown in equation (2) is equivalent to

$$H_0\text{: } \boldsymbol{\tau} = \mathbf{0}. \tag{4}$$

Thus, we should judge whether H_0 in equation (4) is rejected to assess the differences between the true values of thresholds.

2.3. Statistics of $\hat{\boldsymbol{\tau}}$

In this section, we describe the statistics of $\hat{\boldsymbol{\tau}}$ necessary for the test. First, we describe the way in which each estimated threshold $\hat{y}_i$ $(i = 1, 2, \ldots, n)$ is distributed. In many cases, the threshold value can be estimated directly as a parameter of the model fit to the data. However, in this study, we show how the mean and variance of the threshold can be derived under fairly general assumptions even when the model parameters do not directly correspond to the threshold itself but the threshold is expressed as a function of the model parameters. We then apply related procedures to the problem of testing the statistical significance of differences between thresholds. To estimate a threshold as described above, it is necessary to calculate $\hat{\boldsymbol{\theta}}$, the variables of the sigmoid model, to best fit it to the observer's responses derived from the constant stimuli method. To consider the threshold's distribution, the maximum likelihood method, a popular method for fitting, has a convenient characteristic. The distribution of maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ (e.g., $\hat{\theta}_0$ and $\hat{\theta}_1$ in a logistic function) is an asymptotically multivariate normal distribution with mean vector θ and its variance–covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, if many responses of an observer are used for estimating $\hat{\boldsymbol{\theta}}$ (Edwards, 1993). In addition, the estimated variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ can be asymptotically estimated as an inverse of a Fisher information matrix calculated by using the maximum likelihood method (Pratt, 1976). Because $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is typically unknown, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ is used in practice instead of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ as in other statistical testing methods.

To consider the distribution of $\hat{y}_i$, the Delta method (Oehlert, 1992) is very useful. The Delta method can be used to approximately calculate the means and variances of random variables that can be expressed as functions of asymptotically normal random variables. Hence, we can obtain the means and variances of asymptotically normal distributions of thresholds and standard errors (SEs), which are functions of $\hat{\boldsymbol{\theta}}$, from the Delta method.

2.3.1. Theorem 1: The Delta Method
Asymptotically $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_p \end{bmatrix} \sim N \left(\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \right), \tag{5}$$

where '$\sim N$' means the left-hand side is normally distributed and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is a variance–covariance matrix; similarly, a set of functions $\mathbf{g}(\hat{\boldsymbol{\theta}}) = [g_1(\hat{\boldsymbol{\theta}}), g_2(\hat{\boldsymbol{\theta}}), g_3(\hat{\boldsymbol{\theta}}), \ldots, g_q(\hat{\boldsymbol{\theta}})]^t$ $(q < p)$ that can be approximated over the range of interest by linear combinations of the elements of $\hat{\boldsymbol{\theta}}$, is asymptotically represented as

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} g_1(\hat{\boldsymbol{\theta}}) \\ g_2(\hat{\boldsymbol{\theta}}) \\ \vdots \\ g_q(\hat{\boldsymbol{\theta}}) \end{bmatrix} \sim N \left(\begin{bmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \\ \vdots \\ g_q(\boldsymbol{\theta}) \end{bmatrix}, \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{G}^t \right), \tag{6}$$

where $\mathbf{G}$ is a $q \times p$ matrix whose each element G_{ij} is

$$G_{ij} = \left(\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j} \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{7}$$

From equation (6), the asymptotic means and variance–covariance matrix of $\mathbf{g}(\hat{\boldsymbol{\theta}})$ are written respectively as

$$\begin{gathered} [g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), g_3(\boldsymbol{\theta}), \ldots, g_q(\boldsymbol{\theta})]^t \quad \text{and} \\ \boldsymbol{\Sigma}_{\mathbf{g}(\boldsymbol{\theta})} = \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{G}^t. \end{gathered} \tag{8}$$

From Theorem 1, if the distribution of the estimated parameters of a psychometric function is an asymptotically multivariate normal distribution, and each estimated threshold $\hat{y}_i$ $(i = 1, 2, \ldots, n)$ is a function of those parameters as described, then $\hat{y}_i$ is also asymptotically normal with mean y_i and variance $\hat{\sigma}_i^2$ that is estimated from $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ from equation (8) ($\hat{\sigma}_i^2$ corresponds to $\boldsymbol{\Sigma}_{\mathbf{g}(\boldsymbol{\theta})}$, the left-hand side term of equation (8)). Therefore, all the thresholds $\hat{y}_i$ $(i = 1, 2, \ldots, n)$ are asymptotically normal. Thus, the distribution of $\hat{\mathbf{y}}$ is an asymptotically multivariate normal distribution with mean vector $\mathbf{y}$ and variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ is

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{y}} = \begin{bmatrix} \hat{\sigma}_1^2 & & & \mathbf{0} \\ & \hat{\sigma}_2^2 & & \\ & & \ddots & \\ \mathbf{0} & & & \hat{\sigma}_n^2 \end{bmatrix}, \tag{9}$$

if the thresholds are independent, that is, if the covariances between the thresholds are zero.

Finally, we describe the distribution of $\hat{\boldsymbol{\tau}}$, where $\hat{\boldsymbol{\tau}}$ is a function of $\hat{\mathbf{y}}$ as shown in equation (3). If the distribution of $\hat{\mathbf{y}}$ is an asymptotically multivariate normal

distribution, the distribution of $\hat{\boldsymbol{\tau}}$ is also an asymptotically multivariate normal distribution according to the Delta method. The mean vector of the distribution is $\boldsymbol{\tau}$, and the variance–covariance matrix is calculated from $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ and equation (8).

2.4. Characteristics of a Multivariate Normal Distribution

We assume that $\hat{\boldsymbol{\tau}}$ is multivariately normal to assess threshold differences, because the distribution of $\hat{\boldsymbol{\tau}}$ is asymptotically multivariate normal, as described in Section 2.2. A multivariate normal distribution has an important characteristic for testing (Mardia *et al.*, 1979).

2.4.1. Theorem 2: Mahalanobis Squared Distance

For p-dimensional $\hat{\boldsymbol{\tau}}$, whose distribution is a p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\tau}$ and variance–covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis squared distance

$$d^2 = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})^t \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \tag{10}$$

is chi-square distributed with p degrees of freedom.

Again, the null hypothesis is $\boldsymbol{\tau} = \mathbf{0}$. Thus, $\hat{\boldsymbol{\tau}}^t {\boldsymbol{\Sigma}_{\boldsymbol{\tau}}}^{-1} \hat{\boldsymbol{\tau}}$ is chi-square distributed with $n - 1$ degrees of freedom under H_0 from Theorem 2. Hence, H_0 should be rejected if $\hat{\boldsymbol{\tau}}^t {\boldsymbol{\Sigma}_{\boldsymbol{\tau}}}^{-1} \hat{\boldsymbol{\tau}}$ is inside the critical region of a chi-square distribution with $n - 1$ degrees of freedom, and should not be rejected if $\hat{\boldsymbol{\tau}}^t {\boldsymbol{\Sigma}_{\boldsymbol{\tau}}}^{-1} \hat{\boldsymbol{\tau}}$ is outside the critical region.

2.5. Procedure of Testing

The procedure to statistically assess the significance of threshold differences is summarized as follows:

1. Analyze the results of a one-factor experiment with the maximum likelihood method, and estimate thresholds $\hat{\mathbf{y}}$ and a variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ (covariances in $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ are zero if thresholds are independent). The null hypothesis and the alternative hypothesis are

$$H_0\colon\ y_1 = y_2 = y_3 = \cdots = y_n \quad \text{and}$$
$$H_1\colon \text{not } H_0, \quad \text{respectively.}$$

2. Calculate $\hat{\boldsymbol{\tau}}$ with equation (3), and H_0 of step 1 is equivalent to

$$H_0\colon\ \boldsymbol{\tau} = \mathbf{0}.$$

   Calculate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}$, the variance–covariance matrix of $\hat{\boldsymbol{\tau}}$, from equation (8) of the Delta method. Assume that $\hat{\boldsymbol{\tau}}$ is multivariately normal.

3. Calculate $\hat{\boldsymbol{\tau}}^t \boldsymbol{\Sigma}_{\boldsymbol{\tau}}^{-1} \hat{\boldsymbol{\tau}}$, which is chi-squared with $n - 1$ degrees of freedom under H_0 according to the characteristic of a multivariate normal distribution. Calculate the p-value using the chi-square distribution with $n - 1$ degrees of freedom.

4. Reject H_0 if the p-value calculated in step 3 was smaller than the significant level (that is, we judge that threshold differences are statistically significant); do not reject H_0 if the p-value was larger (i.e., we judge that threshold differences are not statistically significant).

2.6. Example of Testing

We show an example of statistical significance testing for thresholds estimated from the constant stimuli method in an experiment in which we measured the detection thresholds of color direction (hue) differences from four background colors (whose color directions are 0°, 90°, 180° and 270°) on an isoluminant plane. The observer's responses were derived from a yes–no task. The number of trials and 'yes' responses for each color direction difference between the test color and the background color for each background color direction are shown in Table 1. Psychometric functions (the logistic function was used) fitted to the results by the maximum likelihood method are shown in Fig. 2. Estimated parameters ($\hat{\theta}_0$ and $\hat{\theta}_1$) of psychometric functions fitted to the results, their variances, and their covariances are shown in Table 2. In a logistic function, a threshold $\hat{y}$ corresponding to 50% 'yes' responses is simply

$$\hat{y} = \hat{\theta}_0, \tag{11}$$

where $\hat{\theta}_0$ is one of the estimated parameters of the fitted logistic function, and their variance $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \mathrm{Var}(\hat{\theta}_0), \tag{12}$$

where $\mathrm{Var}(\hat{\theta}_0)$ is a variance of $\hat{\theta}_0$.

The probit and Weibull models can be used as psychometric functions to estimate a threshold and to assess threshold differences with our proposed method if

Table 1.
Trial numbers and 'yes' responses derived from the constant stimuli method (synthetic)

Color direction difference	Background color direction (°)							
	0		90		180		270	
	# of trials	# of yes responses	# of trials	# of yes responses	# of trials	# of yes responses	# of trials	# of yes responses
0	30	0	30	0	30	0	30	0
2	30	3	30	2	30	7	30	2
4	30	9	30	4	30	9	30	6
6	30	17	30	8	30	15	30	12
8	30	25	30	13	30	18	30	22
10	30	28	30	20	30	22	30	26
12	30	30	30	27	30	27	30	29
14	30	30	30	30	30	29	30	30

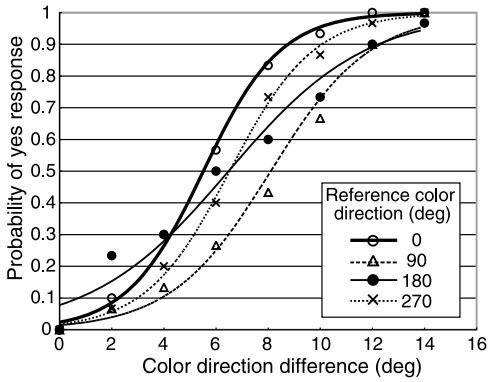

Figure 2. Probabilities of 'yes' responses of Table 1 and the logistic functions fitted to the data as psychometric functions.

Table 2.
Parameters of the logistic functions estimated from the data in Table 1

Color direction difference	Background color direction (°)							
	0		90		180		270	
	Estimate value	Variance	Estimate value	Variance	Estimate value	Variance	Estimate value	Variance
$\hat{\theta}_0$ (threshold)	5.519	0.1	8.100	0.132	6.488	0.192	6.527	0.109
$\hat{\theta}_1$ (covariance)	1.476	0.036	1.912	0.053	2.62	0.103	1.611	0.039
		−0.002		0.004		−0.006		−0.001

the parameter estimations were performed with the maximum likelihood method, although the logistic function is used here.

Differences between thresholds shown in Table 2 could be assessed with the proposed statistical testing method as explained below:

1. Assuming that thresholds are independent, the threshold vector $\hat{\mathbf{y}}$ and its variance–covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{y}}$ are

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \end{bmatrix} = \begin{bmatrix} 5.519 \\ 8.100 \\ 6.488 \\ 6.527 \end{bmatrix} \tag{13}$$

and

$$\hat{\mathbf{\Sigma}}_{\mathbf{y}} = \begin{bmatrix} 0.100 & 0 & 0 & 0 \\ 0 & 0.132 & 0 & 0 \\ 0 & 0 & 0.192 & 0 \\ 0 & 0 & 0 & 0.109 \end{bmatrix}. \tag{14}$$

2. $\hat{\boldsymbol{\tau}}$ are calculated from $\hat{\mathbf{y}}$ and equation (3) as

$$\hat{\boldsymbol{\tau}} = \begin{bmatrix} \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - \hat{y}_2 \\ \hat{y}_1 - \hat{y}_3 \\ \hat{y}_1 - \hat{y}_4 \end{bmatrix} = \begin{bmatrix} -2.581 \\ -0.969 \\ -1.008 \end{bmatrix}. \tag{15}$$

From equation (7), $\mathbf{G}$, which is necessary for the calculation of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}$ (the variance–covariance matrix of $\hat{\boldsymbol{\tau}}$), can be calculated using the Delta method as

$$\mathbf{G} = \begin{bmatrix} \frac{\partial \tau_1}{\partial y_1} & \frac{\partial \tau_1}{\partial y_2} & \frac{\partial \tau_1}{\partial y_3} & \frac{\partial \tau_1}{\partial y_4} \\ \frac{\partial \tau_2}{\partial y_1} & \frac{\partial \tau_2}{\partial y_2} & \frac{\partial \tau_2}{\partial y_3} & \frac{\partial \tau_2}{\partial y_4} \\ \frac{\partial \tau_3}{\partial y_1} & \frac{\partial \tau_3}{\partial y_2} & \frac{\partial \tau_3}{\partial y_3} & \frac{\partial \tau_3}{\partial y_4} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}. \tag{16}$$

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}$ is calculated from $\mathbf{G}$ and equation (8) as

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}} &= \mathbf{G}\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}\mathbf{G}^t \\ &= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0.100 & 0 & 0 & 0 \\ 0 & 0.132 & 0 & 0 \\ 0 & 0 & 0.192 & 0 \\ 0 & 0 & 0 & 0.109 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 0.0232 & 0.010 & 0.010 \\ 0.010 & 0.292 & 0.010 \\ 0.010 & 0.010 & 0.209 \end{bmatrix}. \end{aligned} \tag{17}$$

3. $\hat{\boldsymbol{\tau}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}^{-1} \hat{\boldsymbol{\tau}}$, which is chi-square distributed with three degrees of freedom according to the characteristic of a multivariate normal distribution, is calculated as

$$\begin{aligned} \hat{\boldsymbol{\tau}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}^{-1} \hat{\boldsymbol{\tau}} &= [\, -2.581 \quad -0.969 \quad -1.008 \,] \\ &\quad \times \begin{bmatrix} 0.00177 & 0.00066 & 0.00066 \\ 0.00066 & 0.00209 & 0.00066 \\ 0.00066 & 0.00066 & 0.00191 \end{bmatrix} \begin{bmatrix} -2.581 \\ -0.969 \\ -1.008 \end{bmatrix} \\ &= 28.834. \end{aligned} \tag{18}$$

According to the chi-square table with three degrees of freedom, the p-value corresponding to this $\hat{\boldsymbol{\tau}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}}^{-1} \hat{\boldsymbol{\tau}}$ is

$$p = 2.43 \times 10^{-6}. \tag{19}$$

4. If the significant level is 5%, we reject H_0 according to $p = 2.43 \times 10^{-6} < 0.05$, that is, we judge that the differences of thresholds in Table 2 are statistically significant.

2.7. Multiple Comparison Method

The statistical testing method described above assesses only whether all the true values of thresholds are identical. However, the method does not answer which threshold pairs have significant differences. Multiple comparison methods can be utilized to answer this question.

Many kinds of multiple comparison methods have been published and used. Because our method can test difference between any two thresholds estimated from the constant stimuli method, one can choose a favorable multiple comparison method and combine it with our method. Holm's method (Holm, 1979) is an example of such methods, where significant levels are adjusted so that the total of type I errors in assessments of threshold pairs does not exceed a significant level. If Holm's method is applied to the thresholds in Table 2 with $\alpha = 0.05$, the threshold differences in '0 and 90', '90 and 270' and '90 and 180' are statistically significant, and those in other pairs are not statistically significant.

3. Statistical Significance Testing in a Multifactor Experiment

In the previous section, we described a new statistical significance testing method for thresholds measured in a one-factor experiment. In this section, we propose a testing method for thresholds measured in a factorial experiment. This method corresponds to the multi-way ANOVA.

In a psychophysical experiment, a factor corresponds to a dimension of the experimental conditions whose effects on thresholds are of interest to the investigator. Unlike in a one-factor experiment, in a multifactor experiment, not only the main effect in each factor but also the interaction between factors must be considered. In the proposed method, we assess the main effects and the interactions separately. Using a two-factor experiment as an example, we first describe how to calculate the main effect of each factor and the interaction between factors, and then describe how to assess whether the main effects and the interaction are statistically significant on the basis of the Delta method and the characteristics of a multivariate normal distribution as described in Section 2.

3.1. Main Effect and Interaction

3.1.1. Symbol Introduction

We introduce some symbols. Although we use a two-factor experiment as an example to explain the main effects and interactions, the testing procedure is the same for experiments with more than two factors. Let two factors be denoted by A and B with m and n numbers of levels, respectively, and $m \times n$ thresholds estimated from the constant stimuli method by

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_{1\cdot1} & \hat{y}_{1\cdot2} & \cdots & \hat{y}_{1\cdot n} \\ \hat{y}_{2\cdot1} & \hat{y}_{2\cdot2} & \cdots & \hat{y}_{2\cdot n} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{m\cdot1} & \hat{y}_{m\cdot2} & \cdots & \hat{y}_{m\cdot n} \end{bmatrix}. \tag{20}$$

A level is an experimental condition for each factor, and the number of levels corresponds to the number of experimental conditions for each factor. Then, the grand mean (mean between all the factors) of $\mathbf{y}$ (the true value of $\hat{\mathbf{y}}$) is denoted by μ. The mean of true values of the level i of A between all the factors of B is denoted by $\mu_{i\cdot\cdot}$, and the mean of true values of the level j of B between all the factors of A is denoted by $\mu_{\cdot\cdot j}$. As a result, μ, $\mu_{i\cdot\cdot}$ and $\mu_{\cdot\cdot j}$ are respectively expressed as

$$\mu = \frac{\sum_i \sum_j y_{i\cdot j}}{m \times n}, \tag{21}$$

$$\mu_{i\cdot\cdot} = \frac{\sum_j y_{i\cdot j}}{n}, \tag{22}$$

$$\mu_{\cdot\cdot j} = \frac{\sum_i y_{i\cdot j}}{m}. \tag{23}$$

3.1.2. Main Effect of Each Factor

The main effect of each factor is the influence of the level differences in the factor on the true values of thresholds. The main effect of the level i of A and that of the level j of B are respectively expressed as (Edwards, 1993)

$$\alpha_i = \mu_{i\cdot\cdot} - \mu \tag{24}$$

and

$$\beta_j = \mu_{\cdot\cdot j} - \mu. \tag{25}$$

3.1.3. Interaction between Factors

An interaction is the component of $y_{i\cdot j}$ which cannot be represented by its main effect and the grand mean. The interaction of the level i of A and of the level j of B is (Edwards, 1993)

$$\begin{aligned}(\alpha\beta)_{i\cdot j} &= y_{i\cdot j} - (\alpha_{i\cdot\cdot} + \beta_{\cdot\cdot j} + \mu) \\ &= y_{i\cdot j} - (\mu_{i\cdot\cdot} - \mu) - (\mu_{\cdot\cdot j} - \mu) - \mu \\ &= y_{i\cdot j} - \mu_{i\cdot\cdot} - \mu_{\cdot\cdot j} + \mu. \end{aligned} \tag{26}$$

3.2. How to Assess Main Effects

3.2.1. Null Hypothesis and Alternative Hypothesis

We consider the main effect of factor A. Nonexistence of the main effect of A means that all the main effects of A (equation (24)) are zero. The null hypothesis H_0 and the alternative hypothesis H_1 to test the main effect of A are, respectively,

$$H_0\text{: } \alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_m = 0 \quad \text{and} \quad H_1\text{: not } H_0. \tag{27}$$

In this study, we introduce a new H_0 that is equivalent to equation (27) and easy to assess as in Section 2. Since $\sum_i \alpha_i = 0$, using $\boldsymbol{\tau}_{\mathrm{A}} = (\tau_{\mathrm{A}1}, \tau_{\mathrm{A}2}, \tau_{\mathrm{A}3}, \ldots, \tau_{\mathrm{A}(m-1)})$ whose components are

$$\begin{aligned}\tau_{\mathrm{A}i} &= \alpha_1 - \alpha_{i+1} = (\mu_{1\cdot\cdot} - \mu) - (\mu_{1+1\cdot\cdot} - \mu) \\ &= \mu_{1\cdot\cdot} - \mu_{1+1\cdot\cdot}, \quad i = 1, 2, \ldots, m-1, \end{aligned} \tag{28}$$

the H_0 of equation (27) is equivalent to

$$H_0\colon\ \boldsymbol{\tau}_{\mathrm{A}} = \mathbf{0}. \tag{29}$$

Hence, we can assess the main effect of A by judging whether H_0 of equation (29) is rejected.

3.2.2. Procedure of Testing

The testing procedure is identical to the method described in Section 2. Under the assumption that $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$ is multivariately normal and H_0, $\hat{\boldsymbol{\tau}}_{\mathrm{A}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1} \hat{\boldsymbol{\tau}}_{\mathrm{A}}$ is chi-square distributed with $m - 1$ degrees of freedom according to the characteristic of a multivariate normal distribution, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}$ is a variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$ and can be calculated from $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ (a variance–covariance matrix of $\hat{\mathbf{y}}$) according to the Delta method; then, the procedure to assess the main effect of factor A is summarized as follows:

1. Estimate thresholds $\hat{\mathbf{y}}$ and their variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ (covariances in $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ are zero if the thresholds are independent as in Section 2). Then, H_0 and H_1 are

   $$H_0\colon\ \alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$$

   and

   $$H_1\colon\ \text{not } H_0.$$

2. Calculate $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$. Then, H_0 in step 1 is equivalent to

   $$H_0\colon\ \boldsymbol{\tau}_{\mathrm{A}} = \mathbf{0}.$$

   Calculate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}$, the variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$, according to the Delta method. Assume that $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$ is multivariately normal.

3. Calculate $\hat{\boldsymbol{\tau}}_{\mathrm{A}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1} \hat{\boldsymbol{\tau}}_{\mathrm{A}}$, which is chi-square distributed with $m - 1$ degrees of freedom under H_0. Compute the p-value corresponding to $\hat{\boldsymbol{\tau}}_{\mathrm{A}}^t \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1} \hat{\boldsymbol{\tau}}_{\mathrm{A}}$ from the chi-square distribution.

4. Reject H_0 if the p-value is smaller than the significant level (the main effect of the factor A is statistically significant), and do not reject H_0 if the p-value is larger than the significant level (the main effect of the factor A is not statistically significant).

The assessment of the main effect of B is identical to the assessment of A, which is described above.

3.3. Procedure to Assess Interactions

3.3.1. Null Hypothesis and Alternative Hypothesis

The number of interactions ($(\alpha\beta)_{i\cdot j}$) is $m \times n$ according to the combination of levels of A and B. Nonexistence of the interaction between A and B means that

interactions for all the combinations of A and B are zero. To assess whether there is a significant interaction between A and B, H_0 and H_1 are given as

$$H_0\colon\ (\alpha\beta)_{1\cdot1} = (\alpha\beta)_{1\cdot2} = \cdots = (\alpha\beta)_{2\cdot1} = \cdots = (\alpha\beta)_{m\cdot n} = 0$$

and

$$H_1\colon \text{ not } H_0.$$

We introduce a new H_0 as done in the previous sections. Since $\sum_i (\alpha\beta)_{i\cdot j} = 0$ and $\sum_j (\alpha\beta)_{i\cdot j} = 0$ (Edwards, 1993), using

$$\begin{aligned}\boldsymbol{\tau}_{\mathrm{AB}} = (&\tau_{\mathrm{AB}(1\cdot1)}, \tau_{\mathrm{AB}(2\cdot1)}, \tau_{\mathrm{AB}(3\cdot1)}, \ldots, \tau_{\mathrm{AB}(m-1\cdot1)},\\ &\tau_{\mathrm{AB}(1\cdot2)}, \tau_{\mathrm{AB}(2\cdot2)}, \ldots, \tau_{\mathrm{AB}(m-1\cdot n-1)})\end{aligned}$$

whose components are

$$\begin{aligned}\boldsymbol{\tau}_{\mathrm{AB}(i\cdot j)} &= (\alpha\beta)_{i\cdot1} - (\alpha\beta)_{i\cdot j+1}\\ &= (y_{i\cdot1} - \mu_{i\cdot\cdot} - \mu_{\cdot\cdot1} + \mu) - (y_{i\cdot j+1} - \mu_{i\cdot\cdot} - \mu_{\cdot\cdot j+1} + \mu)\\ &= y_{i\cdot1} - y_{i\cdot j+1} - \mu_{\cdot\cdot1} + \mu_{\cdot\cdot j+1},\\ i &= 1, 2, 3, \ldots, m-1,\ j = 1, 2, 3, \ldots, n-1,\end{aligned} \tag{30}$$

the null hypothesis above is equivalent to

$$H_0\colon\ \boldsymbol{\tau}_{\mathrm{AB}} = \mathbf{0}. \tag{31}$$

Hence, we can assess the interaction between A and B by judging whether H_0 of equation (31) is rejected.

3.3.2. Procedure of Testing

The testing should be performed according to the procedures proposed in the previous sections. Under the assumption that $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ is multivariately normal and H_0, $\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}} \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}} \hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ is chi-square distributed with $(m-1)\times(n-1)$ degrees of freedom according to the characteristic of a multivariate normal distribution, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{AB}}$ is a variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ and can be calculated from $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ (a variance–covariance matrix of $\hat{\mathbf{y}}$) according to the Delta method; then, the procedure to assess the interaction between the factors A and B is summarized as follows:

1. Estimate thresholds $\hat{\mathbf{y}}$ and their variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ (covariances in $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ are zero if thresholds are independent as in Section 2). Then, H_0 and H_1 are

   $$H_0\colon\ (\alpha\beta)_{1\cdot1} = (\alpha\beta)_{1\cdot2} = \cdots = (\alpha\beta)_{2\cdot1} = \cdots = (\alpha\beta)_{m\cdot n} = 0$$

   and

   $$H_1\colon \text{ not } H_0.$$

2. Calculate $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$. Then, H_0 in step 1 is equivalent to

   $$H_0\colon\ \boldsymbol{\tau}_{\mathrm{AB}} = \mathbf{0}.$$

   Calculate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{AB}}$, a variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$, according to the Delta method. Assume that $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ is multivariately normal.

3. Calculate $\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}}\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$, which is chi-square distributed with $(m-1)\times(n-1)$ degrees of freedom under H_0. Calculate the p-value corresponding to $\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}}\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ from the chi-square distribution.
4. Reject H_0 if the p-value is smaller than the significant level (the interaction between A and B is statistically significant), and do not reject H_0 if the p-value is larger than the significant level (the interaction between A and B is not significant).

3.4. Example of Testing

3.4.1. Results for Testing

Color detection thresholds and their variances shown in Table 3 are used as data for testing the main effects and the interaction. The thresholds are estimated as the color differences in the OSA color space corresponding to 50% 'yes' responses with logistic functions, although the number of trials and 'yes' responses are omitted. The thresholds are plotted in Fig. 3. The experimental conditions (factors) are the

Table 3.
Thresholds estimated in a two-factorial experiment from the constant stimuli method (synthetic)

			Direction	
			$+j$	$-j$
Background color	(0, 2, 2)	Threshold	1.96	3.57
		Variance	0.58	0.62
	(0, −2, 2)	Threshold	3.95	1.23
		Variance	0.74	0.65
	(0, 2, −2)	Threshold	2.35	2.10
		Variance	0.77	0.79

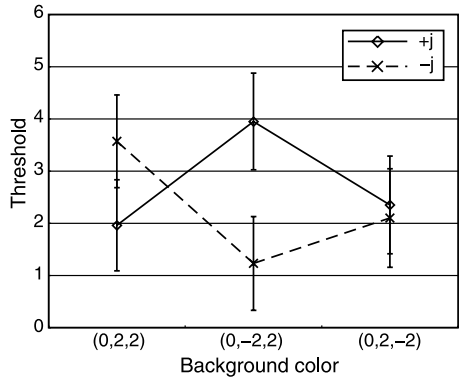

Figure 3. Plots of the thresholds of Table 3. The error bars are the ±1 SEs.

background colors in the OSA color space ($(L, j, g) = (0, 2, 2)$, $(0, -2, 2)$ and $(0, 2, -2)$) and the color shift directions $(+j, -j)$. The background color is identified by the factor A, and the direction by the factor B.

3.4.2. Testing of the Main Effect

Here, we assess the main effect of the background color:

1. Under the assumption that thresholds are independent, from Table 3, the threshold vector $\hat{\mathbf{y}}$ and its variance–covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{y}}$ are

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_{1\cdot1} \\ \hat{y}_{1\cdot2} \\ \hat{y}_{2\cdot1} \\ \hat{y}_{2\cdot2} \\ \hat{y}_{3\cdot1} \\ \hat{y}_{3\cdot2} \end{bmatrix} = \begin{bmatrix} 1.96 \\ 3.57 \\ 3.95 \\ 1.23 \\ 2.35 \\ 2.10 \end{bmatrix} \tag{32}$$

and

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{\mathbf{y}} &= \begin{bmatrix} \hat{\sigma}_{1\cdot1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}_{1\cdot2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}_{2\cdot1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}_{2\cdot2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}_{3\cdot1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\sigma}_{3\cdot2} \end{bmatrix} \\ &= \begin{bmatrix} 0.58 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.62 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.74 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.65 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.77 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.79 \end{bmatrix}. \end{aligned} \tag{33}$$

2. $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$ is calculated from $\hat{\mathbf{y}}$, equation (28) and equation (22), and is given as

$$\hat{\boldsymbol{\tau}}_{\mathrm{A}} = \begin{bmatrix} \hat{\tau}_{\mathrm{A1}} \\ \hat{\tau}_{\mathrm{A2}} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{1\cdot\cdot} - \hat{\mu}_{2\cdot\cdot} \\ \hat{\mu}_{1\cdot\cdot} - \hat{\mu}_{3\cdot\cdot} \end{bmatrix} = \begin{bmatrix} \frac{y_{1\cdot1}+y_{1\cdot2}}{2} - \frac{y_{2\cdot1}+y_{2\cdot2}}{2} \\ \frac{y_{1\cdot1}+y_{1\cdot2}}{2} - \frac{y_{3\cdot1}+y_{3\cdot2}}{2} \end{bmatrix} = \begin{bmatrix} 0.175 \\ 0.54 \end{bmatrix}. \tag{34}$$

From equation (7), $\mathbf{G}$ necessary for calculation of $\hat{\mathbf{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}$ (the variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{A}}$) with the Delta method is

$$\begin{aligned} \mathbf{G} &= \begin{bmatrix} \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{1\cdot1}} & \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{1\cdot2}} & \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{2\cdot1}} & \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{2\cdot2}} & \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{3\cdot1}} & \frac{\partial \tau_{\mathrm{A1}}}{\partial y_{3\cdot2}} \\ \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{1\cdot1}} & \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{1\cdot2}} & \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{2\cdot1}} & \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{2\cdot2}} & \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{3\cdot1}} & \frac{\partial \tau_{\mathrm{A2}}}{\partial y_{3\cdot2}} \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 0.5 & -0.5 & -0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & -0.5 & -0.5 \end{bmatrix}. \end{aligned} \tag{35}$$

$\hat{\mathbf{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}$ is calculated from $\mathbf{G}$ and equation (8) as

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}} = \mathbf{G}\hat{\mathbf{\Sigma}}_{\mathbf{y}}\mathbf{G}^{t} = \begin{bmatrix} 0.6475 & 0.3 \\ 0.3 & 0.69 \end{bmatrix}. \tag{36}$$

3. $\hat{\boldsymbol{\tau}}_{\mathrm{A}}^{t}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1}\hat{\boldsymbol{\tau}}_{\mathrm{A}}$, which is chi-squared with two degrees of freedom according to the characteristic of a multivariate normal distribution, is calculated as

$$\hat{\boldsymbol{\tau}}_{\mathrm{A}}^{t}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1}\hat{\boldsymbol{\tau}}_{\mathrm{A}} = 0.430. \tag{37}$$

According to the chi-square table with two degree of freedom, the p-value corresponding to this $\hat{\boldsymbol{\tau}}_{\mathrm{A}}^{t}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{A}}^{-1}\hat{\boldsymbol{\tau}}_{\mathrm{A}}$ is

$$p = 0.807. \tag{38}$$

4. If the significant level is 5%, we cannot reject H_0 according to $p = 0.807 > 0.05$, that is, we judge that the main effect of the background color is not statistically significant.

The main effect of the direction can be assessed in the same way. The p-value for the main effect of the direction is

$$p = 0.504. \tag{39}$$

3.4.3. Testing of the Interaction

We assess the interaction between the background color and the direction as follows:

1. Under the assumption that thresholds are independent, threshold vector $\hat{\mathbf{y}}$ and its variance–covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ are shown in equations (32) and (33), respectively.

2. $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ is calculated from $\hat{\mathbf{y}}$ and equations (30) as

$$\begin{aligned}\hat{\boldsymbol{\tau}}_{\mathrm{AB}} &= \begin{bmatrix}\hat{\tau}_{\mathrm{AB}(1\cdot1)}\\ \hat{\tau}_{\mathrm{AB}(2\cdot1)}\end{bmatrix} = \begin{bmatrix}\hat{y}_{1\cdot1} - \hat{y}_{1\cdot2} - \hat{\mu}_{\cdot\cdot1} + \hat{\mu}_{\cdot\cdot2}\\ \hat{y}_{2\cdot1} - \hat{y}_{2\cdot2} - \hat{\mu}_{\cdot\cdot1} + \hat{\mu}_{\cdot\cdot2}\end{bmatrix}\\ &= \begin{bmatrix}\hat{y}_{1\cdot1} - \hat{y}_{1\cdot2} - \frac{\hat{y}_{1\cdot1}+\hat{y}_{2\cdot1}+\hat{y}_{3\cdot1}}{3} + \frac{\hat{y}_{1\cdot2}+\hat{y}_{2\cdot2}+\hat{y}_{3\cdot2}}{3}\\ \hat{y}_{2\cdot1} - \hat{y}_{2\cdot2} - \frac{\hat{y}_{1\cdot1}+\hat{y}_{2\cdot1}+\hat{y}_{3\cdot1}}{3} + \frac{\hat{y}_{1\cdot2}+\hat{y}_{2\cdot2}+\hat{y}_{3\cdot2}}{3}\end{bmatrix}\\ &= \begin{bmatrix}-2.06\\ 2.27\end{bmatrix}.\end{aligned} \tag{40}$$

From equation (7), $\mathbf{G}$ necessary for calculation of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{AB}}$ (the variance–covariance matrix of $\hat{\boldsymbol{\tau}}_{\mathrm{AB}}$) with the Delta method is

$$\begin{aligned}\mathbf{G} &= \begin{bmatrix}\frac{\partial\gamma_{\mathrm{AB}(1\cdot1)}}{\partial y_{1\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(1\cdot1)}}{\partial y_{1\cdot2}} & \frac{\partial\tau_{\mathrm{AB}(1\cdot1)}}{\partial y_{2\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(1\cdot1)}}{\partial y_{2\cdot2}} & \frac{\partial\tau_{\mathrm{AB}(1\cdot1)}}{\partial y_{3\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(1\cdot1)}}{\partial y_{3\cdot2}}\\ \frac{\partial\gamma_{\mathrm{AB}(2\cdot1)}}{\partial y_{1\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(2\cdot1)}}{\partial y_{1\cdot2}} & \frac{\partial\tau_{\mathrm{AB}(2\cdot1)}}{\partial y_{2\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(2\cdot1)}}{\partial y_{2\cdot2}} & \frac{\partial\tau_{\mathrm{AB}(2\cdot1)}}{\partial y_{3\cdot1}} & \frac{\partial\tau_{\mathrm{AB}(2\cdot1)}}{\partial y_{3\cdot2}}\end{bmatrix}\\ &= \begin{bmatrix}0.667 & -0.667 & -0.333 & 0.333 & -0.333 & 0.333\\ -0.333 & 0.333 & 0.667 & -0.667 & -0.333 & 0.333\end{bmatrix}.\end{aligned} \tag{41}$$

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{AB}}$ is calculated from $\mathbf{G}$ and equation (8) as

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathrm{AB}} = \mathbf{G}\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}\mathbf{G}^{t} = \begin{bmatrix}0.861 & -0.402\\ -0.402 & 0.924\end{bmatrix}. \tag{42}$$

3. $\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}} \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}} \hat{\boldsymbol{\tau}}_{\mathrm{AB}}$, which is chi-squared with two degrees of freedom according to the characteristic of a multivariate normal distribution, is calculated as
$$\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}} \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}} \hat{\boldsymbol{\tau}}_{\mathrm{AB}} = 7.249. \tag{43}$$
From the chi-square table with two degrees of freedom, the p-value corresponding to this $\hat{\boldsymbol{\tau}}^t_{\mathrm{AB}} \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\tau}\mathrm{AB}} \hat{\boldsymbol{\tau}}_{\mathrm{AB}}$ is
$$p = 0.027. \tag{44}$$
4. If the significant level is 5%, we reject H_0 according to $p = 0.027 < 0.05$, that is, we judge that the interaction between the background color and the direction in the color space is statistically significant.

4. Monte Carlo Simulations

In the proposed testing method, the normality of the estimated thresholds must be assumed. Estimated thresholds, such as maximum likelihood estimators, are asymptotically normal as described above, that is, those thresholds should be normal when trial numbers (and stimulus intensity levels) in the constant stimuli method are enormous. The trial numbers in practical experiments, however, may not be sufficient for such a threshold normality assumption. In addition, the accuracy of the SEs of thresholds estimated from the maximum likelihood method is unclear, even if the threshold normality assumption is valid.

In this section, using simple Monte Carlo simulations, we verify (1) the validity of the threshold normality assumption, (2) the accuracy of the estimated SEs of thresholds on the maximum likelihood method, (3) the comparison of power and type I error between the proposed method and the standard t-test and (4) the comparison of power and type I error between the proposed method and the YKF method. These issues should give some criterions regarding the kind of and the amount of data required to be collected before applying our proposed method.

4.1. Normality Assumption of Thresholds

4.1.1. Method

We tested the normality of thresholds estimated from the constant stimuli method. We used an ideal observer whose psychometric function can be described as a logistic function:
$$f(x) = \frac{1-\lambda}{1+\exp((\theta_1 - x)/\theta_2)}. \tag{45}$$
The parameter λ, which decreases the maximum probability of the observer's responses, was introduced to include effects of involuntary response errors of observers (Wichmann and Hill, 2001a). The parameters of the psychometric function were fixed at $\theta_1 = 1.5, \theta_2 = 0.125$ and $\lambda = 0.01$.

Although it has been demonstrated that different experimental conditions, such as trial numbers and stimulus intensity selections, can affect accuracy and efficiency of threshold estimations (Foster and Bischof, 1991; Garcia-Perez and

Table 4.
Values of the selected stimulus intensities for each CSI

	No.	Stimulus intensity						
		SI1	SI2	SI3	SI4	SI5	SI6	SI7
CSI #4	1	0	0.5	1	1.5	2	2.5	3
	2	0.5	0.833	1.167	1.5	1.833	2.167	2.5
	3	1	1.167	1.333	1.5	1.667	1.833	2
	4	0	0.333	0.667	1	1.333	1.667	2
	5	1	1.333	1.667	1	2.333	2.667	3
	6	0.75	0.917	1.083	1.25	1.417	1,583	1.75
	7	1.25	1.417	1.583	1.75	1.917	2.083	2.25

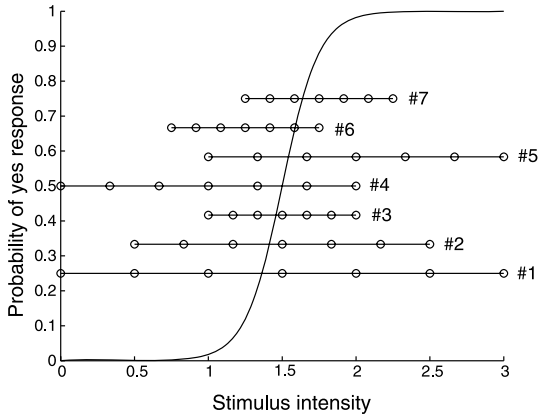

Figure 4. Psychometric function of an ideal observer employed in the simulation of Sections 4.1 and 4.2, and stimulus intensities selected in CSIs #1–#7 (see text). Values of the selected intensities are shown in Table 4.

Alcala-Quintana, 2005), we employed only a few experimental conditions for simplicity. We used seven different combinations of stimulus intensities (CSIs) around the real threshold (1.5; the inflection point of the psychometric function) as shown in Table 4. The psychometric function of the ideal observer and the selected stimulus intensities in CSIs are plotted in Fig. 4. The trial number for each of the stimulus intensities was between 6 and 40.

While estimating each threshold in the simulation, we collected 'yes' responses from ideal observers as in practical psychophysical experiments, and then estimated parameters of psychometric functions utilizing the psignifit toolbox (Wichmann and Hill, 2001a, b), where the shape of the fitted psychometric function was logistic and the parameter lambda was fixed at 0.01. Variance–covariance matrices of the parameters were calculated as inverse of Fisher information matrices derived from the psignifit toolbox. As Wichmann and Hill (2001a) suggested, it should be

preferable to use the psignifit toolbox without fixing the value of lambda so as to minimize errors in the slope estimation. However, we fixed the value of lambda, just for simplicity, so that the thresholds can be defined as stimulus values corresponding to the same 'yes' probability (74.25%). The number of repetition of threshold estimation for each of the CSIs and trial number was 10 000. Thresholds with Fisher information matrices of null were excluded from the following analysis. After estimating all the thresholds, we calculated skewness and kurtosis from the non-excluded thresholds as indices for normality of threshold estimations. These simulations were conducted in Mathworks MATLAB 2007b.

4.1.2. Results

The histogram of estimated thresholds for the CSI #3 and trial number 40 is shown in Fig. 5. This histogram seemed to be derived from a normal distribution. Figure 6

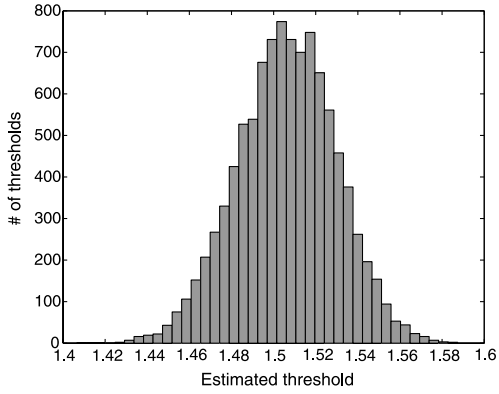

Figure 5. Histogram of estimated thresholds under CSI #3 and trial number for each stimulus intensity of 40.

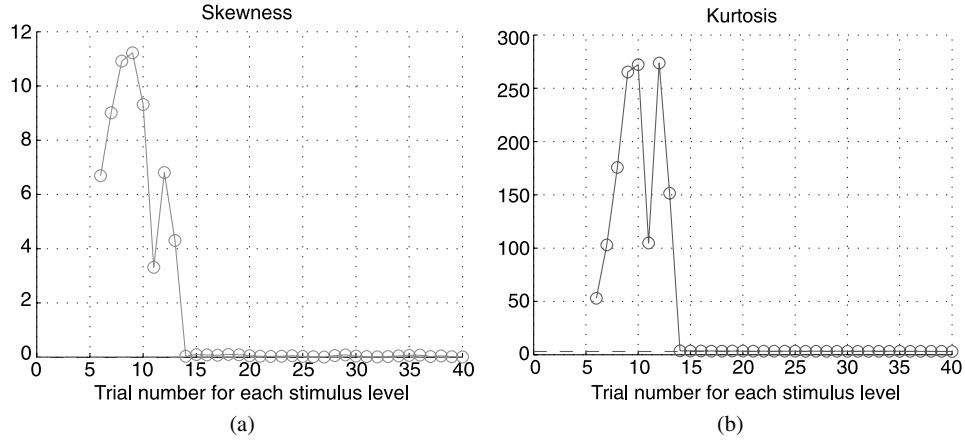

Figure 6. (a) Skewness and (b) kurtosis of 10 000 thresholds estimated under the CSI #2 in the simulation. Dotted lines represent skewness and kurtosis for a normal distribution.

shows the skewness and kurtosis for thresholds of the CSI #2. It is clear that both the skewness and the kurtosis suddenly became stable for trial numbers above a certain value. Indeed, mean values of skewness and kurtosis for small trial numbers largely varied by repeating the same simulations. This tendency was also observed in other CSIs. Therefore, to estimate trial numbers necessary for stable skewness and kurtosis, we arbitrarily defined the 'critical trial number' for each CSI as follows. We used the absolute differences in skewness and kurtosis between adjacent trial numbers (e.g., 23 and 24, or 14 and 15), which we simply call 'adjacent difference', to define the critical trial number. First, 'mean +1 standard deviation' of all of the adjacent differences was defined as a 'criterion difference' in each CSI. Second, a trial number, for trial numbers larger than which the adjacent differences do not exceed the criterion difference, was defined as the critical trial number.

Skewness and kurtosis for trial numbers larger than the critical numbers for all CSIs are shown in Fig. 7(a) and (b), and those with smaller y-axis scales are shown in Fig. 7(c) and (d) ((a) and (c), and (b) and (d) show identical data, respectively). Both the skewness and the kurtosis were much larger for CSI #1 than the other CSIs. This suggests that distribution of the threshold estimations for CSI #1 differs from that for a normal distribution. These results should arise from the fact that the number of stimulus intensities between the intermediate (about 10–90%) 'yes' probability was only one for CSI #1, that is, the sampling of stimulus intensities was too coarse compared to the slope of the psychometric function, leading to inaccurate threshold estimation. In fact, the other finer samplings (#2, 5, 6 and #3, 6, 7) yielded threshold distributions much closer to normal distributions based on skewness and kurtosis. Therefore, multiple stimulus intensities should be allocated around intermediate 'yes' response probabilities to estimate thresholds with stable normality. In addition, skewness and kurtosis came closer to the ideal values of normal distribution (0 and 3, respectively) as trial numbers increased for all CSIs, suggesting that more trial numbers result in better normality of the estimated thresholds, as expected from the asymptotic normality.

Meanwhile, the skewness tended to be slightly biased from 0 depending on the CSIs, indicating slight deviation from the normal distribution. In a similar manner, the kurtosis also tended to be slightly larger than 3, the value corresponding to a normal distribution. Even the finest sampling of stimulus intensities (#3, 6 and 7) yielded slight deviation from normality in skewness and kurtosis. The Jarque–Bera test also rejected the normality of the distribution of the simulated thresholds ($p < 0.01$ for all conditions), although this rejection is easily expected from the huge sample data in the simulation leading to detection of slight difference from normality. We assume, however, that these thresholds are almost normally distributed, at least for testing purpose, because the deviation of those skewness and kurtosis values from normal distribution was very small and they were stable for more than the critical trial numbers. Indeed, the rates of estimated thresholds falling in the 5% rejection regions defined under normality assumption were almost the

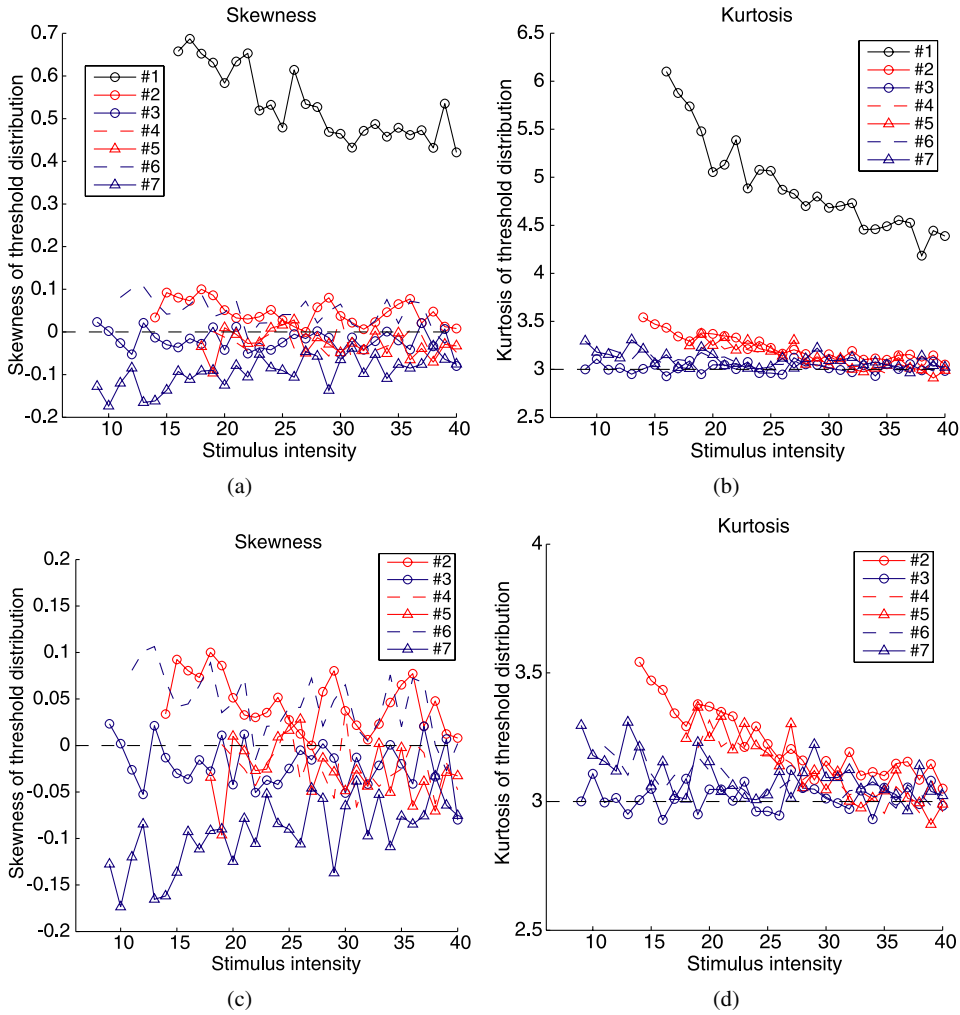

Figure 7. (a) Skewness and (b) kurtosis for trial numbers larger than the critical trial number (see text). (c) Skewness and (d) kurtosis except for the CSI #1 (data is identical to (a) and (b)). This figure is published in colour in the online version.

same as those falling in the 5% rejection regions based on actual threshold distributions themselves (data not shown).

The critical numbers for all CSIs are shown in Fig. 8. These critical numbers were smaller for #3, 6, 7, followed by #2, 4, 5. The critical number for #1 was the largest. In addition, the critical number was the smallest for #2 between #2, 4 and 5, and for #3 between #3, 6 and 7. These results also support the suggestion that stimulus intensity sampling, which is fine and symmetric around the threshold, is more desirable for threshold estimation with stable normality.

In summary, we conclude that the thresholds are almost normally distributed for testing purpose only if the trial number is sufficiently large and the stimulus

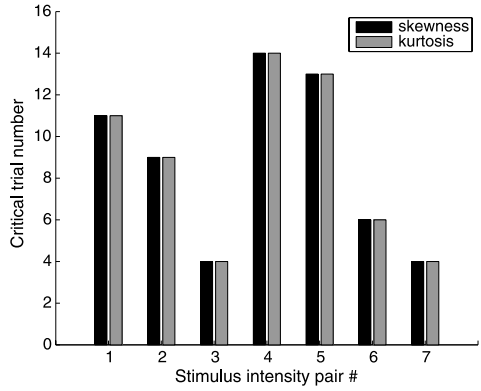

Figure 8. Critical trial numbers calculated for skewness and kurtosis.

intensity selection is appropriate. To achieve stable threshold normality, multiple stimulus intensities should be located around the intermediate 'yes' probabilities, and the trial numbers in each stimulus intensity should be more than 20, leading to typically about 140 trials in total. However, it should be noted that 'sufficient' trial numbers strongly depend on the CSIs and should be larger for alternative forced-choice design (McKee *et al.*, 1985). The proposed method can be applied for thresholds estimated with appropriate stimulus intensities and sufficient trial numbers.

4.2. Estimation of SEs of Thresholds

4.2.1. Method

We examined the accuracy of the SEs of thresholds estimated from the maximum likelihood method by analyzing the simulation results described in Section 4.1. The accuracy of SE estimation of thresholds is directly linked to the accuracy of statistical testing, even if the threshold normality assumption is valid. Although the accuracy of SE estimation has been described in some other studies (e.g., Foster and Bischof, 1991), we retested it for more conditions with more repetitions of the Monte Carlo simulations.

We defined the standard deviation of estimated thresholds (e.g., the standard deviation of threshold samples in distributions of Fig. 5) as an 'actual SE' for each trial number and CSI. Meanwhile, a standard deviation of a threshold estimated from a Fisher information matrix on each threshold estimation was defined as an 'estimated SE'. We calculated means and standard deviations of 'estimated SEs' from 10 000 estimations for each condition, and compared them with the actual SE.

4.2.2. Results

Figure 9 illustrates the real and estimated SEs as a function of the trial number for CSI #2. The estimated SEs approached the real SEs as trial numbers increased, as expected from asymptotic normality of maximum likelihood estimators. In addition, as demonstrated for skewness and kurtosis, the estimated SEs abruptly became

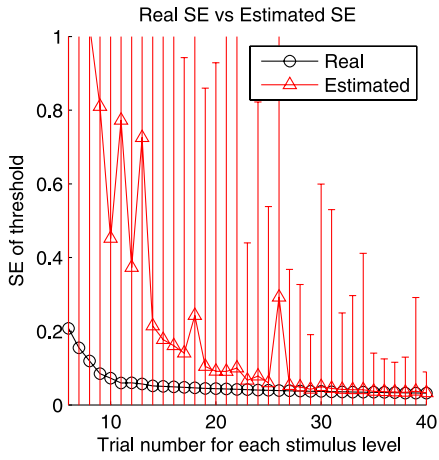

Figure 9. Real SE and estimated SE calculated under CSI #2 as a function of trial number for each stimulus intensity. Black and orange lines represent real and estimated SEs, respectively. Error bars represent standard deviations of estimated SEs. This figure is published in colour in the online version.

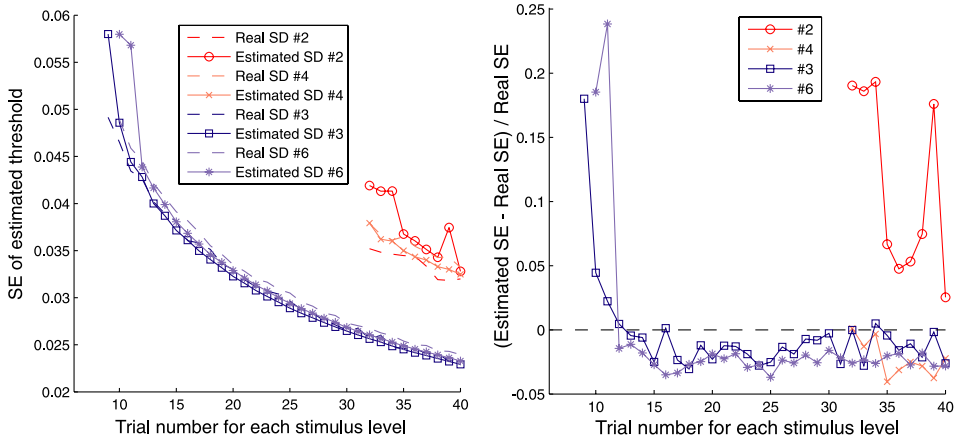

Figure 10. (a) Real and Estimated SE under CSIs #2, 3, 4 and 6 for trial numbers larger than critical number (see text). Dotted lines show real SEs, and solid lines show estimated SEs. Plot colors correspond to CSIs. (b) Difference ratio between real and estimated SEs ((estimated SE − real SE)/real SE). This figure is published in colour in the online version.

stable for trial numbers larger than a certain number. These tendencies were also observed in other CSIs. To investigate the number of trials required to yield stable SE estimations, we arbitrarily defined a 'critical trial number' also for estimated SEs; for all the trial numbers larger than the critical trial number, differences of estimated SEs from real SEs do not exceed 25% of real SEs.

The estimated SEs for trial numbers that exceed the critical trial numbers are shown in Fig. 10(a), and the difference ratios between the estimated and real SEs

((Estimated SE − Real SE)/Real SE) are shown in Fig. 10(b) for CSIs #2, 3, 4 and 6. Again, it is clear that the estimated SEs came closer to the real SEs as the trial number increased. However, for CSIs #2 and 5, the critical trial number and the difference between the real and estimated SEs were larger than that for CSIs #3 and 6. Moreover, we could not find a critical trial number for #1 within the tested trial numbers. These results suggest the necessity of appropriate stimulus intensity selection for usable SE estimation.

Under conditions with large trial numbers, however, the estimated SEs tended to be slightly less than the real SEs. Foster and Bischof (1991) have also reported this tendency. This bias is undesirable for statistical testing. However, the difference between the real and estimated SEs is not large (typically less than 5%). In addition, it was reported that even the bootstrap method could not obviate this estimation bias (rather, bias of bootstrap estimates of SEs tends to be larger than that of asymptotical estimation for large samples; Foster and Bischof, 1991).

Therefore, we consider that these SEs estimated from the maximum likelihood method are not perfectly ideal but acceptable as statistics for hypothesis testing if the trial number is sufficiently large and the stimulus intensity selection is appropriate (multiple stimulus intensities corresponding to intermediate 'yes' response probabilities exist). The effects of this SE estimation bias on hypothesis testing are also discussed in the next subsection.

4.3. Comparison of Proposed Method with t-Test and ANOVA

The previous subsections suggest that the normality assumption of thresholds and accuracy of SE estimations are acceptable for hypothesis testing in the proposed method. Meanwhile, differences between the multiple thresholds estimated from the constant stimuli method can also be tested with the t-test or ANOVA by repeatedly measuring the threshold. In this subsection, we compare the power and type I error of the proposed method with those of the t-test using Monte Carlo simulations.

4.3.1. Methods

We simulated two ideal observers whose psychometric functions were defined by logistic functions, and tested the significant difference between the thresholds of the two observers with the proposed method and the t-test. The parameters of the two psychometric functions were identical to those employed in Sections 4.1 and 4.2 except for values of θ_1; one was 1.45 and the other was 1.55. In particular, they were different only in their thresholds (horizontal positions). The seven kinds of CSIs employed were identical to those in Sections 4.1 and 4.2.

In the simulation, the thresholds for the two observers were estimated, and then the threshold difference was statistically tested. When using our proposed method, the number of the observer's responses at each stimulus parameter was 64, that is, 448 responses were used in total to estimate a threshold. This trial number is much more than the critical trial number of estimated SE for each of the CSIs except for #1, where we could not estimate a critical trial number. After measuring thresholds

for the two observers, the difference between these two thresholds was tested using our proposed method with a significance level of 0.05.

Meanwhile, for the t-test, the number of responses at each stimulus parameter was eight; that is, 56 responses were used in total to estimate a threshold. However, we measured eight thresholds for each condition by repeating this procedure. Therefore, the total number of observer's responses for eight threshold estimations was 448, the same as that obtained when using our proposed method. After estimating eight thresholds in each condition, the difference between the threshold means for the two conditions was tested with the t-test with a significance level of 0.05.

The test was repeated 10 000 times for each of the testing methods. The ratio of the number of 'rejection of null hypothesis (i.e., the threshold difference between the two conditions was significant)' conclusions to the total number of tests was defined as the power of the test. The power was calculated for each testing method.

4.3.2. Results and Discussion

The calculated powers are shown in Fig. 11(a). The powers vary with the stimulus parameter pair, as expected from the SEs of the thresholds estimated in Section 4.2. However, the power of the proposed method is consistently larger than that of the t-test for all the CSIs. This indicates that our proposed method may detect a significant difference between thresholds better than the t-test.

We performed an additional simulation to compare the type I errors between the two testing methods. The thresholds in the two psychometric functions were 1.5, and a ratio of the number of 'reject null hypothesis' conclusions to the total number of the tests was defined as a type I error in this simulation. The other simulation methods were identical to those of the power simulation. The derived type I errors are shown in Fig. 11(b). Although both the type I errors for the CSI #1 were much

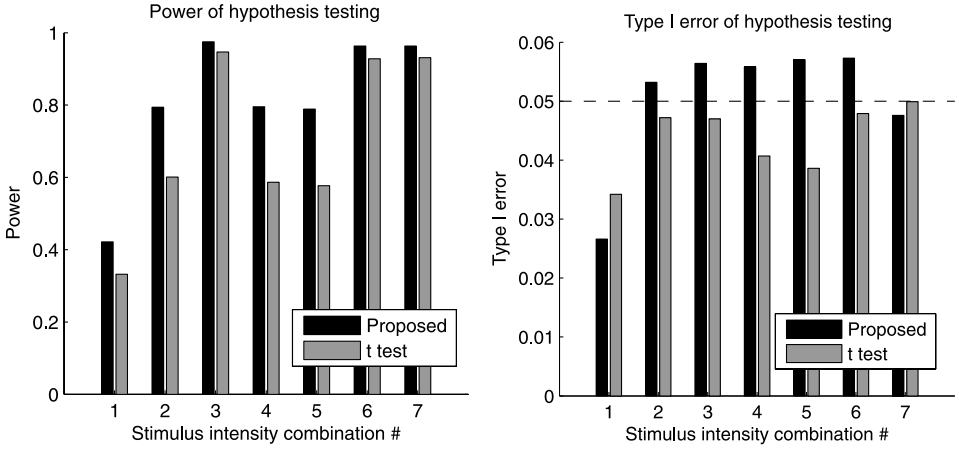

Figure 11. (a) Power and (b) Type I error of the t-test and the proposed method with significant level of 0.05. Black bars represent results of the proposed method, and gray bars represent results of the t-test.

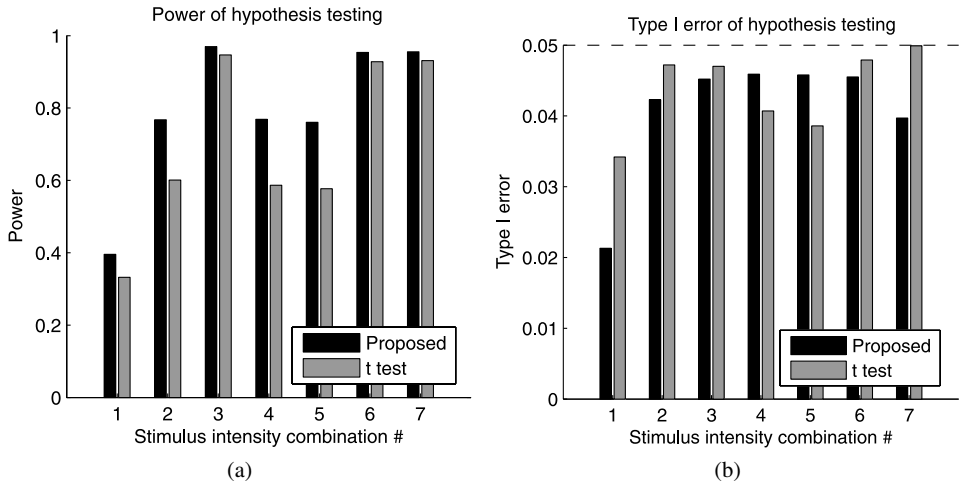

Figure 12. (a) Power and (b) Type I error of the t-test with significant level of 0.05 and of the proposed method with significant level of 0.04.

smaller than 0.05, the significance level in the simulation, those for the other CSIs, were nearly equal to 0.05. However, the type I errors of the proposed method tended to be slightly larger than 0.05. This can be easily expected owing to the fact that the estimated SEs of thresholds were slightly smaller than the real SEs under large trial numbers, as shown in Section 4.2. These results raise a possibility that the larger power of the proposed method originates just from its larger type I error.

To test this possibility, we conducted a similar simulation in which the significance level was decreased to 0.04 only for the proposed method. The powers and type I errors are shown in Fig. 12(a) and (b), respectively. In the figure, the type I errors are nearly equal in the two methods. Nevertheless, the powers were still larger for the proposed method as seen in Fig. 11(a). This dismisses the above possibility, but instead indicates that the proposed method is more efficient than the t-test for testing power. Thus, the proposed method should be more suitable than the t-test or ANOVA, because it can efficiently test the threshold differences for the cases where a sigmoid model can be fitted well to the observer's data and where the number of trials is sufficiently large.

In the above simulation, the sample number used for testing with the t-test was only eight. This small sample number might cause the efficiency of the t-test to be less than that of the proposed method. Therefore, we again performed a similar simulation for the CSI #1, except that the trial number for each stimulus intensity was either 32, 64, 128 or 256, corresponding to 4, 8, 16 or 32 threshold samples for the t-test, respectively, when eight trials for each stimulus intensity were used for each threshold estimation. In the results (data not shown), the power of the t-test increased with the increase in the number of threshold samples as expected. However, the power of our proposed method also abruptly increased with the increase in trial numbers, and rather the difference in power between the two methods in-

creased with an increase in trial numbers. In addition, the power of the t-test was higher for less threshold samples by increasing a trial number for each threshold estimation without changing the total trial number (e.g., for a total trial number of 256 for each stimulus intensity, trial numbers for each threshold estimation can be increased from eight to 16 or 32, leading to numbers of threshold samples from 32 to 16 or 8), but still remained smaller than the proposed method.

Similarly, the yes/no situations in our simulation might tend to benefit the proposed method compared to two-alternative forced choice (2AFC) situations, because rise in lower asymptote of psychometric functions (e.g., 0.5 in 2AFC situations) tends to yield inefficiency of maximum likelihood estimators in psychometric functions (McKee *et al.*, 1985). We again performed simulations with ideal observers whose lower asymptote of psychometric functions was 0.5 instead of 0. In the results (data not shown), the power for asymptote 0.5 was lower than that for 0, as expected for both the two testing methods. However, even in this situation, the power of the proposed method was larger than the t-test in all the CSIs, while type I error was similar to that observed in Fig. 11. The results of these additional simulations suggest that the proposed method is superior to the t-test in their power in a wide range of conditions in stimulus intensities or experimental methods.

Although we tested only the difference between two thresholds here, we would derive similar results if we tested the difference between more than two thresholds using the proposed methods and the ANOVA. As stated above, some psychophysical studies tested the difference between multiple thresholds derived from the constant stimuli method with the t-test or ANOVA by considering each observer's threshold as a sample. However, a sigmoid model should very well fit the data collected from different observers in many cases, unless the observer's psychometric functions are not largely different. In such cases, our method may detect the significance of threshold difference more efficiently than the t-test or ANOVA.

Note that the type I errors of the proposed method were slightly larger than the significant level. This must be because of the small estimation of SEs of thresholds. This property is undesirable as a statistical testing method. However, as stated in Section 4.2, Foster and Bischof (1991) showed that the SEs estimated from the bootstrap method were less than those estimated from the probit analysis when the trial number was sufficiently large. Their simulation results suggest that the bootstrap estimates of SEs cannot improve type I error for large trial numbers, but may possibly worsen it. Again, we argue that if you utilize statistics of psychometric functions for statistical testing, it is better to use maximum likelihood estimates only when the trial number is sufficiently large. In any case, it should be noted that the results should not be conservative but rather slightly 'liberal'.

4.4. Comparison of the Proposed Method with the YKF Method

The previous subsection demonstrated that our proposed method is superior to the t-test with regards to power, especially when stimulus intensity selection is appropriate (this leads to an efficient estimation of SEs from the maximum likelihood

method). As described in the Introduction, Yssaad-Fesselier and Knoblauch (2006) proposed a novel method to evaluate threshold differences derived from psychometric functions. In this section, we compare power and type I error between the proposed method and the YFK method.

4.4.1. Methods

The simulation procedures were identical to those in Section 4.3, except that the t-test was replaced by the YFK method. The YFK method was performed with the PAL_PFLR_ModelComparison function in Palamedes Toolbox (Prins and Kingdom, 2009). We again evaluated the efficiency in testing differences between the two thresholds derived from two different ideal observers (the thresholds were 1.45 and 1.55 for power comparison, and both were 1.5 for type I error comparison). In the PAL_PFLR_ModelComparison function, the two sets of observers' responses were fitted with two different models of psychometric functions. In one model, the two psychometric functions had respective threshold parameters, and in the other they had only one common threshold parameter, corresponding to the null hypothesis. After fitting the functions, likelihood ratio between these two models were calculated. Because this likelihood ratio is chi-squared with one degree of freedom, the p-value was calculated from the likelihood ratio based on the chi-square distribution. Simulations were performed both in the yes/no situations (lower asymptote of the psychometric functions was 0) and the 2AFC situations (lower asymptote was 0.5).

4.4.2. Results and Discussion

The simulated power and type I error for both the proposed method and the YFK method in the yes/no situations are shown in Fig. 13. It is clear that these two meth-

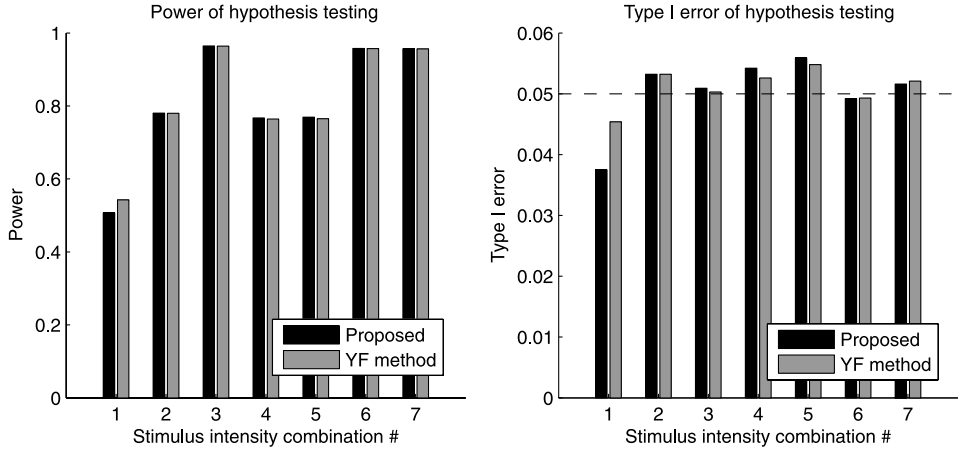

Figure 13. (a) Power and (b) Type I error of the YKF method and the proposed method with significant level of 0.05. Black bars represent results of the proposed method, and gray bars represent results of the YKF method.

ods yielded almost equal power and type I error. These results suggest that these two methods are substantially identical in their efficiency when appropriate stimulus intensities and trial numbers are supplied. Only for the CSI #1, power is a little higher for the YFK method than that for the proposed method. This may be attributed to unstable estimation of SEs as expected from the results of the Section 4.2, where we could not find a critical trial number for SE estimation in the stimulus intensity combination #1; our method is based on the Delta method, which is, in turn, based on asymptotic normality of maximum likelihood estimators, and therefore, the Delta method may accentuate the effects of the inefficient estimated SEs by combining multiple SEs. However, the power of the YKF method for #1 was also small, and power difference from the proposed method was not sufficiently large.

The results for the 2AFC situations (data not shown) were quite similar to the yes/no situations; the power and type I error were almost equal for the two methods except for the CSI #1, although general efficiency was lower than that for the yes/no situations. We assume that testing results of the two methods for different stimulus conditions behave similarly, because both of them rely on the asymptotic theory of the maximum likelihood method in the simulation. Therefore, again it should be very important to adopt appropriate stimulus intensities and trial number when applying these methods. It should be noted that in the YKF method, as implemented in the Palamedes Toolbox, the bootstrap re-sampling of threshold difference parameters or likelihood ratios may help increase its efficiency to some extent when trial numbers for threshold estimation are small.

We conclude that these two methods, our proposed method and the YKF method, are substantially equal in their efficiency when appropriate stimulus intensities and trial numbers are adopted in experiments with constant stimuli methods, although the YKF method can be a little more efficient than our proposed method when accuracy of SE estimation is not sufficient.

5. General Discussion

We proposed a new statistical significance testing method to assess differences between the thresholds estimated from the constant stimuli method using the Delta method and the Mahalanobis squared distance. Although thresholds can be accurately estimated using the constant stimuli method, the differences of the thresholds have been often assessed by less efficient statistical testing methods (see Section 1). The statistical testing method proposed in this paper can efficiently assess differences of thresholds estimated from the constant stimuli method, because it utilizes all the data used for the maximum likelihood estimations, though the thresholds must be assumed to be normally distributed. The proposed method can be utilized in a similar way as the t-test and the ANOVA for the adjustment method. In addition, it also has other advantages; for example, it can assess effects of experimental condition differences on thresholds in individual results, it does not require the assumption of equality of variance, and its testing procedure is simple. Moreover, the

proposed method can be easily modified and applied to the difference of psychometric function slopes between the different experimental conditions by evaluating the slope parameters of a psychometric function model with similar procedures (although not described in detail here).

However, note that our method requires appropriate stimulus intensity selection and sufficient trial numbers as expected from asymptotic normality of maximum likelihood estimators, considering the normality of estimated thresholds and accuracy of SE estimation of thresholds. Normality assumption can be achieved with a relatively small number of trials; 150 trials can be sufficient (Fig. 7, except for the CSI #1). However, many more trials are required for accurate SE estimation. The required trial number critically depends on the selection of the stimulus intensity; if only one stimulus intensity were included within the intermediate (10–90%) 'yes' probabilities, many trial numbers may be required, whereas, if several (5–7) stimulus intensities were included in intermediate 'yes' probabilities, 100 trials may be sufficient from our simulations. Irrespective of the number of intensities, fewer trial numbers tend to yield larger estimated SEs, leading to less efficient results of the proposed method (but still more efficient than the t-test). Although this lower efficiency leads to a smaller type I error and testing power, this does not eliminate the eligibility of the proposed method. Ignoring this inefficiency, only threshold normality should be concerned for eligibility of the proposed method.

There are other possible methods to test the statistical significant differences between the thresholds estimated by the constant stimuli method, such as applying the t-test or ANOVA by measuring multiple thresholds. One can estimate multiple thresholds through the constant stimuli method by estimating a threshold, for example, from the data of each experimental session if there are multiple sessions. In this case, the t-test and ANOVA could be used to assess the significant difference between thresholds of different conditions within an observer. However, the power of this method is less than that of our proposed method for most of the stimulus intensity and trial number conditions, as indicated in our simulations. Furthermore, it will require many trials to derive multiple thresholds from the constant stimuli method. Therefore, it is more appropriate and efficient to use our method rather than estimating multiple thresholds from the constant stimuli method and then applying the t-test or ANOVA. If you prefer the t-test or ANOVA, it may be more efficient to measure thresholds from other psychophysical procedures, for example, adaptive staircase procedures such as the QUEST (Watson and Pelli, 1983).

Recently, the bootstrap method has received a lot of attention for evaluating variances of the threshold estimations (Efron, 1982; Wichmann and Hill, 2001a, b). It has been used in many psychophysical studies (Allen *et al.*, 2003; Carlson *et al.*, 2006; Khang *et al.*, 2003) for the construction of confidence intervals for thresholds. The procedure for the bootstrap method is very simple, and it is very powerful for estimating confidence intervals of a variety of statistical values, such as the mean and median, without calculating complex parameters of statistical distributions. Because an assumption of a statistical distribution is not required for the bootstrap

method, this method is more appropriate than our proposed method when the number of trials is small and the threshold normality assumption is implausible. It has been indicated that estimation of threshold variances from the bootstrap method is more robust than that from the asymptotic normality assumption for small data sets (Foster and Bischof, 1987, 1991). The bootstrap method is very compelling for its robustness against the number of experimental samples and its flexibility is hardly affected by the assumption of certain statistical distributions. However, standard error estimation by the bootstrap method is not always desirable for our method. Variance estimation from the bootstrap method is less accurate than that from the asymptotic normality analysis with large sample sizes, such as 300–900 trials for each threshold (Foster and Bischof, 1991). Therefore, it is more efficient to estimate the standard error of the threshold on the basis of asymptotic normality as in our proposed method, rather than using the bootstrap method, if the data amount is sufficient for the asymptotic normality assumption.

The YKF method is another valuable method for testing threshold differences. In the YKF method, parameters corresponding to threshold differences between different experimental conditions are included in the model of psychometric functions. This approach makes it possible to test not only threshold differences, but also other different features in psychometric function models by defining parameters of interest in the models. The efficiency of our method and the YKF method is substantially equal in most of the stimulus intensity selections and trial numbers, as demonstrated in our simulation (Section 4.4). Thus, one can choose either of these two methods for testing threshold differences in practical psychophysical experiments, because both these methods should yield substantially equal results (e.g., the p-values calculated from the two methods for the data in Table 1 is almost equal). However, the YKF method can be a little more efficient than our method when SE estimations are less trustworthy, although in this case the efficiencies are less than satisfactory for both methods. Therefore, only for thresholds with small trial numbers, it can be desirable to use the YKF method and apply bootstrap re-sampling to evaluate variance of the parameters used for testing.

In conclusion, our method can be used for statistical significance testing of the difference between the thresholds estimated from the constant stimuli method. In addition, our method has larger power than the t-test or ANOVA in which each threshold is considered just as a sample, and substantially equal power as the YKF method for appropriate stimulus selections and sufficient trial numbers. Although large trial numbers and appropriate stimulus intensity selection are necessary when applying our method, it should be a very useful tool to analyze and interpret experimental results from the constant stimuli method.

Acknowledgements

We thank Kazuo Shigemasu for his helpful comments regarding the logistic function and the maximum likelihood method, and Donald I. A. MacLeod, Michael

Falconbridge, David H. Foster, and two reviewers for their valuable suggestions and corrections on this paper.

References

Allen, H. A., Hess, R. F., Mansouri, B. and Dakin, S. C. (2003). Integration of first- and second-order orientation, *J. Optic. Soc. Amer. A* **20**, 974–986.

Carlson, T. A., Schrater, P. and He, S. (2006). Floating square illusion: perceptual uncoupling of static and dynamic objects in motion, *J. Vision* **6**, 132–144.

Edwards, L. K. (Ed.) (1993). *Applied Analysis of Variance in Behavioral Science*. Marcel Dekker, New York, NY, USA.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia, PA, USA.

Foster, D. H. and Bischof, W. F. (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance function, *Biol. Cybern.* **57**, 341–347.

Foster, D. H. and Bischof, W. F. (1991). Thresholds from psychometric functions — superiority of bootstrap to incremental and probit variance estimators, *Psychol. Bull.* **109**, 152–159.

Garcia-Perez, M. A. and Alcala-Quintana, R. (2005). Sampling plans for fitting the psychometric function, *Spanish J. Psychol.* **8**, 256–289.

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Stat.* **6**, 65–70.

Khang, B. G., Koenderink, J. J. and Kappers, A. M. L. (2003). Perception of surface reflectance of 3-D geometrical shapes: influence of the lighting mode, *Perception* **32**, 1311–1324.

Kingdom, F. A. A. and Kasrai, R. (2006). Colour unmasks dark targets in complex displays, *Vision Res.* **46**, 814–822.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York, NY, USA.

McKee, S. P., Klein, S. A. and Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions — implications of Probit analysis, *Percept. Psychophys.* **37**, 286–298.

Nagy, A. L., Neriani, K. E. and Young, T. L. (2005). Effects of target and distractor heterogeneity on search for a color target, *Vision Res.* **45**, 1885–1899.

Oehlert, G. W. (1992). A note on the delta method, *Amer. Stat.* **46**, 27–29.

Prins, N. and Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data, available at: http://www.palamedestoolbox.org.

te Pas, S. F. and Koenderink, J. J. (2004). Visual discrimination of spectral distributions, *Perception* **33**, 1483–1497.

Pratt, J. W. (1976). F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation, *Ann. Stat.* **4**, 501–514.

Watson, A. B. and Pelli, D. G. (1983). Quest — a Bayesian adaptive psychometric method, *Percept. Psychophys.* **33**, 113–120.

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

Wichmann, F. A. and Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling, *Percept. Psychophys.* **63**, 1314–1329.

Yssaad-Fesselier, R. and Knoblauch, K. (2006). Modeling psychometric functions in R, *Behav. Res. Methods* **38**, 28–41.

Appendix

A.1. Matlab scripts and Excel macro

We created Matlab scripts and Excel macros to perform the statistical testing proposed in this paper. Those files can be downloaded at http://www.uchikawa.ip.titech.ac.jp/pgs.html.

They perform statistical significance tests for thresholds measured in a one- or a two-factor experiment (these tests are similar to one-way and two-way ANOVAs), and multiple comparison for thresholds measured in a one-factor experiment.

Estimations of thresholds and their SEs by a maximum likelihood method are required before testing the files, because they use thresholds and their SEs as their inputs. The explanations about how to use the files may also be downloaded from the same website.