The Spanish Journal of Psychology
2009, Vol. 12, No. 1, 3-11


ISSN 1138-7416

# Empirical Performance of Optimal Bayesian Adaptive Estimation

Miguel Ángel García-Pérez and Rocío Alcalá-Quintana
Universidad Complutense (Spain)

Simulation studies have shown how Bayesian adaptive estimation methods should be set up for optimal performance. We assessed the extent to which these results hold up for human observers, who are more subject to failure than simulation subjects. Discrimination and detection experiments with two-alternative forced-choice (2AFC) tasks were used for that purpose. Forty estimates of the point of subjective equality (PSE, or the 50% correct point on the psychometric function for discrimination) and 32 estimates of detection threshold (the 80% correct point on the psychometric function for detection) were taken for each of four observers with the optimal Bayesian method, while data for fitting the psychometric function $\Psi$ were gathered concurrently with an adaptive method of constant stimuli governed by fixed-step-size staircases. The estimated parameters of the psychometric function served as a criterion for comparison. In the discrimination task, PSEs for each observer were distributed around the independently estimated 50% correct point on $\Psi$ and their variability was occasionally minimally larger than simulation results indicated it should be. In the detection task, the distribution of threshold estimates was consistently above the independently estimated 80% correct point on $\Psi$ and their variability was as expected from simulations. A close analysis of these results suggests that the optimal Bayesian method is affected by growing inattention or fatigue in detection tasks (factors that are not considered in simulations), and limits the practical applicability of Bayesian estimation of detection thresholds.
*Keywords: bayesian adaptive methods, sequential procedures, psychophysics, threshold, point of subjective equality, 2AFC tasks*

Los métodos bayesianos de estimación adaptativa han sido optimizados en varios estudios de simulación. En este trabajo evaluamos hasta qué punto los resultados obtenidos en las simulaciones son aplicables a observadores humanos. Para ello se sometió a cuatro observadores a dos tipos de experimento (discriminación y detección) con la tarea de elección forzada entre dos alternativas (2AFC). La configuración óptima del método bayesiano sirvió para obtener, por cada observador, 40 estimaciones del punto de igualdad subjetiva (PSE, que es el punto de la función psicométrica que lleva aparejado un porcentaje de éxito del 50% en un experimento de discriminación) y 32 estimaciones del umbral de detección, definido como el punto de la función psicométrica cuyo porcentaje de éxito asociado es el 80%. Simultáneamente, se utilizó el método adaptativo de los estímulos constantes para obtener una estimación independiente de los parámetros la función psicométrica $\Psi$ de cada observador que sirviera como criterio de comparación. En la tarea de discriminación, y para todos los observadores, las distribuciones de los PSE se situaron en torno a los puntos del 50% de $\Psi$ estimados de manera independiente y la variabilidad fue sólo ligeramente superior a la esperada a partir de las simulaciones. Por el contrario, en la tarea de detección, las distribuciones de estimaciones del umbral se situaron consistentemente por encima de los puntos del 80% de $\Psi$, aunque su variabilidad fue similar a la registrada en las simulaciones. Un análisis minucioso de estos resultados sugiere que el método bayesiano óptimo se ve muy afectado por la creciente falta de atención y la fatiga en las tareas de detección (factores que no fueron contemplados en las simulaciones), lo que limita la aplicabilidad de los métodos bayesianos en la estimación práctica de umbrales de detección.
*Palabras clave: métodos adaptativos, procedimientos Bayesianos, escaleras de paso fijo, estimación de umbrales, elección forzada entre dos alternativas*

---

This research was partly supported by grants BSO2001-1685 from Ministerio de Ciencia y Tecnología and SEJ2005-00485 from Ministerio de Educación y Ciencia

Correspondence concerning this article should be addressed to Rocío Alcalá-Quintana, Departamento de Metodología, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 Madrid (Spain). Phone: 34-913-943061. Fax: 34-913-943189. E-mail: ralcala@psi.ucm.es

How to cite the authors of this article: García-Pérez, M.A., Alcalá-Quintana, R.

Estimating detection and discrimination thresholds or points of subjective equality is one of the basic activities in psychophysical research. Recent simulation studies (Alcalá-Quintana & García-Pérez, 2004a, 2005, 2007; García-Pérez & Alcalá-Quintana, 2007) have demonstrated that these parameters can be efficiently and accurately estimated with a suitably configured Bayesian adaptive method, which we will refer to as O-BEST (Optimal Bayesian Estimation by Sequential Testing). However, statistical properties determined through simulation have been shown to not always hold up in actual practice, undoubtedly because of discrepancies between the idealized response models used in simulations and the actual processes that take place in psychophysical research with human observers (see, e.g., Alcalá-Quintana & García-Pérez, 2007; Green, 1990, 1993; Kollmeier, Gilkey, & Sieben, 1988; Madigan & Williams, 1987; Simpson, 1989; Stillman, 1989).[1] Particularly for the case of O-BEST, Alcalá-Quintana and García-Pérez (2007) reported that the property of consistency that is observed in simulations—whereby bias disappears and variance decreases as the number of trials increases—is not observed in actual practice: As the number of trials increases, the mean of threshold estimates drifts linearly and consistently without traces of convergence, and the variance of threshold estimates stabilizes without traces of further reduction.

Given this state of affairs, practical use of O-BEST is not justified unless empirical evidence is found to the effect that it has adequate psychometric properties. In particular, it should be proved that O-BEST estimates are unbiased and that their variance (which determines the efficiency of the procedure) is related to the spread of the underlying psychometric function as simulation results indicate it should. It must be borne in mind that the simulation studies referred to in the above paragraph determined the psychometric properties of O-BEST estimates using the true parameters as referents, but those true parameters are not available in an empirical study involving human observers. Then, checking out the properties of O-BEST estimates in empirical research with human observers requires obtaining independent and dependable estimates of the entire psychometric function involved, whose parameters are necessary for this evaluation.

This paper presents the results of an empirical test of the theoretical performance of O-BEST both in 2AFC detection tasks (for estimation of the detection threshold) and in 2AFC discrimination tasks (for estimation of the point of subjective equality, or PSE). The psychometric properties of O-BEST are determined by repeated application of the procedure to the same observers, whereas the criterion parameters for comparison are obtained by fitting psychometric functions to data gathered concurrently with an adaptive method of constant stimuli (which we will refer to as AMOCS) optimally configured as described by García-Pérez and Alcalá-Quintana (2005). The data were used to test predictions based on simulations analogous to those described in Alcalá-Quintana and García-Pérez (2004a, 2005) but adjusted to the conditions in which the empirical data were gathered in this study and using true parameter values identical to the AMOCS estimates for each participant. Our results indicate that O-BEST performs as expected in 2AFC discrimination tasks, but that its performance does not match expectations in 2AFC detection tasks. Speculations on the reasons for this differential behavior are given in the Discussion section.

Some of these results have been presented in abstract form (Alcalá-Quintana & García-Pérez, 2004b).

## Method

### *Apparatus and Stimuli*

All experiments were controlled by a PC equipped with *VisionWorks* (Swift, Panish, & Hippensteel, 1997). Stimuli were displayed on a 20-inch Clinton Monoray (Richardson Electronics Ltd., LaFox, IL) monochrome monitor (model M20ECD5RE, DP104 phosphor) with a spatial resolution of 1024 × 600 pixels (horizontal × vertical), a luminance resolution of $2^{15}$ gray levels, and a frame rate of 122 Hz. The voltage-to-luminance non-linearity was compensated for via look-up tables arising from a calibration procedure that rendered a correlation of 0.999986 between actual and nominal luminance.

In the detection experiment, the target stimulus was a Gabor patch with a vertical carrier of 1 c/deg and a circular Gaussian envelope with a standard deviation of 2 deg. The stimulus was displayed with a mean luminance of 157 cd/m$^2$ that blended in with a uniform 157-cd/m$^2$ background covering the entire image area. The center of the monitor displayed a small cross that the observers fixated throughout the experiment. In each trial, the target always appeared centered on the fixation cross (which was not extinguished during stimulus presentation) and with a contrast level dictated by the applicable psychophysical procedure to be described below. The temporal course of stimulus presentation was a Gaussian pulse with a standard deviation of 100 ms.

The discrimination experiment included two conditions. In one of them, the target was as described in the preceding paragraph and the standard was analogous except that its carrier was horizontally oriented; in the other, the orientations

[1] It should also be noted that simulation results have been confirmed in some other empirical tests (see, e.g., García-Pérez, 2000; Laming & Marsh, 1988; Lesmes, Jeon, Lu, & Dosher, 2006; Schlauch & Rose, 1990).

of target and standard were swapped. In either case, the standard was always displayed with a fixed Michelson contrast $m = 0.1$ (i.e., a log contrast $x = -1$ log units). The temporal course of stimulus presentation was a rectangular pulse of ~131 ms (16 video frames).

### *Configuration of O-BEST*

The detection threshold was determined with the optimal configuration of the Bayesian procedure determined for 2AFC detection tasks by Alcalá-Quintana and García-Pérez (2004a, 2005; see also García-Pérez & Alcalá-Quintana, 2007), namely, a uniform prior on the interval [–3, 0], the prior mean as a placement rule, the posterior mean as final estimate, a fixed number of 70 trials, and a logistic model function with a spread of 1.81 units, a lower asymptote at $p = 0.5$ and an upper asymptote at $p = 0.96$ (thus implying a finger-error parameter $\lambda = 0.04$; see Equation 1 below). The procedure was set up to track the 80% correct point on the underlying psychometric function.

The PSE was analogously determined except that, owing to the peculiarities of 2AFC discrimination tasks and given a standard at $x = -1$ log units, the uniform prior was defined on the interval [–2, 0], the fixed number of trials was reduced to 30, and the logistic model function had a spread of 1 unit, a lower asymptote at $p = 0.06$ and an upper asymptote at $p = 0.94$ (thus implying a finger-error parameter $\lambda^* = 0.06$; see Equation 2 below. The procedure was logically set up to track the 50% correct point on the underlying psychometric function.

### *Configuration of AMOCS*

The psychometric function for detection was estimated from data gathered with one of the procedures recommended by García-Pérez and Alcalá-Quintana (2005), namely, adaptive fixed-step-size staircases involving the 1–up/1–down rule and using a step up of 0.6 log units and a step down of 0.2 log units. Each staircase proceeded until 30 reversals had occurred. Four staircases used a starting level of –1.6 log units and another four used a starting level of –1.5 log units.

The psychometric function for discrimination was analogously estimated except that the adaptive staircases used steps up and down of 0.1 log units, proceeded until 11 reversals had occurred, and their starting levels were either –1 log units (for ten staircases) or –0.95 log units (for another ten).

### *Experimental Procedure*

The monitor was allowed to warm up for no less than half an hour before any session started. Binocular viewing with natural accommodation and pupils was used. Observers sat 75 cm away from the display and their head was not restrained although they were asked to maintain a fixed viewing distance throughout the experiment. The room was dark except for the light from the display monitor. The background luminance and the fixation cross were present throughout the experimental session.

All data were gathered with a temporal 2AFC paradigm. A temporal 2AFC trial consisted of two presentations in only one of which was the target displayed (newly decided with equiprobability on each trial), whereas the other interval displayed mean luminance (in detection experiments) or the standard stimulus (in discrimination experiments). The two intervals were marked by beeps of different pitch and were separated by gaps of ~115 ms (14 frames) in the detection experiment and ~615 ms (75 frames) in the discrimination experiment. The observer's task was to indicate by a key press either the interval in which the target had been presented (in detection experiments) or the interval in which the stimulus had higher contrast (in discrimination experiments). If both intervals appeared to have displayed a stimulus with the same contrast (or a blank), the observer was asked to guess at random. If observers missed a trial for whatever reason, they could use a third key to ask for the trial to be discarded and repeated (not necessarily immediately afterwards). The session was self-paced, as the next trial did not start until the observer had responded. Error feedback was not provided.

Detection data were collected in two repeat sessions, each consisting of 12 blocks of trials (8 blocks governed by O-BEST and 4 blocks governed by AMOCS, administered to each observer in a newly decided random order). Each O-BEST block randomly interweaved two identical 70-trial runs; each AMOCS block randomly interweaved two staircases that differed only as to their starting point, as described above. Thus, O-BEST blocks consisted of exactly 140 trials whereas AMOCS blocks (whose staircases were not set to finish after a fixed number of trials but after 30 reversals) varied between 126 and 169 trials. This design thus yielded, for each observer, 32 separate estimates of detection threshold obtained with O-BEST and a separate data set consisting of 1112–1214 trials (from eight 30-reversal staircases) for fitting the psychometric function for detection.

Discrimination data were similarly collected in five repeat sessions, each consisting of 10 blocks of trials (8 blocks governed by O-BEST and 2 blocks governed by AMOCS, administered to each observer in a newly decided random order). Each O-BEST block randomly interweaved two 30-trial runs, one for the horizontally-oriented standard and one for the vertically-oriented standard; each AMOCS block randomly interweaved two staircases that differed as to their starting point (as described above) for each of the conditions (standard stimulus oriented vertically or horizontally). Thus, O-BEST blocks consisted of exactly 60 trials whereas AMOCS blocks (comprising four 11-reversal staircases) varied between 59 and 98 trials. This design thus yielded, for each

observer and condition, 40 separate estimates of the PSE obtained with O-BEST and a separate data set consisting of 327–417 trials (from 20 11-reversal staircases) for fitting the psychometric function for discrimination.

### *Fitting the Psychometric Function*

Data from all repeat AMOCS sessions in a condition (detection or discrimination with given standard and test orientations) were pooled and binned by contrast level to fit psychometric functions. In either case the analysis was carried out separately for each observer.

For detection data, the fitted psychometric function had the logistic form

$$\Psi(x) = \frac{1}{2} + \frac{1/2 - \lambda}{1 + \exp[-(x-\eta)/\beta]}, \quad (1)$$

and estimates of $\lambda$, $\eta$, and $\beta$ were obtained with maximum-likelihood methods using NAG subroutine E04JYF (Numerical Algorithms Group, 1999), which allows constrained optimization. We imposed the natural constraints $\hat{\beta} > 0$ and $\hat{\eta} < 0$ and, following the recommendations of Wichmann and Hill (2001) regarding the upper asymptote, we also constrained $0 \leq \hat{\lambda} \leq 0.06$. Discrimination data were fitted by the alternative logistic function

$$\Psi^*(x) = \lambda^* + \frac{1 - 2\,\lambda^*}{1 + \exp[-(x-\eta^*)/\beta^*]}, \quad (2)$$

because the lower and upper asymptotes are both determined by the finger-error parameter $\lambda^*$ in 2AFC discrimination experiments. Maximum-likelihood estimates of $\lambda^*$, $\eta^*$, and $\beta^*$ were similarly obtained.

In either case, the location parameter $\eta$ (or $\eta^*$) was transformed into a threshold (or PSE) parameter $\theta$ (or $\theta^*$) that satisfies the applicable definition (i.e., the stimulus level at which the probability of success in the task is $\pi$) through (see García-Pérez & Alcalá-Quintana, 2005)

$$\hat{\theta} = \hat{\eta} + \hat{\beta}\ \ln\left[\frac{\pi - 1/2}{1 - \hat{\lambda} - \pi}\right], \quad (3)$$

$$\hat{\theta}^* = \hat{\eta}^* + \hat{\beta}^*\ \ln\left[\frac{\pi - \hat{\lambda}^*}{1 - \hat{\lambda}^* - \pi}\right], \quad (4)$$

where $\pi = 0.8$ in case of detection and $\pi = 0.5$ in case of discrimination. Similarly, the spread $\sigma$ (or $\sigma^*$) of the psychometric function was computed through (see García-Pérez & Alcalá-Quintana, 2005)

$$\hat{\sigma} = 2\,\hat{\beta}\ln 99, \quad (5)$$

$$\hat{\sigma}^* = 2\,\hat{\beta}^*\ln 99. \quad (6)$$

### *Observers*

Six observers with normal or corrected-to-normal vision were recruited, but only two of them (the authors, who were aware of the design and goals of the study) took part in all experiments. The remaining four observers were naïve to the purpose of the study and were not familiar with psychophysical experiments.

## Results

### *Detection*

Each panel in Figure 1 shows results from AMOCS and O-BEST for one of the participants in the detection experiment. Perhaps the most salient outcome is that the (within-subject) average O-BEST estimate of threshold is systematically higher than the estimate of threshold obtained from AMOCS data (compare the location of the short vertical segment through the triangles above each panel with the location of the vertical line in the panel), and that the latter is not contained in the 95% confidence interval calculated from the former (represented by the width of the vertical gray bar in each panel). The confidence interval should strictly be compared with true $\theta$ and not with an estimate thereof, but it is nevertheless striking that the displacement occurs systematically for all observers.

With regard to the relation of the standard deviation of O-BEST estimates to the spread of the psychometric function, results reported in Figure 1 imply ratios between 0.072 and 0.118, which are slightly higher than the figure 0.05–0.06 arising from 70-trial simulation runs (see Alcalá-Quintana & García-Pérez, 2005). As noted in Introduction, it should be recalled that the latter figure involves a comparison with the actual spread of the psychometric function, whereas the figures reported here arise from a comparison with an estimate of this spread (which is, hence, affected by sampling error also). To evaluate whether this procedural difference might explain the higher ratios that were observed empirically (and, also, the discrepancies between average O-BEST estimates and AMOCS estimates), we ran simulations thoroughly analogous to those in Alcalá-Quintana and García-Pérez (2005), with the only difference that (1) simulated O-BEST and AMOCS sessions were designed to match exactly the empirical sessions in these experiments; (2) for each of four simulated observers, the true parameters of their psychometric functions were taken to be the AMOCS parameter estimates of its reference empirical observer; (3) data were subjected to the same analyses described here; and (4) the ratio of the standard deviation of O-BEST estimates to the estimated spread of the psychometric function was computed for each replicate. The results of

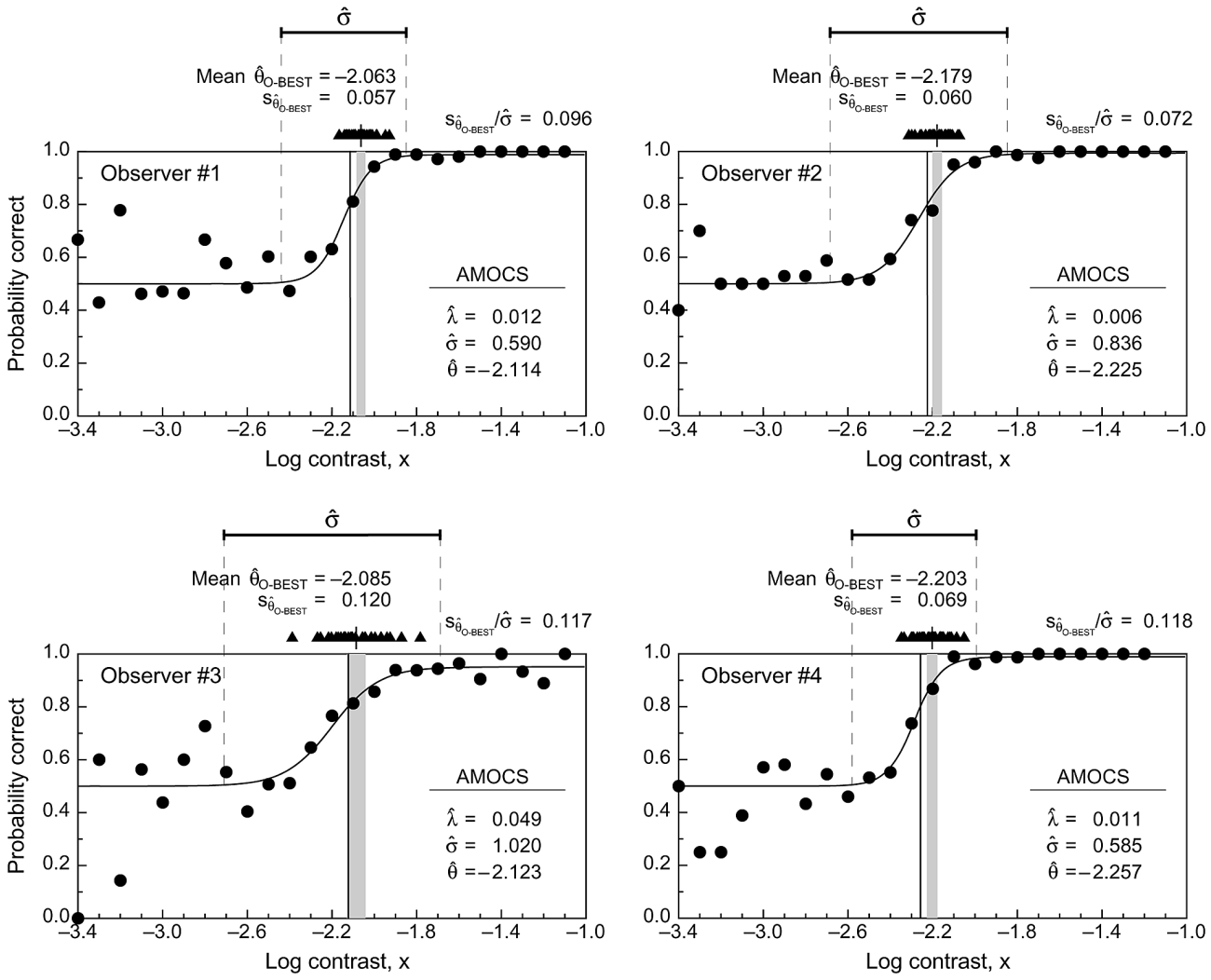


*Figure 1.* Results of the detection experiment for each observer. Binned AMOCS data are shown within the panel and indicate percentage correct in the 2AFC task at each stimulus level. The sigmoidal curve through the data is the best-fitting logistic function in Equation 1 with estimated (or derived through Equations 3 and 5) parameters shown in the inset. The location of threshold is indicated by a vertical line through the panel at an abscissa $x = \hat{\theta}$, and the estimated spread $\hat{\sigma}$ of the psychometric function is indicated as a segment at the far top of the panel. The 32 separate O-BEST estimates of threshold are indicated by solid triangles immediately above each panel, with their mean indicated by a short vertical segment. Values for the mean and standard deviation of these estimates are also printed above the symbols. The ratio of this standard deviation to the estimated spread of the psychometric function is printed outside the top-right corner of each panel. The width of the vertical grey bar in each panel gives the 95% confidence interval for the location of threshold, obtained from O-BEST data.

these simulations (see Figure 2) show that the average O-BEST estimate virtually matches the AMOCS estimate of threshold on a subject by subject basis (a feature not observed in our empirical results), and that the central 95% range of estimates of threshold from either procedure is almost identical; on the other hand, and also on a subject by subject basis, the central 95% range of the ratios of the standard deviation of O-BEST estimates to the AMOCS estimate of spread included the particular ratio observed in our empirical study.

### *Discrimination*

Figure 3 shows results for each observer in the discrimination task in the same graphical format as in Figure 1, but now there are two sets of results per panel owing to the two different conditions in the discrimination experiment (vertical or horizontal orientation for the standard stimulus and orthogonal orientation for the test). The most salient characteristic now is that the average O-BEST estimate of the PSE matches much more accurately

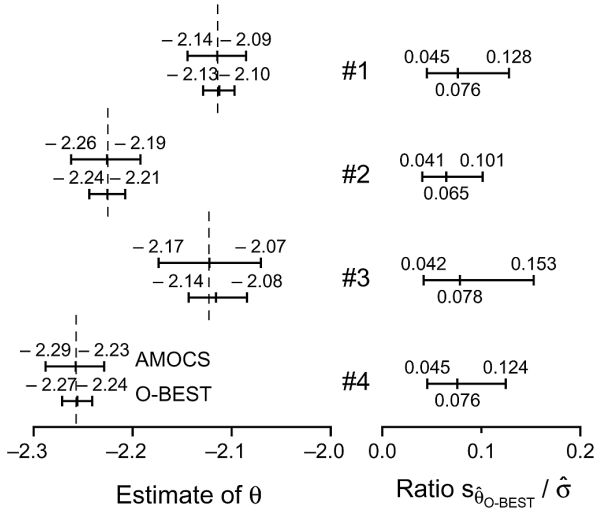


*Figure 2.* Results of a simulation of our detection experiment. Horizontal segments on the left side indicate, for each observer (vertically stacked), the central 95% range of the distributions of AMOCS and O-BEST estimates of threshold around the true threshold (vertical dashed segments) defined for each observer to be the AMOCS estimate shown in the corresponding panel of Figure 1. Horizontal segments on the right side indicate, also for each observer, the central 95% range of the distribution of the ratio of the standard deviation of O-BEST estimates of threshold to the AMOCS estimate of spread. In all cases, results are based on 5,000 replicates.

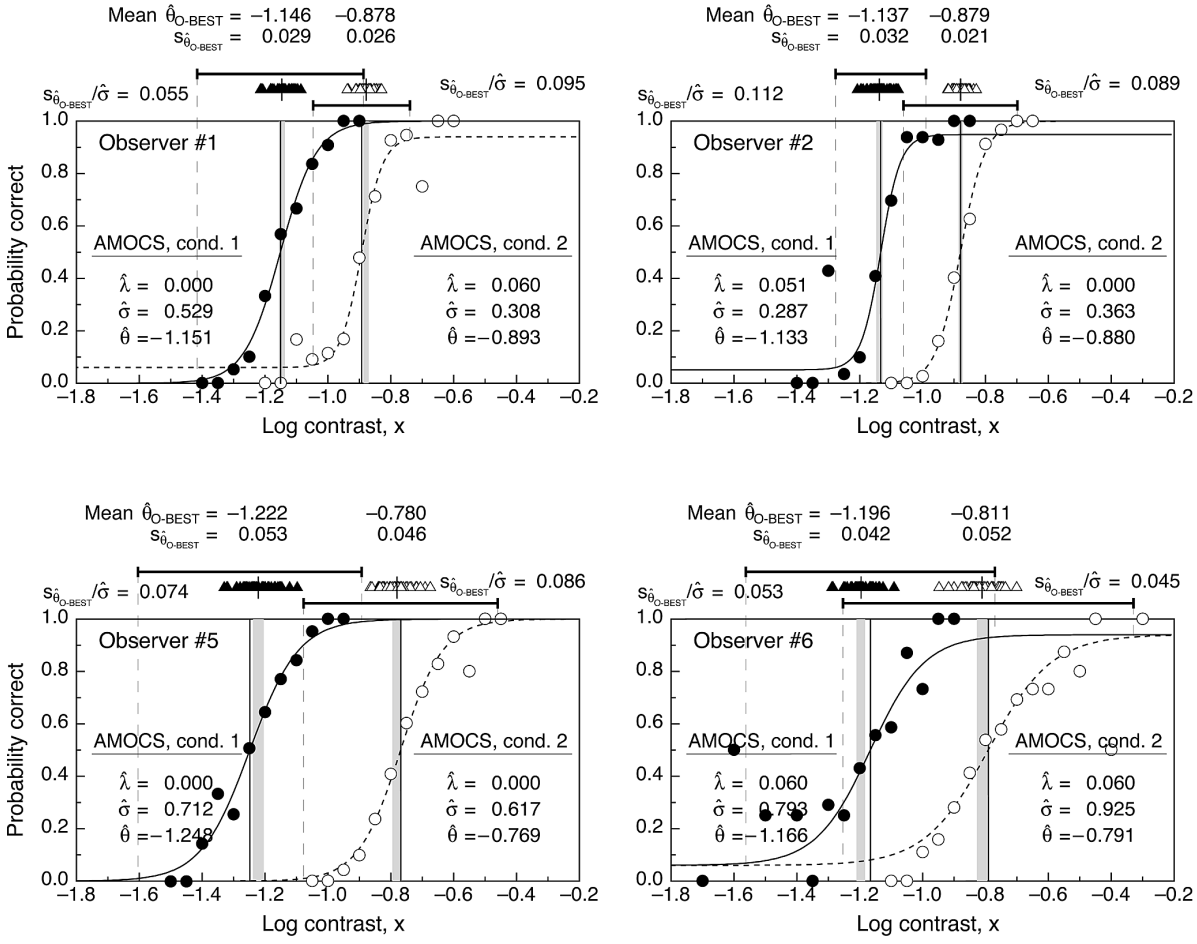


*Figure 3.* Results of the discrimination experiment for each observer. Observers #1 and #2 are the same as in Figure 1; observers #5 and #6 did not participate in the previous experiment. Each panel shows two sets of results: solid symbols and continuous lines for the condition in which the standard stimulus was horizontal and the test stimulus was vertical (condition 1) and open symbols and dashed lines for the condition in which the standard was vertical and the test was horizontal (condition 2). Numerical results printed on the left (alternatively, right) correspond to condition 1 (alternatively, 2). For simplicity, stars used to designate the parameters of the psychometric function for discrimination (see Equations 2, 4, and 6) have been removed. All graphical conventions as in Figure 1.

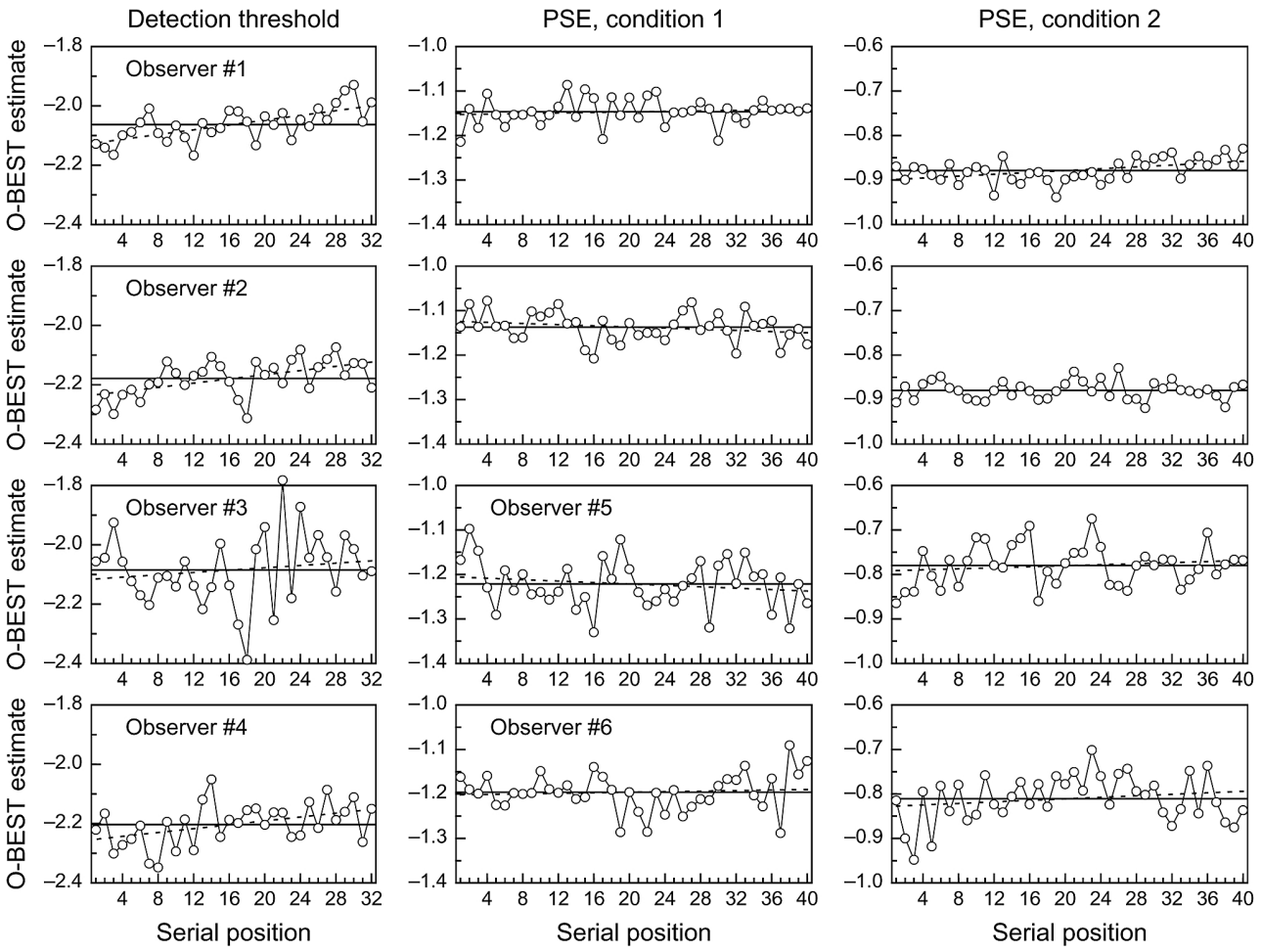


*Figure 4.* O-BEST estimates of detection threshold (left column) and PSE (centre and right columns) for each observer (rows) as a function of serial position along the experiment. Observers #3 and #4 only participated in the detection experiment whereas observers #5 and #6 only performed the discrimination experiment. The solid horizontal line across each panel is drawn at the ordinate of the average O-BEST estimate; the dashed line is the least-squares regression of O-BEST estimate on serial position. Note that estimates of detection threshold tend to increase as the experiment progresses, particularly for observers #1, #2, and #4. Conversely, estimates of the PSE do not show any noticeable trend

the value of the independent estimate of the PSE from AMOCS data (in the sense that they are close to each other for each observer and condition and that neither of them is systematically above or below the other), and that the 95% confidence interval for the PSE from O-BEST data generally includes the value that is independently estimated from AMOCS data.

On the other hand, the ratios of the standard deviation of O-BEST estimates to the spread of the underlying psychometric function range from 0.045 to 0.113 which, again, are occasionally larger than the figures 0.05–0.06 reported for 30-trial simulation runs by Alcalá-Quintana and García-Pérez (2005) when the referent was the actual spread of the psychometric functions. Simulations carried out along the lines described in the preceding section revealed that the central 95% range of the ratios computed with an AMOCS estimate of spread as referent almost always included the value that was empirically obtained. These simulations also confirmed that the average O-BEST estimate of the PSE matches the AMOCS estimate.

## Discussion

The results of detection and discrimination experiments demonstrated that the standard deviation of O-BEST estimates of threshold (or PSE) is related to the spread of the underlying psychometric function just as simulation studies have shown it should be. On the other hand, a discrimination experiment has also revealed that O-BEST estimates of the PSE are unbiased when the criterion is an independent estimate obtained with AMOCS, but O-BEST estimates of detection threshold are systematically higher than a criterion threshold location obtained from AMOCS data.

Thus, this empirical test confirms the utility of O-BEST for estimating the PSE in discrimination experiments, but its validity for estimating the detection threshold is suspect. In principle, the fact that the average O-BEST estimate of detection threshold is higher than a separate estimate of detection threshold obtained independently and concurrently with AMOCS indicates either that O-BEST overestimates the detection threshold or that AMOCS underestimates it. The rest of this section argues that the former alternative is more likely.

A closer look at our results reveals that individual O-BEST estimates of the detection threshold generally increase along the experiment (see the left column of Figure 4), whereas O-BEST estimates of the PSE do not follow a systematic pattern as the experiment progresses (see the center and right columns in Figure 4). In particular, increases in detection threshold (as indicated by least-squares regression; see the dashed line in each panel of Figure 4) across 32 occasions[2] range from 0.06 units (for observer #3) to 0.13 (for observer #1) through 0.10 (for observer #4) and 0.11 (for observer #2). On the contrary, changes in location of the PSE across 40 occasions[3] are much smaller and of both signs, ranging from –0.03 (for observer #5 in condition 1) to 0.04 (for observer #1 in condition 2).

Then, the average O-BEST estimate of threshold is clearly higher the larger the number of O-BEST blocks involved in the detection experiment. On the assumption that the *underlying threshold* does not actually change over time,[4] these results might reflect that the observer is either less attentive or more tired as the detection experiment progresses and, thus, has lapses leading to response errors that end up masquerading as an increase in threshold (see Peli & García-Pérez, 1997; Stuart, McAnally, & Castles, 2001).

As regards AMOCS, on the other hand, a similar serial analysis cannot be carried out because there is a single AMOCS estimate obtained by fitting the psychometric function to data from all applicable trials. At the same time, splitting these trials into subsets to fit separate psychometric functions at various stages along the detection experiment will reduce the dependability of the resultant estimates substantially owing to the reduced number of responses used to fit each function (see García-Pérez & Alcalá-Quintana, 2005). Yet, if the same inattention or fatigue discussed in the preceding paragraph occurred during the collection of AMOCS data (something that is tenable given that O-BEST and AMOCS blocks of trials were randomly interwoven in experimental sessions), response errors would have the same effect and will also end up masquerading as elevated thresholds (see Madigan & Williams, 1987).

In other words, inattention or some form of perceptual fatigue would cause O-BEST and AMOCS estimates of detection threshold to increase compared to the true underlying threshold. Under this interpretation, we must conclude that O-BEST is quantitatively more seriously affected by response errors and, thus, that O-BEST tends to overestimate the detection threshold. The hypothesis that inattention or fatigue are involved does not conflict with the fact that analogous effects were not observed in our discrimination experiment. Indeed, Meese (1995) has shown that response errors do not cause any effect on estimates of the PSE obtained with the Best PEST (Pentland, 1980; a method that is similar to O-BEST in many respects), and García-Pérez and Alcalá-Quintana (2005) have also shown that they do not affect estimates of the PSE obtained with AMOCS. Of course, the speculation that O-BEST is more affected than AMOCS by fatigue or inattention can be tested in a number of ways. First, a model of fatigue or inattention could be set forth and simulated to determine whether both psychophysical methods are differentially affected; some progress along this line has been made and very preliminary (but promising) results have been reported by Alcalá-Quintana and García-Pérez (2004b). Second, and empirically, interweaving O-BEST and AMOCS trials in the same block (instead of running them in separate blocks as in our experiments) would distribute the effects of these factors evenly across psychophysical methods and, arguably, would affect both of them identically.

In sum, the similarities and differences that have been reported here between simulation and empirical results on the performance of O-BEST in 2AFC detection and discrimination tasks seem to be explained on the assumption of growing inattention and fatigue along experimental sessions (something that simulations do not consider). These factors do not have any effect on estimates of the PSE in discrimination tasks (whether they are obtained with O-BEST or with AMOCS), but they do have a large effect in estimates of the detection threshold obtained with O-BEST. Although the ultimate cause of the inferior empirical performance of O-BEST in detection tasks has not been elucidated in the present paper, the fact that this performance is inferior seems well supported by our empirical and simulation results.

---

[2] Strictly speaking, there are only 16 measurements as far as serial position is concerned because O-BEST blocks consisted of two interwoven and concurrent runs, as described in Method.

[3] In this case there are indeed 40 measurements at different occasions in time because the two interwoven and concurrent runs in each O-BEST block pertained to different conditions, as described in Method.

[4] When threshold is operationally defined as a percent point on the psychometric function, a change in parameter $\lambda$ can affect the location of this operationally defined threshold without any actual change in the sensory threshold that would be observed for perfectly reliable observers for whom $\lambda = 0$. The sensory threshold obtained in these ideal conditions is the underlying threshold that we refer to in this discussion.

## References


Alcalá-Quintana, R. & García-Pérez, M. A. (2004a). The role of parametric assumptions in adaptive Bayesian estimation. *Psychological Methods*, *9,* 250–271.

Alcalá-Quintana, R. & García-Pérez, M. A. (2004b). Empirical performance of optimal Bayesian adaptive psychophysical methods. *Perception, Suppl.*, *33*, 178.

Alcalá-Quintana, R. & García-Pérez, M. A. (2005). Stopping rules in Bayesian adaptive threshold estimation. *Spatial Vision*, *18*, 347–374.

Alcalá-Quintana, R. & García-Pérez, M. A. (2007). A comparison of fixed-step-size and Bayesian staircases for sensory threshold estimation. *Spatial Vision*, *20*, 197–218.

García-Pérez, M. A. (2000). Optimal setups for forced-choice staircases with fixed step sizes. *Spatial Vision*, *13*, 431–448.

García-Pérez, M. A. & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, *8*, 256-289.

García-Pérez, M. A. & Alcalá-Quintana, R. (2007). Bayesian adaptive estimation of arbitrary points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, *60*, 147–174.

Green, D. G. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, *87*, 2662–2674.

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

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *Journal of the Acoustical Society of America*, *83*, 1852–1862.

Laming, D. & Marsh, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics*, *44*, 99–107.

Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick *TvC* method. *Vision Research*, *46*, 3160–3176.

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

Meese, T. S. (1995). Using the standard staircase to measure the point of subjective equality: A guide based on computer simulations. *Perception & Psychophysics*, *57*, 267–281.

Numerical Algorithms Group (1999). *NAG Fortran Library Manual, Mark 19*. Oxford: Author.

Peli, E. & García-Pérez, M. A. (1997). Contrast sensitivity in dyslexia: Deficit or artifact?. *Optometry and Vision Science*, *74*, 986–988.

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

Schlauch, R. S. & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, *88*, 732–740.

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

Stillman, J. A. (1989). A comparison of three adaptive psychophysical procedures using inexperienced listeners. *Perception & Psychophysics*, *46*, 345–350.

Stuart, G. W., McAnally, K. I., & Castles, A. (2001). Can contrast sensitivity functions in dyslexia be explained by inattention rather than a magnocellular deficit? *Vision Research*, *41*, 3205–3211.

Swift, D., Panish, S., & Hippensteel, B. (1997). The use of *VisionWorks*™ in visual psychophysics research. *Spatial Vision*, *10*, 471–477.

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


Received March 14, 2008
Revision received July 27, 2008
Accepted September 26, 2008