A Hierarchical Bayesian Approach to Adaptive Vision Testing: 

A Case Study with the Contrast Sensitivity Function

by

Hairong Gu¹#, Woojae Kim¹, Fang Hou¹, Luis Lesmes², Mark A. Pitt¹, Zhong-Lin Lu¹, and

Jay I. Myung¹

¹ Department of Psychology, The Ohio State University, Columbus, OH

² Adaptive Sensory Technology, Boston, MA

(# Corresponding author: Department of Psychology, 225 Psychology Building, The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210. Email: gu.124@osu.edu.)
Abstract

Measurement efficiency is of concern when a large number of observations are required to obtain reliable estimates for parametric models of vision. The standard entropy-based Bayesian adaptive testing procedures addressed the issue by selecting the most informative stimulus in sequential experimental trials. Non-informative, diffuse priors were commonly used in those tests. Hierarchical adaptive design optimization (HADO; Kim et al., 2014) further improves the efficiency of the standard Bayesian adaptive testing procedures by constructing an informative prior using data from observers who have already participated in the experiment. The present study is an empirical validation of HADO in estimating the human contrast sensitivity function (CSF). The results show that HADO significantly improves the accuracy and precision of parameter estimates, and therefore requires much fewer observations to obtain reliable inference about contrast sensitivity, comparable to the quick CSF method (Lesmes et al., 2010), which uses the standard Bayesian procedure. The improvement with HADO was maintained even when the prior was constructed from heterogeneous populations or a relatively small number of observers. These results, taken together, justify that HADO can be used in Bayesian adaptive testing by replacing non-informative, diffuse priors with statistically justified informative priors without introducing unwanted bias.

Keywords: visual psychophysics, Bayesian adaptive estimation, hierarchical Bayesian modeling, informative priors, contrast sensitivity.
**Introduction**

Measurement in controlled experiments serves as a rigorous and objective avenue to obtain reliable inferences in scientific and clinical investigations. In vision, a general interest is to measure human performance in visual tasks to infer how external visual stimuli are transformed, and processed through the visual system to yield perceptual experience. To accurately measure how visual performance, e.g., detection threshold, varies within a multi-dimensional feature space of stimuli, normally a large number of stimuli covering the region of interest in the feature space need to be presented to an observer and tested. In most cases, using a pre-selected, fixed set of stimuli, a.k.a. the method of constant stimuli, is not practical given the restrictions on the time and cost of data collection. The drawback of the method of constant stimuli is that the stimuli selected before the experiment are not equally informative for different individuals, especially when clinical populations are considered. To combat this heterogeneity among individuals, numerous adaptive methods have been developed to improve the efficiency and precision of vision testing.

**Adaptive Testing in Vision**

The non-parametric staircase procedure has been the dominant adaptive method of testing human sensory abilities (Dixon & Mood, 1948; Kesten, 1958; Taylor & Creelman, 1967). Despite its popularity, the staircase procedure typically only accommodates one-dimensional variation of stimuli, and is used for estimating thresholds at predetermined performance levels on the psychometric function. Further, although staircase procedures with various down-up rules have been proposed and their properties have been systematically studied (Levitt, 1971; Kaernbach, 1991), their theoretical underpinnings concerning optimality are not well-grounded in probability theory.
A new generation of adaptive testing methods overcomes these limitations by taking advantage of statistical models that capture the regularities of the performance with parametric functions to improve the testing efficiency. A statistical model reduces the description complexity to a few parameters in a mathematical function, thereby requiring fewer observations for accurate and precise estimation. These methods include best-PEST (Pentland, 1980), QUEST (Watson & Pelli, 1983) and ZEST (King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994). Nevertheless, these methods are limited by the strong assumption that an estimate of the underlying threshold is always optimal to be used in a new trial, without rigorously quantifying the usefulness of each stimulus in improving statistical inferences.

A more theoretically sound adaptive procedure is a class of entropy-based Bayesian adaptive procedures, such as the psi method (Kontsevich and Tyler, 1999), the quick method (Lesmes, et al, 2006, 2010, 2015; Lu & Dosher, 2014), Bayesian adaptive estimation (Kujala & Lukka, 2006), adaptive design optimization (ADO, Cavagnaro, Myung, Pitt, & Kujala, 2010), and active data collection (DiMattina & Zhang, 2008, 2011; DiMattina, 2015). Formulated under an information-theoretic Bayesian framework, these procedures update the posterior distributions of the parameters in a psychological function sequentially with incoming observations. The usefulness of each stimulus on a given trial is quantified by the expected reduction of entropy of the posteriors, or the uncertainty of the parameters. In a general term, the utility function measures the usefulness of a given stimulus choice, written in the following form:

\[ U(s) = \int_y \int_\theta u(s, y, \theta) p(y|s, \theta)p(\theta)dyd\theta, \]  

where \( u(s, y, \theta) \), called the sample utility, is a function of stimulus \( s \), observation \( y \) and parameter \( \theta \), \( p(y|s, \theta) \) is the statistical model, and \( p(\theta) \) is the prior distribution of \( \theta \). For example, a
psychometric function $p(y|s, \theta)$ describes the probability of correct response for a given stimulus $s$, with the parameter $\theta$ containing a threshold and slope, and the response $y$ being either a correct response or an incorrect response. The sample utility $u(s, y, \theta)$ quantifies the usefulness of stimulus $s$ with a specific parameter value $\theta$ and a potential response $y$. A particular specification of $u(s, y, \theta)$ is

$$u(s, y, \theta) = \log \frac{p(\theta|y,s)}{p(\theta)},$$

(2)

in which $p(\theta)$ is the prior distribution of $\theta$, and $p(\theta|y,s)$ is the posterior. Therefore $u(s, y, \theta)$ can be interpreted as the reduction in the uncertainty about parameter $\theta$ after a new trial with a stimulus $s$ and an observation $y$. By taking the integral of the sample utility over all possible observation $y$ and parameter $\theta$, the derived expected utility, $U(s)$, in Eq. (1) measures the expected information gain of stimuli $s$ (Cover & Thomas, 1991). The design that maximizes the expected utility is selected and presented in the next experimental trial. Hence, the optimal stimulus is expected to yield the largest information gain about the psychological function upon the observation of a response in the next trial.

The computation of Eqs. (1) – (2) requires the specification of $p(\theta)$, the prior distribution that represents the current state of knowledge about model parameters. For each trial during an adaptive testing session, the prior distribution is updated in a straightforward way by applying Bayes rule with incoming data. However, the initial prior at the beginning of an experiment must be specified a priori by researchers. Commonly in Bayesian adaptive testing of visual functions, conservative priors, either uniform (Kontsevich & Tyler, 1999; Kujala & Lukka, 2006; Lesmes, Jeon, Lu, & Dosher, 2006) or diffuse (Lesmes, Lu, Baek, & Albright, 2010; Hou, Huang, Lesmes, Feng, Tao, Zhou, & Lu, 2010; Lesmes et al, 2015), have been used.
One advantage of Bayesian inference is that statistical inferences can be enhanced by starting each new inference from previously collected observations, quantified as an informative prior (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Although these adaptive methods adopt full Bayesian procedures for parameter estimation, the advantage of informative priors is yet to be evaluated. Consequently, the standard Bayesian adaptive testing methods are needlessly conservative in that every experiment starts its inference anew with a ground-zero state of knowledge, even if there might be several observers’ worth of data already in hand.

A Hierarchical Adaptive Approach

To further improve the current Bayesian testing procedure, we recently introduced a hierarchical Bayesian adaptive testing method, *hierarchical adaptive design optimization* (HADO; Kim, Pitt, Lu, Steyvers, & Myung, 2014), which achieves even greater efficiency by applying informative priors constructed using data from observers who have previously conducted the same task. The improvement is brought about by integrating hierarchical Bayesian modeling (HBM) with the standard entropy-based Bayesian adaptive testing procedures. In Kim et al. (2014), HADO was applied to estimating the contrast sensitivity function (CSF) in a series of simulations with CSF data from 67 amblyopic and 80 normal eyes, using as a baseline the quick CSF procedure (quick CSF, Lesmes et al., 2010), which embodies the standard Bayesian testing procedure to measure the CSF. A leave-one-out paradigm was used to compare HADO with quick CSF by treating 146 subjects as being previously tested and the remaining subject as a new individual to be measured subsequently. We showed that HADO achieved around 2dB decrease in the root-mean-square-error (RMSE) in the estimation of CSF during the first 40 trials, and saved over 20 trials to reach a 90% correct classification as an amblyopic eye versus a normal eye, compared to the quick CSF procedure in a two alternative forced-choice task.
Although HADO achieves greater efficiency in simulations, some of its working assumptions are in need of further investigation. For example, it was assumed in HADO that the observers used for constructing informative priors were from the same population as the new observers, and the number of observers was large enough to represent the target population. The purpose of the present study is to empirically validate HADO as a reliable and general testing procedure by evaluating those assumptions. In what follows, we first provide a brief overview of the methodological foundation of HADO, and then present two experiments that applied HADO to estimate the CSF.

**Hierarchical Adaptive Design Optimization (HADO)**

The methodological foundation of HADO (Kim et al., 2014) features the integration of the standard entropy-based Bayesian method with another statistical tool, hierarchical Bayesian modeling (HBM; Bernardo & Smith, 1994; Rouder & Lu., 2005; Lee, 2006), a statistical technique that improves the precision of inferences by accounting for dependencies in data. Particularly, observers from the same population or tested in the same experimental condition are expected to perform similarly to each other than those from different populations or conditions. The innovative application of HBM in HADO is to model the similarity, or the common information shared by observers, and turn it into informative priors for new observers.

The framework of HADO is illustrated in Figure 1. As a sub-routine in HADO, the shaded area represents the standard entropy-based Bayesian adaptive testing procedure, which consists of three basic steps that are repeated in each trial: 1) design optimization (finding the optimal stimuli); (2) measurement (presenting stimuli and collecting response); (3) Bayesian updating of prior to posterior. On the \(t\)-th trial, the prior is expressed as \(p(\theta_n | y_n^{(1:t-1)})\), in which \(y^{(1:t-1)}\) denotes responses in the previous trials, and \(n\) indexes observers. The utility of a
stimulus \( U(s_t) \) quantifies the expected reduction of entropy of parameters each stimulus can potentially bring, calculated by Eq. (1). Then the optimal stimulus \( s^*_t \) corresponding to the maximal \( U(s_t) \) is administered and a response \( y^{(t)}_n \) is observed. The posterior distribution of parameters is calculated by Bayes rule \( p(\theta_n|y^{(1:t)}_n) \propto p(y^{(t)}_n|\theta_n, s^*_t)p(\theta_n|y^{(1:t-1)}_n) \). The posterior \( p(\theta_{n}|y^{(1:t)}_n) \) is subsequently used as the prior for the next, \((t+1)\)-th trial. These steps repeat until a given number of trials are executed or a given criterion for accuracy and precision of estimation is reached.

Figure 1: The illustration of HADO algorithm. The standard Bayesian adaptive procedure, e.g., the \( \psi \) method (the shaded area), is an integral part of the whole HADO algorithm. See the text for additional details. Reprint from Kim et al. (2014, Figure 1).

The standard Bayesian adaptive testing procedure leaves open the option of priors at the beginning of an experiment, to which hierarchical Bayesian modeling (HBM) contributes. The
upper loop of Figure 1 represents the HBM part of HADO that draws information from
previously run observers to provide an informative prior for a new observer. Given a dataset of \( n \)
observers, a hierarchical Bayesian model can be formulated as

\[
\begin{align*}
\eta & \sim p(\eta) \\
\theta_i | \eta & \sim p(\theta_i | \eta) \\
y | s, \theta_i & \sim p(y | s, \theta_i)
\end{align*}
\]

(3)

The lower level, represented in the third line, contains an individual-level data model \( p(y | s, \theta_i) \)
that describes how a response \( y \) is generated given stimulus \( s \) and the \( i \)th observer’s parameter \( \theta_i \).
A higher-level model \( p(\theta_i | \eta) \), in the second line, defines the assumption of the dependency
among individual-level parameters \( \theta_i \)'s, conditional on the higher-level parameter \( \eta \) associated
with the population. Depending on a researcher’s assumption, \( p(\theta_i | \eta) \) may be modeled by a
parametric distribution. For instance, if we assume that the CSFs of a population follow a
Gaussian distribution \( N(\theta_i | \mu, \Sigma) \), then \( \eta \) is the mean \( \mu \) and variance \( \Sigma \). Alternatively, a
nonparametric model such as kernel density estimator can be specified if the underlying
distribution is believed to be significantly deviant from standard parametric distributions. The
first line of Eq. (3) specifies the prior distribution of the higher-level parameter \( \eta \). When \( n \)
observers worth of data are collected with observed responses \( y_{1:n} \) (the subscript \( 1:n \) denotes a
collection of observations from \( n \) observers), the posterior distribution of the parameter \( \eta \) is
obtained by

\[
p(\eta | y_{1:n}) \propto \int \prod_{i=1}^{n} p(y_i | \theta_i) p(\theta_i | \eta) p(\eta) \, d\theta_1 \ldots d\theta_n.
\]

(4)

Subsequently, the prediction of the parameter \( \theta_{n+1} \) for a new observer is made by

\[
p(\theta_{n+1} | y_{1:n}) = \int p(\theta_{n+1} | \eta)p(\eta | y_{1:n}) \, d\eta.
\]

(5)
It is important to note that the above $p(\theta_{n+1}|y_{1:n})$ will serve as an informative prior for the next, $(n+1)$-th observer in the experiment. It can be expected that with the increase in the number of collected observers $n$, $p(\theta_{n+1}|y_{1:n})$ contains more information and therefore becomes more concentrated. On the other hand, when no prior data are available (i.e., $n = 0$), HADO is reduced to the standard Bayesian adaptive testing procedure method with a non-informative, diffuse prior.

The HADO algorithm is implemented through two sub-routines, the standard Bayesian testing procedure (e.g., the quick CSF method) and HBM, each corresponding to the adaptive and hierarchical component, respectively. In the current implementation, the Bayesian testing procedure is executed online during an experiment to select an optimal stimulus on each trial, while HBM is executed offline after the experiment, to update the population-level structure with existing datasets to construct a prior for a new observer.

To evaluate the performance of HADO in an empirical study, we compared a HADO-informed CSF with the standard Bayesian testing procedure.

**Contrast Sensitivity Function (CSF)**

The CSF describes how contrast sensitivity changes as a function of spatial frequency, and can serve as a comprehensive assessment of spatial vision. The CSF is closely related to daily visual functions and has been used to characterize both normal and impaired vision (Ginsburg, 1981, 2003; Hess, 1981). Accurate measurement of CSFs is of keen interest for the purpose of diagnosing visual deficits because various visual pathologies are associated with characteristic changes in the CSF (Hess, 1978; Jindra & Zemon, 1989; Marmor, 1981; Wolkstein, Atkin, & Bodis-Wollner, 1980). However, measuring CSFs with adequate precision typically requires a large number of observations because of its complex shape and the
requirement of proper sampling in both contrast and spatial frequency domains. To reduce the
data collection burden, adaptive testing methods have been extensively studied and exploited in
measuring CSF (Dorr, Lesmes, Lu, & Bex, 2013; Hou et al., 2010; Lesmes et al., 2010; Hou et
al, 2015).

Lesmes et al. (2010) proposed a quick CSF method that adopts the entropy-based
Bayesian testing procedure to estimate the contrast sensitivity function. The regularities in
contrast sensitivity were modeled by a truncated log-parabola model with four parameters
(Watson & Ahumada, 2005):

\[
S(f) = \begin{cases} 
\gamma_{max} - \delta & \text{if } f < f_{max} - \frac{\beta}{2} \sqrt{\frac{\delta}{2 \log_{10} 2}}, \\
\gamma_{max} - (\log_{10} 2)(\frac{f-f_{max}}{\beta/2})^2 & \text{otherwise.}
\end{cases}
\]  

(6)

Figure 2a illustrates the parametrization of the CSF in terms of the four parameters ($\gamma_{max}$, peak
sensitivity; $f_{max}$, peak frequency; $\beta$, bandwidth at half of the peak sensitivity; $\delta$, low-frequency
truncation level). The contrast sensitivity function $S(f)$ in Eq. (6) is the reciprocal of the contrast
threshold, corresponding to stimulus contrasts that are associated with a pre-defined performance
level.
Figure 2: (a) Contrast sensitivity function $S(f)$ as a function of spatial frequency $f$, parameterized by peak sensitivity $\gamma_{\text{max}}$, peak frequency $f_{\text{max}}$, bandwidth $\beta$, and truncation level $\delta$ (Watson & Ahumada, 2005); (b) The psychometric function $p(c, f)$ for a 10AFC task that describes the probability of detecting stimuli of contrast $c$, with the threshold determined by contrast sensitivity function $S(f)$.

The statistical model in the quick CSF method is a psychometric function defined by a cumulative Gaussian function (Alcalá-Quintana & García-Pérez, 2004):

$$p(c, f) = G + (1 - G - L)\Phi\left(\frac{c - S(f)}{\sigma}\right), \quad (7)$$

where $p(c, f)$ is the probability of generating a correct response at a specific contrast level $c$ and a spatial frequency $f$, $G$ is the guessing rate, $L$ is the lapse rate, $\Phi(\cdot)$ is the cumulative Gaussian function, and $\sigma$ determines the slope of the psychometric function. An example of a cumulative Gaussian psychometric function is shown in Figure 2b. The guessing rate is the probability of making a correct response when the contrast of stimuli approximates zero. In an $N$-alternative, forced-choice ($N$-AFC) task, it is reasonable to set the guess rate to $1/N$. Figure 2b shows a guess rate of .1 for a 10-AFC task (Hou, Lesmes, Bex, Dorr, & Lu, 2015). The lapse rate restrains the maximum probability of correct response to account for response errors caused by inattention. The slope of the psychometric function is pre-set to a value obtained from pilot experiments.

Eq. (6) and Eq. (7), combined as $f(y|c, f, \gamma_{\text{max}}, f_{\text{max}}, \beta, \delta, G, L, \sigma)$, mathematically describes how external stimuli variables $c$ and $f$ are transformed into underlying visual sensitivity, tuned by parameters $(\gamma_{\text{max}}, f_{\text{max}}, \beta, \delta, G, L, \sigma)$ specific to each observer being tested, and finally reflected as the probability of correct responses. The parameters to be estimated in the
current study are $\gamma^{max}, f^{max}, \beta,$ and $\delta$. The other three parameters $G, L, \sigma$ are fixed in the current study ($G=0.1; L=0.04; \sigma=0.42$) (Hou et al, 2015). In a hierarchical modeling context, the model $f(y|c, f, \gamma^{max}, f^{max}, \beta, \delta)$ describes the individual-level data, shown in the third line in Eq. (3).

The quick CSF method, as the other standard entropy-based Bayesian testing procedures, assumes non-informative prior at the beginning of the experiment. In the present study, the quick CSF procedure was used as a benchmark to compare with the HADO procedure in an empirical validation.

**Hierarchical Adaptive Estimation of the CSF**

The soundness of HADO hinges upon the validity of two main assumptions it makes. First, informative prior obtained based on data from existing data should be representative of a new observer. In other words, new observers are assumed to come from the same population as earlier observers. If this assumption is violated, the prior is no longer informative, and even would bias the estimate of new observers. The second assumption is that to construct an informative prior, the size of the data sample (i.e., number of existing observers) should be large enough to well approximate the population distribution. If data from only a small number of observers are available, the prior constructed from the data might be either too diffuse because of the lack of information from existing data, or more susceptible to bias when these observers are not representative of the population.

With these assumptions in mind, we designed two validation experiments by manipulating the extent of agreement with the theoretical assumptions of HADO described above (i.e., sample representativeness and sample size). Data from a baseline experiment, in which a large number of observers took participated in CSF experiments under three different luminance conditions, reported in Hou, Lesmes, Kim, Gu, Pitt, Myung, & Lu (under review),
were used to build the informative priors for the two experiments in the present study. In
Experiment 1, priors constructed from different luminance conditions were applied to assess the
effect of priors’ representativeness on HADO performance. In Experiment 2, priors constructed
from samples of different sizes were used to gauge how large a sample needs to be to construct
an effective prior. In what follows, we first provide a brief description of the baseline experiment
upon which the two validation experiments were built.

Baseline Experiment: Data Collection for Prior Construction

Data from the baseline experiment served the purpose of establishing priors, i.e., \( p(\theta|\eta) \)
in Eq. (3), to be used in both validation experiments of HADO. For that purpose, the CSFs of
112 observers were measured under three different luminance conditions using the quick CSF
method with a 10 alternative forced-choice task (Lesmes et al., 2010; Hou, et al, 2015). For
specific details and analysis of this experiment, readers are directed to Hou et al. (under review).
In the following, only the information relevant to the present study is described.

Each observer received six blocks of quick CSF measurements in four different viewing
conditions: low luminance (L), medium luminance (M), high luminance (H) and low pass (LP).
In the H condition, subjects viewed the display through uncovered goggles. In the M condition,
subjects viewed the display binocularly through goggles with neutral density filters with an
attenuation factor of 0.67 decimal log units. In the L condition, subjects viewed the display
through goggles fit with two neutral density filters with a total optical density of 1.56 log units.
Bangerter occlusion foils were used as the low pass filter in the LP condition. The equivalent
mean luminance in the L, M, H and LP conditions was 2.62, 20.4, 95.4 and 95.4 cd/m²,
respectively. The order of the test blocks was L, L, M, H, LP and H. The first L condition was
used for observers to dark-adapt and practice the quick CSF test, and the two H conditions were
included to assess the test-retest reliability of the quick CSF method. In each test block, the quick
CSF procedure with a 10AFC letter identification task was used to measure the CSF in 50 trials.
Each observer finished the six blocks in approximately 70 minutes. For details about the quick
CSF procedure (i.e., use of a diffuse prior, adaptive stimuli selection, and Bayesian estimation),
please refer to Hou et al. (2015).

In the current study, we used data from 100 of the 112 observers in the baseline
experiment for prior construction, and only used data in the H, M and L luminance conditions.
For each observer, the point estimates of the four parameters of the truncated log-parabola CSF
model conditions were computed as the Bayesian posterior means under each luminance
condition. The estimates across all 100 observers in each luminance condition were then pooled
together. Nonparametric kernel density estimation (KDE, Scott, 2009) was then applied to
estimate the population-level distribution of the parameters, which is the higher-level model
referred to earlier (i.e., $p(\theta | \eta)$) in Eq. (3). To visualize these higher-level distributions, we
mapped the estimates of the four parameters onto two summary statistics of a CSF, the area
under the log CSF (AULCSF: Applegate, Howland, Sharp, Cottingham, & Yee, 1997; Oshika,
Okamoto, Samejima, Tokunaga, & Miyata, 2006) and the cutoff spatial frequency (cutSF:
Huang, Tao, Zhou, & Lu, 2007; Zhou, Huang, Xu, Tao, Qiu, Li, & Lu., 2006), both of which are
diagnostic measures of contrast sensitivity (Lesmes et al, 2010; Hou et al., 2010; Hou, et al,
2015). As such, the four dimensional distributions of CSF parameters were transformed into a
two dimensional distribution of AULCSF and cutSF. Figure 3 shows the 75% equal-density
contours of these distributions corresponding to the three conditions. Differences among the
distributions are clearly visible in their locations in the parameter space, which are attributable to
the experimental manipulations. Given that larger values of AULCSF and cutSF indicate better
vision, the distribution of CSFs in the H condition locates upper right in the space. Distributions representing the M and L conditions are located in regions covering smaller values of AULCSF and cutSF, exhibiting the expected ordering based on our luminance manipulations.

Figure 3: Equal-density contours of the estimated population distributions of AULCSF and cutSF under the H, M and L luminance conditions estimated from the data of 100 observers in the baseline experiment.

Experiment 1: Effect of Different Types of Priors

The goal of Experiment 1 was to assess whether the use of informative priors, estimated from data in the baseline experiment, can help achieve greater efficiency in the estimation of CSF, compared to the quick CSF method that assumes non-informative, diffuse priors, and if so, to determine the size of the benefit. The choice of priors is straightforward, which is to use the H prior to new observers in the H luminance condition, and the L prior to observers in the L luminance condition. However, in practice, the choice of priors may not always be clear. For
example, if an observer comes from an unknown population, the imposition of a prior of strong beliefs risks introducing an unjustified bias in parameter estimation (e.g., an L prior given to a CSF measurement in the H condition). In such a case, it may be wiser to use a prior constructed from the collapsed data from different populations, which may result in a prior still more informative than a diffuse prior. To investigate this possibility, we assessed the influence of additional types of priors, e.g., mis-specified and mixed-population priors, on HADO performance.

**Methods**

- **Observers**
  
  10 college students from The Ohio State University, participated to obtain partial course credit in an Introductory Psychology course. All observers had normal or corrected-to-normal vision, and were naive to the purpose of the study. Verbal consent was obtained prior to participation. The study protocol was approved by the institutional review board of human subjects research of the Ohio State University.

- **Apparatus**

  The experiment was implemented in Matlab R2013a (The Mathworks Corp., Natick, MA) with the Psychtoolbox subroutines (Kleiner, Brainard, Pelli, Ingling, Murray & Broussard, 2007) on a PC. Stimuli were presented on a gamma-corrected Samsung UN55FH6030 55” monitor with a 1920 ×1080 pixel resolution and a vertical refresh rate of 60 Hz. The mean luminance of the monitor was 95.4 cd/m² (measured by a Tektronix J17 photometer). A bit-stealing algorithm was used to achieve 9-bit gray-scale resolution (Tyler, 1997). Observers viewed the display binocularly from a distance of 4 meters in a dark room. A chin-rest was used
to help observers fix their head position relative to the screen. Two luminance conditions, H and L, the same as those in the baseline experiment, were tested.

- Stimuli

Ten filtered Sloan letters, C, D, H, K, N, O, R, S, V and Z (shown in Figure 4), were used as stimuli (Alexander, Xie, & Derlacki, 1994; Hou, Lu, & Huang, 2014; Hou et al, 2015). All filtered letter stimuli had a center frequency of 3.3 cycles per object (cpo) and bandwidth (of half height) of one octave. The filtered letters had narrow band spectrum in the spatial frequency domain and were found to assess contrast sensitivity in different central spatial frequencies equivalently as the conventional gratings (Alexander et al., 1994; McAnany & Alexander, 2006; Hou et al, IOVS 2014: ARVO E-Abstract 770). The pixel intensity of each filtered image was normalized by the maximum absolute intensity of the image. After normalization, the maximum absolute Michelson contrast of the image is 1.0. Stimuli with different contrasts were obtained by scaling the intensities of the normalized images with corresponding values. The filtered images were rescaled to 19 different sizes to generate stimuli with 19 evenly spaced (in log space) central spatial frequencies ranging from 1.19 to 30.95 cycles per degree (cpd) for the quick CSF procedure.
Figure 4: a) Ten filtered letters. b) Illustration of filtered letter ‘C’ in various spatial frequency conditions.

- Design and procedure

Three different informative priors (H, L and Mixture), plus the diffuse prior as a baseline, were constructed in the current experiment. The H and L priors were exactly the estimated population distributions of the CSFs under the H and L luminance conditions as shown in Figure 3. The Mixture prior was obtained by averaging the H, M, and L distributions with equal weights. Therefore the Mixture prior contains information about all three conditions (i.e., a general population) so as to represent a wide spread of beliefs in the parameter space but still more informative beliefs than a diffuse prior. The diffuse prior was the same as used in Hou, et al. (2015), which is a hyperbolic secant function close to being flat over the parameter space.

Each observer ran two testing blocks with a break in between, one under the H luminance condition and the other under the L luminance condition. In each block, four independent quick CSF experiments with different priors were conducted simultaneously with stimuli from different prior conditions interleaved. Each quick CSF experiment consisted of 50 trials, and
therefore in total each observer received a total of 200 trials under each of the two luminance conditions.

The rest of the experimental setup was the same as in the baseline experiment. The stimuli were defined on discrete grids in each dimension, with 128 grids in the contrast dimension (evenly spaced on a base-10 log scale from 0.2% to 100%) and 19 grids in the spatial frequency dimension (evenly spaced on a base-10 log scale from 1.19 to 30.95 cpd). To improve the experience of observers, on each trial, two other easier stimuli, assuming 2 and 4 times of the optimal luminance contrast (capped at 0.9) were presented along with the stimulus with optimal contrast selected by the quick CSF algorithm, and all three stimuli had the same optimal spatial frequency. The stimuli were represented by three letters randomly chosen with replacement among the ten Sloan letters and displayed in a row with a center-to-center distance of 1.1 times letter size.

Results

Ideally, assessing the performance of CSF estimation would require knowing the true, underlying CSF of an observer. However, given that the underlying CSF is unknown, it was approximated as follows. First, for each observer under a particular luminance condition, all 200 trials from the four prior conditions were collapsed into one data set. The posterior distribution of the CSF parameters was then obtained by applying Bayes rule to the data under a diffuse prior. Finally, the posterior mean was taken as our estimate of the true CSF. This method was applied to the data from each observer under each luminance condition to obtain separate true CSFs in the H and L conditions.
Point estimates of the CSF parameters were obtained for each experimental trial, and transformed to a summary statistics, AULCSF. To assess the quality of the estimates, the root-mean-square-error (RMSE) was calculated for each experimental trial \( (j = 1, \ldots, 50) \) by

\[
\text{RMSE}_j = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (\hat{\theta}_{i,j} - \hat{\theta}_{i,T})^2} \quad j = 1, \ldots, 50;
\]

where \( \hat{\theta}_{i,T} \) denotes the approximation of the true AULCSF of the \( i \)-th observer estimated as described above given a luminance condition, and \( \hat{\theta}_{i,j} \) is the estimate of the AULCSF of the \( i \)-th observer on the \( j \)-th trial. Note that the RMSE reflects a combination of accuracy and precision in measurement theory (or equivalently, bias and variance in statistical inference) in a single summary statistic (Wackerly, Mendenhall & Scheaffer, 2007). Accordingly, it can be considered an empirical instantiation of the mean squared error in the theory of point estimation in statistics (Lehmann & Casella, 1998).

Figures 5a and 5c (left-hand column) show the RMSE profiles over trials based on quick CSF with different priors under the H and L luminance conditions, respectively. Overall, regardless of the priors, the estimation error decreases as observations accumulate over trials. The estimation errors under all prior types were smaller than 1.8 dB under the H luminance condition, and 1.2 dB under the L condition at the end of the experiment (i.e., 50 trials). The effects of informative priors were assessed by comparing their RMSEs with those from the diffuse prior. Further, the effect of priors is evident in the graphs, and largest at the beginning of the experiment. When priors matched the condition (or population), specifically when the H prior is used under the H luminance condition (black curve in Figure 5a) or the L prior is used under the L luminance condition (red curve in Figure 5c), the measurement error is smallest across trials. Estimation with the Mixture prior performs worse in both luminance conditions.
(green curves in both plots) than the correctly specified priors, but the results are still far superior to the performance under the diffuse priors (blue curves). In a misspecification scenario, specifically when the L prior is used in the H luminance condition (red curve in Figure 5a) and the H prior is used in the L condition (black curve in Figure 5c), the estimation error becomes larger than with the correctly specified prior and the Mixture prior. In the H luminance condition, in particular, the error with the mis-specified, L prior is overall comparable to that with the diffuse prior and gets even worse after several trials. By contrast, in the L condition, the use of the mis-specified, H prior fares better. This can be explained by the general, asymmetric shape of these prior distributions which are skewed toward poor vision, revealed in Figure 3 and Figure 6. That enables the H prior to have better coverage of the L CSFs than the L prior does to the H CSFs.

These empirical results were obtained from only 10 observers and thus tend to be noisy due to idiosyncratic sampling errors. To provide clarity, simulations were performed to complement the experimental findings. We used the approximated true CSFs of the 10 observers (estimated from all responses with a diffuse prior) to generate simulated responses. For each CSF, Experiment 1 was executed 100 times with the simulated data. The RMSEs across all observers and replications were computed.
Figure 5: RMSE plots of AULCSF estimation in Experiment 1 with the four different types of priors under the H (a, b) and L (c, d) luminance conditions.

Figures 5b and 5d (right-hand column) are the plots of RMSEs for the H and L luminance conditions obtained from the simulation. As can be seen, the simulation results are qualitatively indistinguishable from the experimental ones shown on the left panel, but much smoother due to more replications in each condition.

In Table 1, the average reduction of RMSEs by using informative priors from using the diffuse prior based on the simulation data was shown. After 10 trials, the correctly specified priors (the H prior for the H condition and the L prior for the L condition) reduce error by 4.58
dB in the H condition and 4.99 dB in the L condition compared to the diffuse prior. Using the Mixture prior makes 4.29 dB and 4.54 dB reduction in the H and L luminance conditions, respectively. By contrast, the use of mis-specified priors (the L prior in the H condition and the H prior in the L condition) result in only 0.28 dB and 2.88 dB reductions in the H and L conditions, respectively. After the entire 50 trials, the difference in estimation between priors becomes smaller because of the contribution of data to the estimation. The mis-specified priors performed nearly on par with the diffuse prior in the end, although the L prior in the H condition is slightly worse by 0.78 dB and the H prior in the L condition is slightly better by 0.62 dB.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Prior</th>
<th>After Trial 10 (dB)</th>
<th>After Trial 50 (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>Diffuse</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>-4.58</td>
<td>-1.90</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>-4.29</td>
<td>-1.86</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-0.28</td>
<td>0.78</td>
</tr>
<tr>
<td>L</td>
<td>Diffuse</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>-2.88</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>-4.54</td>
<td>-1.59</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>-4.99</td>
<td>-1.71</td>
</tr>
</tbody>
</table>

Table 1: Simulation results of the reduction of RMSE in the estimation of AULCSF by using the H, Mixture and L priors compared to the diffuse prior, in the H and L luminance conditions.

To summarize the main findings from Experiment 1, the results show that using an informative prior, estimated from data belonging to the same population as new observers,
greatly improves the efficiency of measurement, reflected in requiring fewer trials to attain the
same quality of parameter estimation. In other words, with the same number of observations, a
representative informative prior attains greater accuracy and precision in estimation. As shown in
Figure 5, it takes 20 to 30 trials for the diffuse priors to reach the initial error level of estimation
found with the correctly specified priors at the beginning of an experiment. Although CSF
estimates obtained by using any priors gradually converge to stationary values as more
observations are collected, the initial advantage of using an informative prior is valuable when
there is a restriction on the time of the testing session.

Another point is that using a mis-specified prior can produce an even worse outcome than
using a diffuse prior if a large bias is contained in the prior. The identification of a suitable prior
(or identify the population for an observer) may not be straightforward in practical situations.
One feasible solution suggested in the current results is to use a mixture prior that represents a
wide range of observers in the population. Although a mixture prior did not provide as much
improvement in estimation as the correctly specified informative prior, it can help mitigate the
problem of misspecification and at the same time still outperforms non-informative diffuse
priors.

Experiment 2: Effects of Sample Sizes with an Informative Prior

In Experiment 1, the informative priors were constructed using the data from 100
observers in the baseline experiment. The large sample size ensured that the sample was
sufficiently representative of the population, and the estimated informative prior, \( p(\theta|\eta) \) in Eq.
(3), was much more concentrated than the diffuse prior. In practice, however, it may not be
feasible to collect that many observers before constructing an informative prior in HADO. The
question of great interest is then how large the sample size the informative priors should be based
on in order to gain efficiency in new measurements. A small sample may result in a prior either being too diffuse or potentially biased from outliers in a small sample. In short, the purpose of Experiment 2 was to investigate how the priors constructed from different sample sizes affect the estimation errors. To do that, we conducted experiments to measure CSF of observers in the H luminance condition using the priors constructed from the data with different sample sizes in the same H condition. As such, the only possible source of measurement differences in different conditions should come from the amount of information in the priors.

**Methods**

- Observers

A total of ten observers were recruited for the experiment, five of which had participated in Experiment 1. All observers had normal or corrected-to-normal vision, and were naive to the subject of the current study. Verbal consent was obtained before the experiment. The study protocol was approved by the institutional review board of human subjects research of the Ohio State University.

- Apparatus and stimuli

The apparatus and stimuli were the same as in Experiment 1

- Design and procedure:

Informative priors of varying sample sizes of observers were constructed using the data in the H luminance condition in the baseline experiment. In a clinical context, a practical approach would be to use the first-arriving observers to construct priors for subsequent observers. However, the degree of representativeness of such a sample to the population is totally random. In order to minimize such sampling bias and select a sample being moderately representative to the population, we devised the following sampling strategy for selecting a
sample of a given size from the original pool of 100 observers. We first selected a large number of subsets of observers with a given sample size, with each subset randomly selected without replacement (e.g., 1000 subsets of 5 observers), and then obtained a non-parametric (i.e., kernel density estimator based) estimate of the distribution of the CSF parameters for each subset. Separately, we obtained a similar distribution based on all 100 observers as a proxy for the underlying true population. We then calculated the distance between the mean of the sample distribution of each subset and that of the proxy population distribution, and also calculated the divergence of each sample distribution by the determinant of its variance-covariance matrix. Finally, we chose one sample whose bias and divergence were both closest to the median among all the samples. In this way, we selected three samples of differing sample sizes (i.e., n = 5, 12, and 30) from the 100 observers in the baseline experiment.

Figure 6 shows the 75% equal-density contours of the prior distributions of AULCSF and cutSF estimated from the observers of different sample sizes, and the total 100 observers, along with the diffuse prior. As reflected in the plot, with the increase of sample size, the prior distributions become more concentrated, and therefore contain more certainty about an average CSF in the corresponding population. On the other hand, the diffuse prior covers a much wider range of the parameter space, even though some part of its coverage is highly unlikely for a person with normal vision.

Each of the ten observers received a single session of 250 trials in the H luminance condition. Within the session, five independent quick CSF experiments (50 trials for each) were interleaved, each corresponding to the diffuse prior condition (i.e., sample size of zero) and the four informative prior conditions that differed in sample sizes of 5, 12, 30, and 100 observers. The stimuli presentation paradigm was the same as in Experiment 1.
Figure 6: Equal-density contours of the diffuse and informative priors of sample sizes 5, 12, 30 and 100 used in Experiment 2.

Results

Error measures were defined and computed in the same way as in Experiment 1 (i.e., RMSEs of estimated AULCSFs from the approximated true CSFs, as described in Eq. (8)). Figure 7a shows the comparison of estimation quality across the five different prior conditions. As expected, given any prior, the estimates gradually converged to a stable value after sufficient observations were made. Compared to the diffuse prior, all informative priors achieved smaller errors at the beginning of the experiments. There seem to be only small differences among the sample sizes, suggesting that even a very small sample (n = 5) can be quite effective in improving estimation.
Figure 7: RMSE plots of the estimation of AULCSF in Experiment 2 with the diffuse prior ($n=0$) and the informative priors of different sample sizes ($n=5, 12, 30$ and $100$).

As in Experiment 1, a simulation was also conducted to compare with the experimental data. To assess the random effects of the selection of different observers for priors, many priors were created from different selection of observers for each sample size. To do that, we drew 500 subsets of observers for each sample size of 5, 12 and 30, with each subset randomly sampled from the 100 observers without replacement. For each subset, an informative prior was constructed. Hence, the total number of priors for the simulation was $500 \times 3 + 2 = 1502$, with the additional two being the diffuse prior and the prior constructed from the total 100 observers. Ten simulated observers assumed to have the CSFs of the real observers. With each of the 1502 priors, each simulated observer went through the same experimental procedure as in the human experiment 50 times. The RMSE of estimation was calculated across the replications, priors and 10 observers for each sample-size condition.
The results of the simulation revealed a clearer effect of sample size on the quality of estimation (Figure 7b). All informative priors started with smaller errors than the diffuse prior. The general pattern showed that the larger the sample size, the smaller RMSE the estimation achieved. The advantage of using informative priors remained until the end of the session, even though the differences between priors diminished. As shown in Table 2, for the first 10 trials, the average reduction of RMSE by using a prior constructed from 5 existing observers compared to using a diffuse prior was 3.30 dB, 4.06 dB for the sample size of 12, 4.56 dB for the sample size of 30, and 4.72 dB for the sample size of 100. The average reduction of RMSE through 50 trials by using informative priors were smaller (see Table 2) because the effect of priors became less influential on later trials. As a measure of precision, we calculated the standard deviation of the RMSEs on each experimental trial across all subsets of the priors. For the first ten trials, the standard deviations of RMSE for the diffuse prior was 1.80 dB, and decreased to 1.24 dB, 0.74 dB, 0.46 dB and 0.35 dB for the priors constructed from the sample sizes of 5, 12, 30, and 100, respectively. After 50 trials, the difference of the variability among priors was smaller because the estimation becomes more stable when more observations were collected.

Several conclusions can be drawn from the results of Experiment 2. First, prior constructed from at least five observers is sufficient to provide significant improvement in CSF estimation of new observers. Even though a small sample may not accurately represent the population, our results showed that the prior constructed from a small sample was still relatively divergent (the blue line in in Figure 6), and therefore sampling error may not induce overly large bias to the priors. On the other hand, a diffuse prior (the yellow line in Figure 6) tends to be excessively divergent for the true human vision so that information in a small sample is still better than none. Second, priors constructed from larger sample size led to more accurate and
precise estimation, demonstrating the benefit of measurement when more prior knowledge is available. Third, the gain from increasing sample size stabilizes when the sample size is large enough. In the current study, a sample size of 30 appeared to perform comparably to the estimation with a sample size of 100.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>After Trial 10 (dB)</th>
<th>After Trial 50 (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-3.30</td>
<td>-1.25</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-4.06</td>
<td>-1.53</td>
</tr>
<tr>
<td>30</td>
<td>-4.56</td>
<td>-1.75</td>
</tr>
<tr>
<td>100</td>
<td>-4.72</td>
<td>-1.88</td>
</tr>
<tr>
<td>0</td>
<td>1.80</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>1.24</td>
<td>0.62</td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.74</td>
<td>0.48</td>
</tr>
<tr>
<td>30</td>
<td>0.46</td>
<td>0.38</td>
</tr>
<tr>
<td>100</td>
<td>0.35</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 2: The RMSE of AULCSF and the precision (variability) of estimation in HADO experiments with the diffuse prior and the informative priors constructed from the observers of sample sizes of 5, 12, 30 and 100. The values of precision are the standard deviations of RMSEs from replicated measurements.
Discussion

In vision research, we are often faced with the dilemma between the quality of measurement and the cost of additional data collection. The current study focused on a newly proposed method called hierarchical adaptive design optimization (HADO, Kim et al, 2014), which improves both efficiency and accuracy of the measurement by incorporating previous data as informative priors in the test. Using the case study of measuring CSF, both experiments and simulations showed that HADO can indeed yield more accurate and precise estimates than the existing adaptive methods given the same number of observations. The improvement from HADO is manifested to the greatest degrees in early trials of the experiment and is equivalent to the amount of information that estimation with a diffuse prior can acquire after 20 to 30 trials. HADO shows its advantage clearly even when a mixture prior constructed from heterogeneous populations is employed (Experiment 1) or when only a small number of observers contribute to the construction of priors (Experiment 2). To put these results in perspective, the current study demonstrates that the theoretical diffuse priors, which have been adopted to represent conservative prior beliefs and thus avoid inferential bias, can be replaced by statistically justifiable informative priors without committing unwanted bias to achieve the desired improvement in measurement.

There are many ways that HADO can be implemented and extended further than shown in the current study. HADO can potentially gain more efficiency if more covariates relevant to the to-be-measured functional vision characteristics can be included to build more sophisticated hierarchical models. For example, if CSF is assumed to vary linearly with a covariate $cov$, the second line in Eq. (3) can be formulated as $\theta | a, b, \sigma \sim N(a + b cov, \sigma^2)$, a normal distribution with its mean modeled by a linear regression with $cov$ as the regressor. Plausibly, the cost for
measuring these covariates could be very low. In the current study, individual variables such as
age, gender, visual acuity, and eyeglass prescriptions are likely to co-vary with the CSF
characteristics and are very easy to obtain. Therefore, even more efficiency could be gained
when these variables are included as covariates in a group-level model to determine more
specific priors for a new measurement.

As demonstrated in Experiment 1, a mis-specified prior may impose a large error on the
statistical inference of new observers. A mixture prior that represents a larger population is a
better choice when the group membership of a new observer is unknown. In the present study,
the weights of the components in a mixture prior (H, M and L) were chosen to be uniform.
Improvement can be made to change the weights of the components according to the known or
cheap-to-measure covariates.

The use of kernel density estimation (KDE) for prior construction (described in the
baseline experiment) may be considered a crude form of empirical Bayes methods, which
approximates a full Bayesian treatment of hierarchical modeling (Casella, 1985). It was a
necessary choice due to the prohibitive computational burden that would be on many instances of
Bayesian computations in real and simulated experiments in the current study. Technically, the
distribution resulting from KDE estimation is not precisely the same entity as the prior
distribution shown in Eq. (5) which represents a full Bayesian treatment. Despite the theoretical
difference between the two methods, however, our empirical Bayes priors, as an approximation,
were shown to attain similar improvement in subsequent measurements to a Gaussian prior
estimated from a full Bayesian hierarchical model.

Another approach to build a mixture prior is to fit all datasets across populations with
Bayesian nonparametric methods (Bush & MacEachern, 1996; Teh, Jordan, Beal & Blei, 2006;
Rodriguez, Dunson, & Gelfand, 2008) instead of estimating each component distribution separately and combining them with weights. The Bayesian nonparametric approach represents a theoretically better-grounded mechanism for constructing a mixture distribution without the assumption of any parametric form of the group-level distribution. However, its implementation demands more sophisticated statistical modeling and estimation techniques.

The advantage of hierarchical Bayesian modeling, an integral part of HADO, can also be further explored. The current study only focused on an application in which all observers take only a single kind of a vision test (i.e., CSF). There are many other functional vision characteristics, e.g., visual acuity, stereo acuity, Vernier acuity, binocular combination. Together with CSF, they define a more complete assessment of visual characteristics. Potential inferential benefit can be gained by using data from one test to construct informative priors for a different test. The solution is to build parallel hierarchical models for different tests, and link these tests with shared parameters in the higher-level distribution. This way, better measurements and inferences are made not only by having one person’s data inform another person’s test but also by having one kind of test data inform another kind. Ultimately, when this HBM approach is combined with model-based adaptive testing, a powerful system of comprehensive visual assessment could be established.

Besides CSF, the estimation of other visual psychophysical models such as threshold versus external noise contrast functions (TvC; Lesmes et al., 2006), sensory memory decay (Sperling, 1960; Lu, Neuse, Madigan, & Dosher, 2004), color matching ellipses (Wyszecki & Stiles, 1982) all face the same problem of efficient measurement. Bayesian adaptive methods were adopted for estimating these functions (Baek, Lesmes, & Lu, 2014; Kujala & Lukka, 2006;
Lesmes et al., 2010). Applying the current HADO procedure in these testing scenes would also gain benefits on testing efficiency.

In conclusion, the application of HADO in this behavioral study demonstrates a statistically justified way to incorporate information in previously collected data into a new test rather than starting with a purely non-informative prior. As an extension to the standard Bayesian adaptive testing method, HADO can be implemented with a moderate amount of modeling effort on top of the current adaptive testing framework, with a noticeable gain in efficiency. By combining the advantage of hierarchical Bayesian modeling and the adaptive testing procedure, HADO is a powerful and flexible statistical tool that can be applied for more realistic modeling and efficient measurement.
Acknowledgments

This research is supported by NIH grant R01-MH093838 to JIM and MAP and NIH grant R01-EY021553 to ZLL.

Disclosure

Luis Lesmes: Commercial Relationship(s), Adaptive Sensory Technology, Code I, E, P;
Zhong-Lin Lu: Commercial Relationship(s), Adaptive Sensory Technology, Code I, P; Rest of authors: None

References


