A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function

Hairong Gu Department of Psychology, The Ohio State University, Columbus, OH, USA

Woojae Kim Department of Psychology, The Ohio State University, Columbus, OH, USA; Present address: Department of Psychology, Howard University, Washington, DC, USA

Fang Hou Department of Psychology, The Ohio State University, Columbus, OH, USA

Luis Andres Lesmes Adaptive Sensory Technology, Boston, MA, USA

Mark A. Pitt Department of Psychology, The Ohio State University, Columbus, OH, USA

Zhong-Lin Lu Department of Psychology, The Ohio State University, Columbus, OH, USA

Jay I. Myung Department of Psychology, The Ohio State University, Columbus, OH, USA

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. Noninformative, diffuse priors were commonly used in those tests. Hierarchical adaptive design optimization (HADO; Kim, Pitt, Lu, Steyvers, & Myung, 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 represents an empirical validation of HADO in estimating the human contrast sensitivity function. The results show that HADO significantly improves the accuracy and precision of parameter estimates, and therefore requires many fewer observations to obtain reliable inference about contrast sensitivity, compared to the method of quick contrast sensitivity function (Lesmes, Lu, Baek, & Albright, 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 of this case study support the conclusion that HADO can be used in Bayesian adaptive testing by replacing noninformative, diffuse priors with statistically justified informative priors without introducing unwanted bias.

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

Citation: Gu, H., Kim, W., Hou, F., Lesmes, L. A., Pitt, M. A., Lu, Z.-L., & Myung, J. I. (2016). A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function. *Journal of Vision, 16*(6):15, 1–17, doi:10.1167/16.6.15.

doi: 10.1167/16.6.15 Received September 25, 2015; published April 22, 2016 ISSN 1534-7362

accurately measure how visual performance—e.g., detection threshold—varies within a multidimensional 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 preselected, fixed set of stimuli (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 and improve the efficiency and precision of vision testing, numerous adaptive testing methods have been developed.

Adaptive testing in vision

The nonparametric 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—i.e., the threshold—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 (Garcia-Perez, 2001; Kaernbach, 1991; Levitt, 1971), their theoretical underpinnings concerning optimality are not well grounded in probability theory.

A new generation of adaptive testing methods overcome these limitations by taking advantage of statistical models that capture the regularities of the underlying mechanism 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 still limited to the estimation of psychometric functions with one parameter, and further restricted by the strong assumption that the stimulus corresponding to the estimated threshold is always the optimal stimulus to be presented on the next trial, without rigorously formalizing the *usefulness* of each stimulus in improving statistical inferences.

A theoretically more sound adaptive procedure is a class of entropy-based Bayesian adaptive procedures, such as the psi method (Kontsevich & Tyler, 1999), the quick methods (Lesmes, Jeon, Lu, & Dosher, 2006; Lesmes, Lu, Baek, & Albright, 2010; Lesmes et al., 2015; Lu & Dosher, 2013), Bayesian adaptive estimation (Kujala & Lukka, 2006), adaptive design optimization (Cavagnaro, Myung, Pitt, & Kujala, 2010), and active data collection (DiMattina, 2015; DiMattina & Zhang, 2008, 2011). Formulated within 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 equivalently the uncertainty of the parameters. In general terms, the *utility function* measures the usefulness of a given stimulus choice, written in the following form:

$$U(s) = \int_y \int_\theta u(s, y, \theta) p(y|s, \theta) p(\theta)\, dy\, d\theta, \quad (1)$$

where $u(s, y, \theta)$, called the sample utility, is a function of stimulus s, observation y, and parameter θ; $p(y|s, \theta)$ is the statistical model; and $p(\theta)$ is the prior distribution of θ. For example, a psychometric function $p(y|s, \theta)$ describes the probability of correct response for a given stimulus s, with the parameter θ 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 θ and a potential response y. A particular specification of $u(s, y, \theta)$ is

$$u(s, y, \theta) = \log \frac{p(\theta|y, s)}{p(\theta)}, \quad (2)$$

in which $p(\theta)$ is the prior distribution of θ, and $p(\theta|y, s)$ is the posterior. Therefore, $u(s, y, \theta)$ can be interpreted as the reduction in the uncertainty about parameter θ after a new trial with a stimulus s and an observation y. By taking the integral of the sample utility over all possible observations y and parameters θ, the derived *expected utility* $U(s)$ in Equation 1 measures the *expected information gain* brought by the stimulus 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 in the response on the next trial.

The computation of Equations 1 and 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's 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 (Kont-

sevich & Tyler, 1999; Kujala & Lukka, 2006; Lesmes et al., 2006) or diffuse (Hou, Huang, Lesmes, Feng, Tao, Zhou, & Lu, 2010; Lesmes et al., 2010; Lesmes et al., 2015)—have been used.

One advantage of Bayesian statistics is that inferences can be enhanced by starting each new inference from previously collected observations, quantified as an informative prior (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Although the standard Bayesian adaptive testing methods we have reviewed adopt full Bayesian procedures for parameter estimation, the advantage of informative priors has yet to be utilized. Consequently, those 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 results from integrating hierarchical Bayesian modeling (HBM) with the standard entropy-based Bayesian adaptive testing procedures. In the work by 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 healthy eyes, using as a baseline the quick CSF (qCSF) procedure (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 qCSF by treating 146 subjects as being previously tested and the remaining subject as a new individual to be measured subsequently. The results showed that HADO achieved a decrease of around 2 dB (1 dB = 0.05 decimal log units) in the root-mean-square error (RMSE) in the estimation of CSF during the first 40 trials, and saved more than 20 trials to reach a 90% correct classification as an amblyopic eye versus a healthy eye, compared to the qCSF 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 further, that a sufficiently large number of observers would be available to represent the target population. The purpose of the present study is to empirically validate HADO by evaluating those assumptions in a case study of CSF measurement. 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 HBM (Bernardo & Smith, 1994; Lee, 2006; Rouder & Lu, 2005), a statistical technique that improves the precision of inferences by accounting for dependencies in data. Particularly, observers from the same population or who were tested in the same experimental condition are expected to perform more similarly to each other than to those from different populations or conditions. The innovative application of HBM in HADO is to extract 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 subroutine 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: (a) design optimization (finding the optimal stimuli), (b) measurement (presenting stimuli and collecting responses), and (c) Bayesian updating of prior to posterior. On trial t, the prior is expressed as $p\left(\theta_n|y_n^{(1:t-1)}\right)$, in which $y_n^{(1:t-1)}$ denotes responses in the previous trials, and n indexes observers. The utility $U(s_t)$ of a stimulus quantifies the expected reduction of entropy of parameters each stimulus can potentially bring, calculated by Equation 1. Then the optimal stimulus s_t^* corresponding to the maximal $U(s_t)$ is administered and a new response $y_n^{(t)}$ observed. The posterior distribution of parameters is calculated by Bayes's rule: $p\left(\theta_n|y_n^{(1:t)}\right) \propto p\left(y_n^{(t)}|\theta_n, s_t^*\right) p\left(\theta_n|y_n^{(1:t-1)}\right)$. The posterior $p\left(\theta_n|y_n^{(1:t)}\right)$ is subsequently used as the prior for the next trial $t+1$. These steps repeat until a given number of trials are executed or a given criterion for accuracy and precision of estimation is reached.

The standard Bayesian adaptive testing procedure leaves open the option of priors at the beginning of an experiment, to which HBM contributes. The upper loop of Figure 1 represents the HBM component of HADO that draws information from previously run observers to provide an informative prior for a new

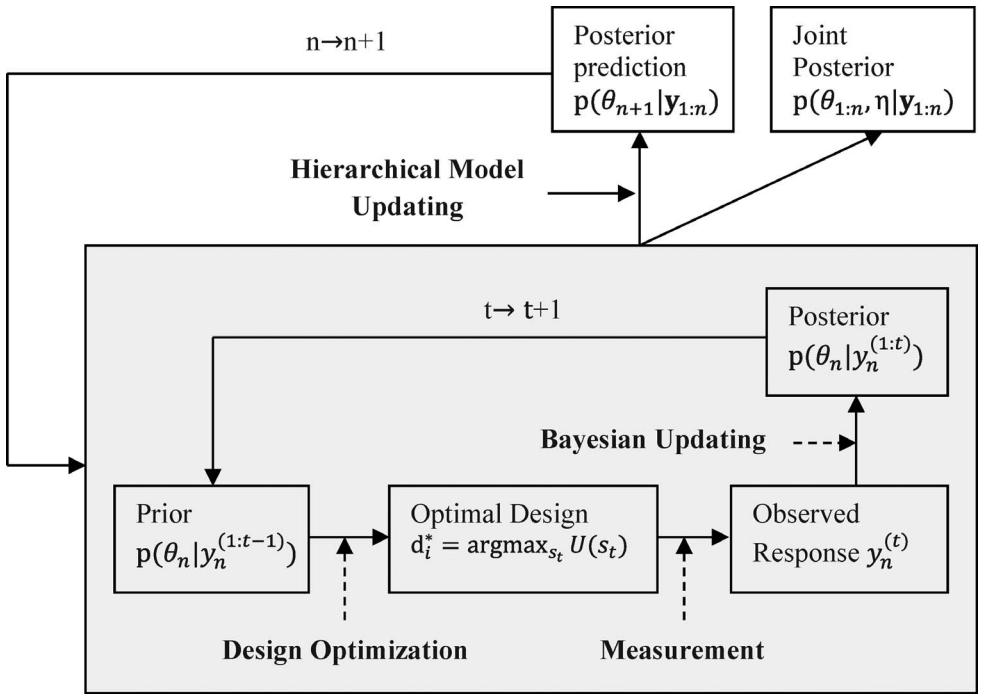

Figure 1. Illustration of the HADO algorithm. The standard Bayesian adaptive procedure—e.g., the psi method (the shaded area)—is an integral component of the HADO algorithm. See the text for additional details. Reprinted with permission from Kim, W., Pitt, M. A., Lu, Z. L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation,* 26, 2468.

observer. Given a data set of n observers, a hierarchical Bayesian model can be formulated as

$$\begin{cases} \eta \sim p(\eta) \\ \theta_i|\eta \sim p(\theta_i|\eta) \\ y|s, \theta_i \sim p(y|s, \theta_i) \end{cases} \qquad i = 1, \ldots, n. \quad (3)$$

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 a stimulus s and the ith observer's parameter θ_i. A middle-level model $p(\theta_i|\eta)$, in the second line, defines the dependency among individual-level parameters θ_i, conditional on the higher level parameter η 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 η would be the mean μ and variance Σ. Alternatively, a nonparametric model such as a kernel density estimator can be specified if the underlying distribution is believed to deviate significantly from standard parametric distributions. The first line of Equation 3 specifies the prior distribution of the higher level parameter η. When n observers' worth of data are collected with observed responses $y_{1:n}$ (the subscript "1:n" denotes a collection of observations from a total of n observers), the posterior distribution of the parameter η is obtained by

$$p(\eta|\ y_{1:n}) \propto \int\int \left[\prod_{i=1}^{n} p(y_i|\ \theta_i) p(\theta_i|\ \eta)\right] p(\eta)\, d\theta_1 \ \ldots \ d\theta_n. \quad (4)$$

Subsequently, the prediction of the parameter θ_{n+1} for a new observer is made by

$$p(\theta_{n+1}|\ y_{1:n}) = \int p(\theta_{n+1}|\eta) p(\eta|y_{1:n})\, d\eta. \quad (5)$$

It is important to note that $p(\theta_{n+1}|y_{1:n})$ serves as an informative prior for the next observer $n + 1$ 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 noninformative, diffuse prior.

The HADO algorithm is implemented through two subroutines—the standard Bayesian testing procedure (e.g., the qCSF method) and HBM—corresponding to the adaptive and the hierarchical component, respectively. In a typical implementation, the Bayesian testing procedure is performed online during an experiment to select an optimal stimulus on each trial, while HBM is performed off-line after an experiment, to update the population-level structure with existing data sets and construct a prior for a new observer.

In evaluating the performance of HADO, we compared HADO and the standard Bayesian testing procedure in an empirical case study of CSF measurement. The standard Bayesian testing procedure of CSF is reviewed in the next section.

Contrast sensitivity function (CSF)

The CSF describes how contrast sensitivity (reciprocal contrast threshold) 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 the CSF is of keen interest for the purpose of diagnosing visual deficits, because various visual pathologies are associated with characteristic changes in it (Hess, 1978; Jindra & Zemon, 1989; Marmor, 1981; Wolkstein, Atkin, & Bodis-Wollner, 1980). However, measuring CSFs typically requires a large number of observations, given the complex shape and the requirement of proper sampling in both contrast and spatial-frequency domains (Lesmes et al., 2010). To reduce the data-collection burden, adaptive testing methods have been extensively studied and exploited for measuring CSF (Dorr, Lesmes, Lu, & Bex, 2013; Hou et al., 2010; Hou, Lesmes, Bex, Dorr, & Lu, 2015; Lesmes et al., 2010).

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

$$S(f) = \begin{cases} \gamma^{\max} - \delta & \text{if } f < f^{\max} - \frac{\beta}{2}\sqrt{\frac{\delta}{\log_{10}2}}; \\ \gamma^{\max} - (\log_{10}2)\left(\frac{f - f^{\max}}{\beta/2}\right)^2 & \text{otherwise.} \end{cases} \quad (6)$$

Figure 2a illustrates the parametrization of the CSF in terms of the four parameters (peak sensitivity $\gamma^{\max}$, peak frequency $f^{\max}$, bandwidth β at half of the peak sensitivity, and low-frequency truncation level δ). The SCF $S(f)$ in Equation 6 is the reciprocal of the contrast threshold, corresponding to the level of contrast that is associated with a predefined performance level.

The psychometric function in the qCSF method is defined by a cumulative Gaussian function (Alcalá-Quintana & García-Pérez, 2004)

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

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 guess rate, L is the lapse rate, $\Phi(\cdot)$ is the cumulative Gaussian function, and σ determines the slope of the psychometric function. An example of a cumulative Gaussian psychometric function is shown in Figure 2b. The guess rate is the probability of making a correct response when the contrast of stimuli approximates zero. In an N-alternative, forced-choice (NAFC) task, the guess rate is assumed to be equal to $1/N$. Figure 2b shows a guess rate of 0.1 for a 10AFC task (Hou et al., 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 preset to a value obtained from previous studies (Hou et al., 2015).

Equations 6 and 7, combined as $f(y|c, f, \gamma^{\max}, f^{\max}, \beta, \delta, G, L, \sigma)$, mathematically describe how external stimulus variables c and f are transformed into underlying visual sensitivity, tuned by the parameters ($\gamma^{\max}, f^{\max}, \beta, \delta, G, L, \sigma$) specific to each observer being tested, and finally reflected as the probability of correct responses. The four parameters to be estimated in the present study are $\gamma^{\max}$, $f^{\max}$, β, and δ. The other three parameters G, L, and σ are fixed as $G = 0.1$, $L = 0.04$, and $\sigma = 0.42$ following Hou et al. (2015). Note that in a hierarchical modeling context, the model $f(y|c, f, \gamma^{\max}, f^{\max}, \beta, \delta)$ describes the individual-level data, corresponding to the third line in Equation 3.

The qCSF method, as with the other standard entropy-based Bayesian testing procedures, assumes a noninformative prior at the beginning of an experiment. In the present empirical validation study, we used the qCSF procedure as a benchmark to compare its performance with that of the HADO procedure.

Hierarchical adaptive estimation of the CSF

The soundness of HADO hinges upon the validity of two main assumptions it makes. First, the informative prior obtained based on 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 would contain mismatched information, giving a biased view of new observers. The second assumption of HADO is that to construct an

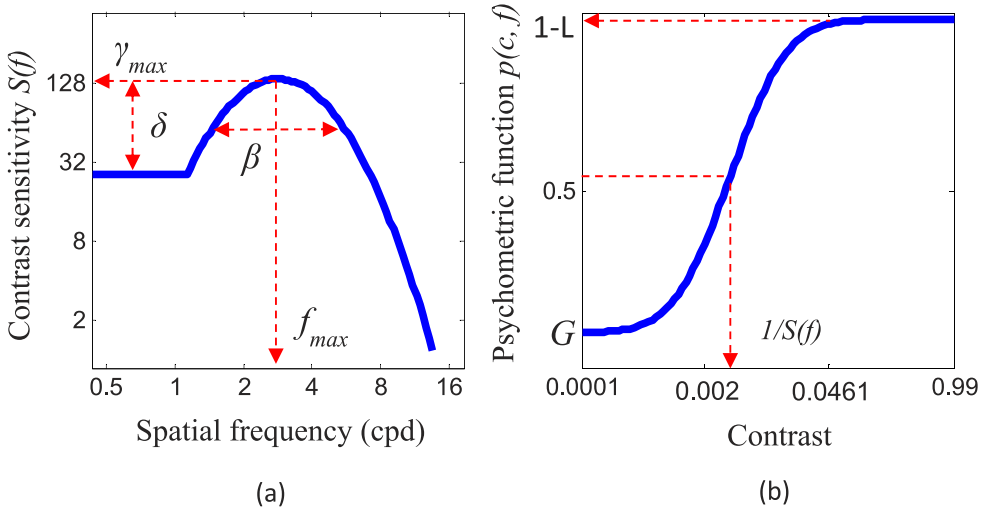

Figure 2. (a) Contrast sensitivity function $S(f)$ as a function of spatial frequency f, parameterized by peak sensitivity γ^{max}, peak frequency f^{max}, bandwidth β, and truncation level δ (Watson & Ahumada, 2005). (b) The psychometric function $p(c, f)$ for a 10AFC task that describes the probability of detecting stimuli of contrast c, with the threshold determined by $1/S(f)$.

informative prior, the size of the data sample (i.e., number of existing observers) should be large enough to well approximate the underlying population. If data are available from only a small number of observers, the resulting prior either may become too diffuse to be useful for an inference purpose or may lead to undesirable bias when these observers are not representative of the population.

One specific goal of the current study was to investigate the robustness of HADO when these two assumptions are violated to different degrees. While this question might be partly addressed by a simulation study, we opted to integrate such tests into the design of the current empirical validation. More specifically, we assumed neither true parameter values nor correct models for the purpose of data generation in both HADO calibration and validation phases (we need modeling assumptions for the purpose of estimation). This aspect had important implications. First, we believed that the benefit of HADO should be demonstrated by testing real, out-of-sample observers whose behavior is not necessarily predicted by the model being used for estimation. Second, since true CSFs are unknown in reality, we treated large-sample estimates as a ground truth against which the accuracy and reliability of parameter estimates under various, specific conditions were evaluated.

To conduct an empirical study on the robustness of HADO, we designed two validation experiments by manipulating the extent of agreement with the theoretical assumptions of HADO (i.e., sample representativeness and sample size). Data from a previously conducted baseline experiment, in which a large number of observers participated in the qCSF task under three different luminance conditions (reported separately by Hou et al., 2016), were used to build the informative priors for Experiments 1 and 2 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 several samples of different sizes were used to gauge how large a sample needs to be to construct an effective prior. In what follows, we begin with a brief overview of the baseline experiment upon which the two validation experiments were built.

Baseline experiment: Data collection for prior construction

In the baseline experiment (Hou et al., 2016), the CSFs of 112 observers were measured under three different luminance conditions using the standard Bayesian testing procedure (qCSF) in a 10AFC task (Hou et al., 2015; Lesmes et al., 2010). Test–retest analyses were carried out to investigate the precision for detecting changes in CSFs.

In the HADO validation experiments, we used the data from the baseline experiment to construct various forms of informative priors—i.e., $p(\theta|\eta)$ in Equation 3. In the following presentation, only the information relevant to the present study is described. For details of the baseline experiment, readers are directed to Hou et al. (2016).

Each naïve observer received six blocks of qCSF measurements in four different viewing conditions: low luminance (L), medium luminance (M), high luminance (H), and low pass (LP). In the H condition, observers viewed the display through uncovered goggles. In the

M condition, they viewed the display binocularly through goggles with neutral density filters with an attenuation factor of 0.67 decimal log units. In the L condition, they viewed the display through goggles fitted with the neutral density filters with a total attenuation factor of 1.56 decimal 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/m^2, respectively. The order of the test blocks was L, L, M, H, LP, H. The first L condition was used for observers to dark-adapt and practice the qCSF test, and the two H conditions were included to assess the test–retest reliability of the qCSF method. In each test block, the qCSF 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 min. For additional details about the procedure (i.e., use of a diffuse prior, adaptive stimuli selection, and Bayesian estimation), please refer to Hou et al. (2015).

In the present study, we used data from 100 of the 112 observers in the baseline experiment for prior construction, and only those data from the H, M, and L conditions. For each observer, the point estimates of the four parameters of the truncated log-parabola CSF model were computed by 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, 1992) 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 Equation 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 et al., 2006)—both of which are diagnostic measures of contrast sensitivity (Hou et al., 2010; Hou et al., 2015; Lesmes et al., 2010). The four-dimensional distributions of CSF parameters were thus 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 the CSFs in the H condition is located in the 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.

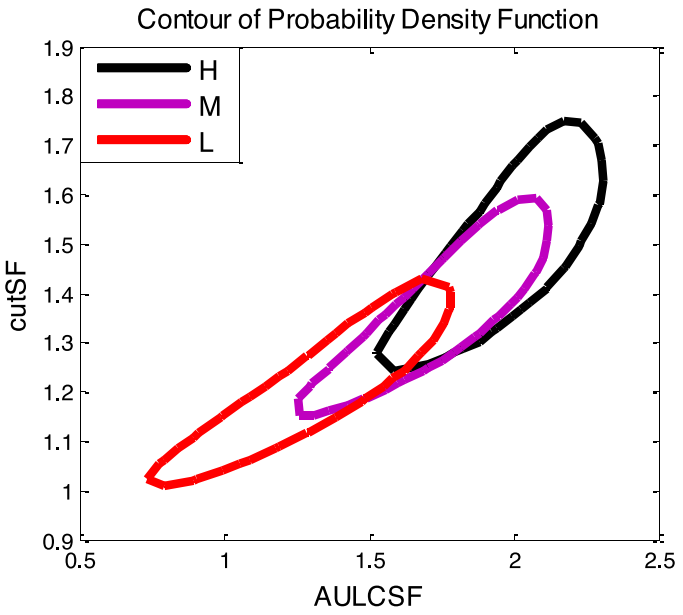

Figure 3. Equal-density contours of the estimated population distributions of AULCSF and cutSF under the H, M, and L luminance conditions estimated from the data of 100 observers in the baseline experiment.

Experiment 1: Effect of different types of priors

The relatively large sample size (100) of CSFs measured under each luminance condition in the baseline experiment can serve as a proxy for the corresponding population. The goal of Experiment 1 was to assess whether the use of informative priors, estimated from the baseline data, can help achieve greater efficiency in the estimation of CSF compared to the qCSF method, which assumes noninformative, diffuse priors, and if so, to determine the size of the benefit. The choice of priors is straightforward: Use the H prior for new observers in the H luminance condition, the L prior for observers in the L luminance condition, etc. An improvement of the estimation when the correct informative priors were applied would be expected naturally. In practice, however, 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 can be misspecified and 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 an informative prior constructed from the collapsed data from different populations, which may result in a prior still more informative than a noninformative diffuse prior. To investigate this

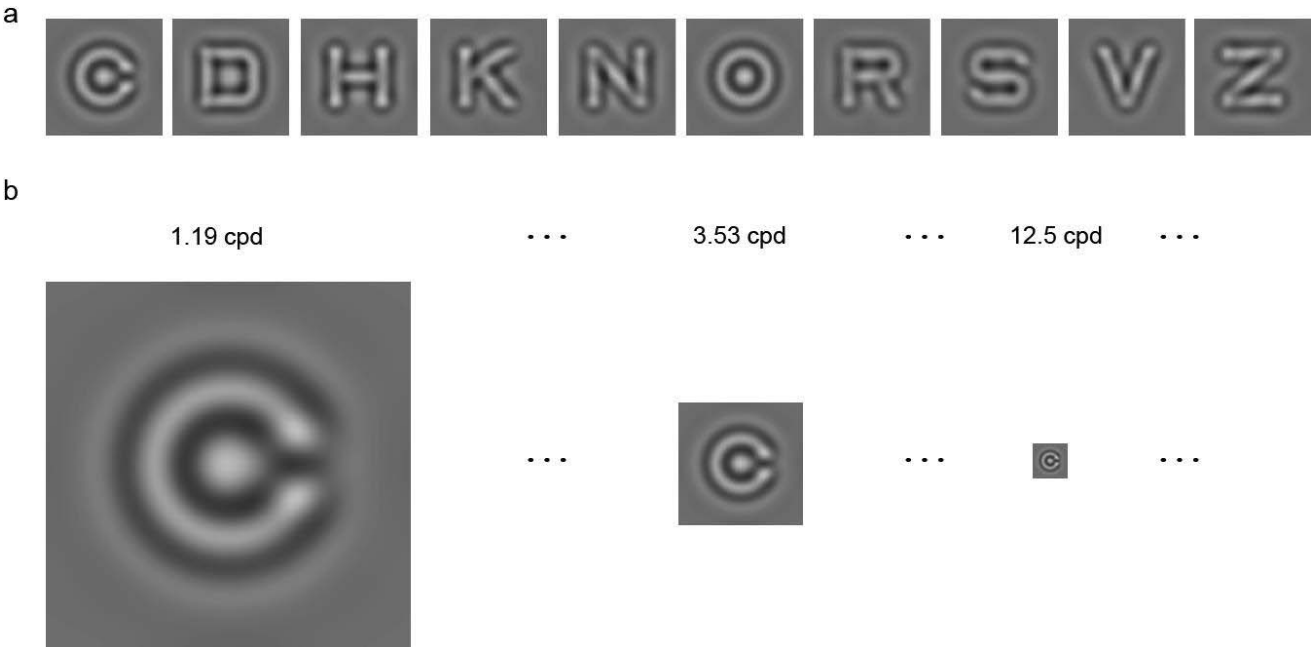

Figure 4. (a) Ten filtered letters. (b) Illustration of filtered letter "C" in various spatial-frequency conditions.

possibility, we assessed the influence of additional types of priors—i.e., misspecified and mixed-population priors—on HADO performance.

Methods

Observers: Ten 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 unaware of 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 (MathWorks, Natick, MA) with the Psychtoolbox subroutines (Kleiner et al., 2007) on a PC. Stimuli were presented on a gamma-corrected Samsung UN55FH6030 55-in. 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/m^2 (measured by a Tektronix J17 photometer). A bit-stealing algorithm was used to achieve 9-bit grayscale resolution (Tyler, 1997). Observers viewed the display binocularly from a distance of 4 m 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 and a bandwidth (of half height) of one octave. The filtered letters had a narrowband spectrum in the spatial-frequency domain and were found to assess contrast sensitivity in different central spatial frequencies equivalently to the conventional gratings (Alexander et al., 1994; McAnany & Alexander, 2006). 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 each image was 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 from 1.19 to 30.95 c/° for the qCSF procedure.
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 belief but still more informative belief about the parameters than a diffuse prior, which is noninformative by construct. The diffuse prior was the same as that used by Hou et al. (2015): a hyperbolic secant function close to being flat over the parameter space.

All stimuli in the experiment were selected from the ten Sloan letters. Observers were instructed to familiarize themselves with the letter set before the start of the experiment and respond only with letters from the set during the test. The task was therefore a 10AFC identification task. To improve user experience, three letter stimuli with the same central spatial frequency were presented in the same row with a center-to-center distance of 1.1 times the letter size in each trial. Each letter was randomly chosen with replacement from among the ten Sloan letters. The contrast of the letter on the right was determined by the qCSF algorithm, and the contrasts of the letters on the left and in the middle were 2 and 4 times of the optimal contrast, with an upper limit of 0.9. Observers were asked to identify the three letters one by one among the 10 alternatives.

Each observer received 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 qCSF experiments with different priors were conducted simultaneously, with stimuli from different prior conditions interleaved. Each qCSF experiment consisted of 50 trials, and therefore in total each observer performed 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 (Hou et al., 2016). 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 in the spatial-frequency dimension (evenly spaced on a base-10 log scale from 1.19 to 30.95 c/°).

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's 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 trials and transformed to the AULCSF summary statistic. To assess the quality of the estimates, the RMSE was calculated for each experimental trial ($j = 1, \ldots, 50$) by

$$\mathrm{RMSE}_j = 20 \times \sqrt{\frac{1}{10}\sum_{i=1}^{10}(\hat{\theta}_{i,j} - \hat{\theta}_{i,T})^2} \qquad j = 1, \ldots, 50, \tag{8}$$

where $\hat{\theta}_{i,T}$ denotes the approximation of the true AULCSF of the ith observer estimated as already described in a given luminance condition, and $\hat{\theta}_{i,j}$ is the estimate of the AULCSF of the ith observer on the jth trial. The constant 20 is multiplied to scale the results in decibel (dB) units, given that the parameter values are base-10 logarithms. 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).

Figure 5a and c (left-hand column) shows the RMSE profiles over trials based on the qCSF procedure 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. The effect of priors is evident in the graphs, and largest at the beginning of the experiment. When priors match 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 misspecified 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 misspecified H prior fares better. This can be explained by the general asymmetric shape of these prior distributions, which are skewed toward poor vision, as revealed in Figures 3 and 6. This enables the H prior to have better

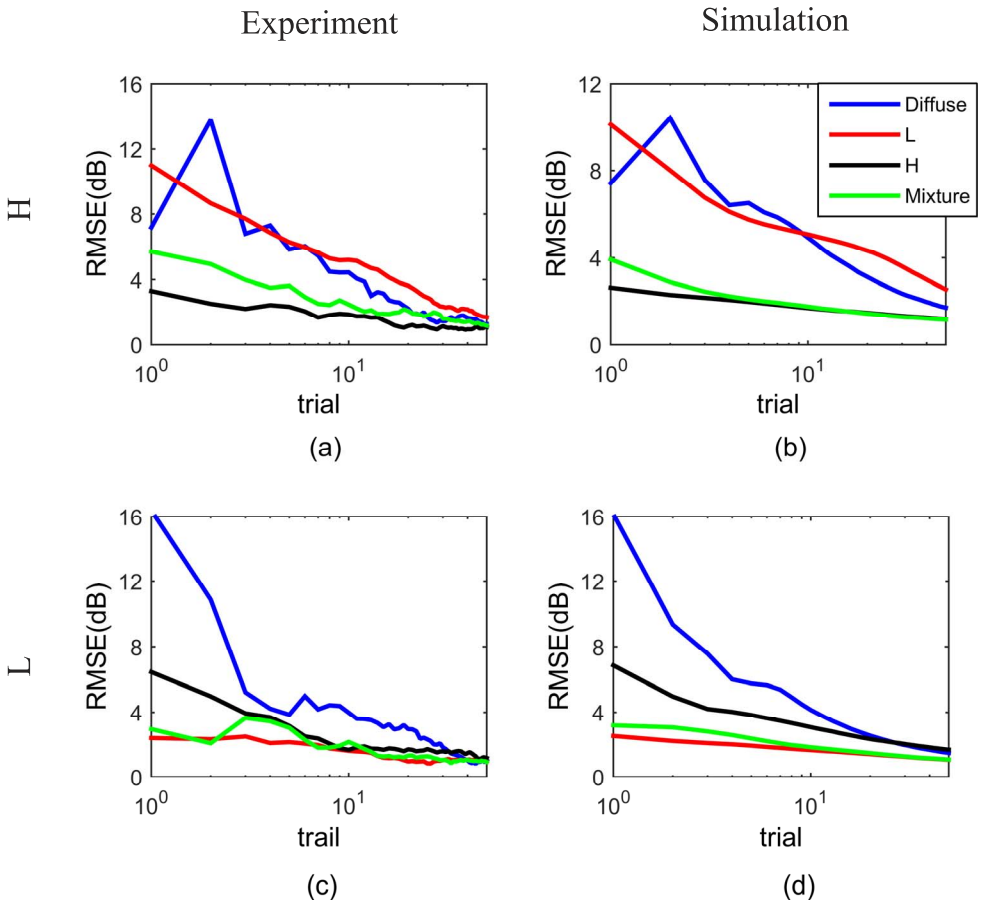

Figure 5. RMSE plots of AULCSF estimation in Experiment 1 with the four different types of priors under the H (a–b) and L (c–d) luminance conditions.

coverage of the L CSFs than the L prior does of 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.[1] 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 5b and d (right-hand column) gives 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 in the left panel, but much smoother due to the large number of replications in each condition.

We note an interesting pattern observed under the diffuse prior in the H condition (Figure 5a, b). The RMSE started small but became larger in the first few trials before decreasing in later trials. We believe this abnormal pattern is pure coincidence and of no theoretical interest. The pattern is mostly likely due to the fact that the mean of the diffuse prior just happens to be close to the true CSFs, thereby yielding a smaller RMSE at the beginning of the experiment. The subsequent large RMSEs are driven by the posterior updating with the initial observations. Note that the abnormal pattern is not observed under other priors in the H condition, nor any prior in the L condition, because the means of those priors did not happen to be close to the true CSFs. The phenomenon can also be observed in Experiment 2. The relationship of the diffuse prior and the true CSFs can be observed next in Figure 6, in which the CSFs in the H condition are closer to the mean of the diffuse prior than in the other conditions.

Table 1 shows the average reduction of RMSEs by using informative priors from using the diffuse prior based on the simulation data. 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 produces reductions of 4.29 and 4.54 dB in the H and L luminance conditions, respectively. By

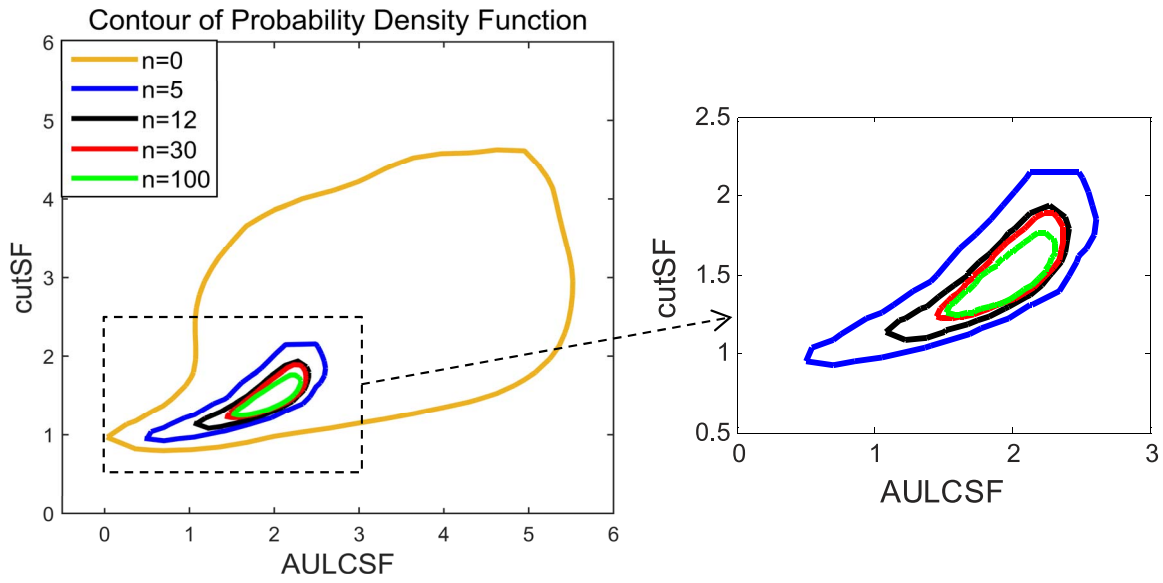

Figure 6. Equal-density contours of the diffuse and informative priors of sample sizes 5, 12, 30, and 100 used in Experiment 2.

contrast, the use of misspecified priors (the L prior in the H condition and the H prior in the L condition) results in reductions of only 0.28 and 2.88 dB 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 misspecified 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.

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 misspecified 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 of 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 noninformative diffuse priors.

Condition	Prior	After trial 10 (dB)	After trial 50 (dB)
H	Diffuse	0	0
	H	−4.58	−1.90
	Mixture	−4.29	−1.86
	L	−0.28	0.78
L	Diffuse	0	0
	H	−2.88	−0.62
	Mixture	−4.54	−1.59
	L	−4.99	−1.71

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.

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 Equation 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. Because a

small sample contains limited information about its population, the prior constructed from the sample may still be somewhat diffuse, or potentially biased if there are outliers in the sample. The question of greatest interest is then how large the sample size should be in order to achieve sufficient efficiency, or whether a prior constructed from a relatively small sample will be even worse off. The goal of Experiment 2 was thus to investigate how the priors constructed from different sample sizes affect the estimation errors. The experiment measured CSFs of observers in the H luminance condition using the priors constructed from the data of different sample sizes in the same H condition. Therefore, the only possible source of measurement differences across conditions should be the amount of information in the priors.

Methods

Observers: A total of 10 observers were recruited for the experiment, five of whom had participated in Experiment 1. All observers had normal or corrected-to-normal vision and were unaware of 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 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 would be totally random. In order to minimize such sampling bias and select a sample that was moderately representative of 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 of a given size, with each subset randomly selected without replacement (e.g., 1,000 subsets of five observers), and then obtained a kernel density 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 measured the representativeness of the KDE sample distribution constructed from each subset by its bias and divergence compared to the proxy population distribution. The bias was calculated by the distance between the mean of the sample distribution and that of the proxy population distribution. The divergence was calculated by the determinant of the variance–covariance matrix of the sample distribution. We found that with the increase in sample size, the average bias and divergence of the sample distributions gradually decreased. Finally, we chose one sample of a given sample size 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 ($n = 5$, 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. With the increase of sample size, the prior distributions become more concentrated and therefore contain more certainty about the 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 10 observers received a single session of 250 trials in the H luminance condition. Within the session, five independent qCSF experiments (50 trials for each) were interleaved, each corresponding to the diffuse prior condition (i.e., sample size of 0) and the four informative prior conditions (sample sizes of 5, 12, 30, and 100). The stimulus-presentation paradigm was the same as in Experiment 1.

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 Equation 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.

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 selections 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 $500 \times 3 + 2 = 1{,}502$, with the additional two being the diffuse prior and the prior constructed from the total 100 observers. Ten simulated observers were assumed to

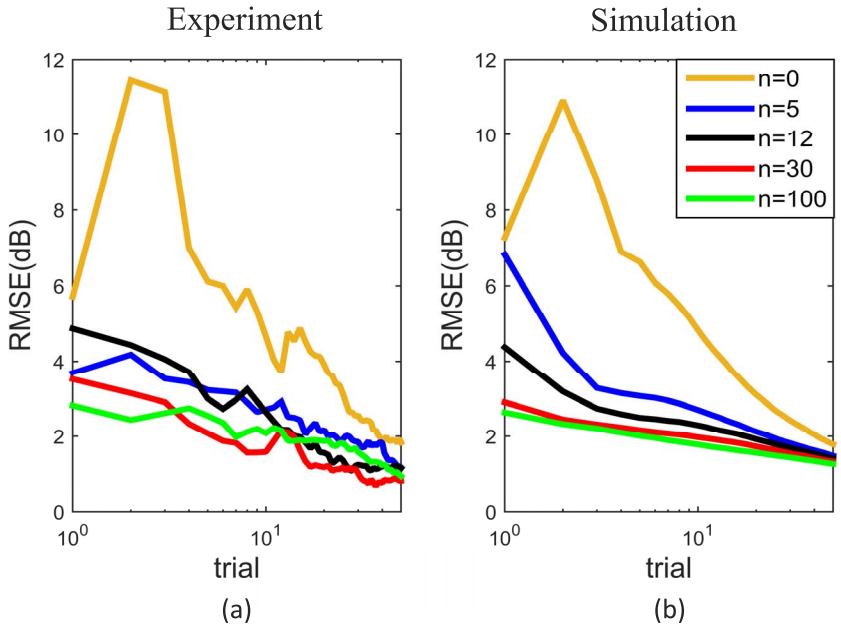

Figure 7. RMSE plots of the estimation of AULCSF in Experiment 2 with the diffuse prior ($n = 0$) and the informative priors of different sample sizes ($n = 5$, 12, 30, and 100).

have the CSFs of the real observers. With each of the 1,502 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 five existing observers compared to using a diffuse prior was 3.30 dB; it was 4.06 dB for the sample size of 12; it was 4.56 dB for the sample size of 30; and it was 4.72 dB for the sample size of 100. The average reduction of RMSE through 50 trials by using informative priors was 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 10 trials, the standard deviations of RMSE for the diffuse prior was 1.80 dB, decreasing to 1.24, 0.74, 0.46, and 0.35 dB, respectively, for the priors constructed from the sample sizes of 5, 12, 30, and 100. After 50 trials, the difference of the variability among priors was smaller because the estimation became more stable when more observations were collected.

Several conclusions can be drawn from the results of Experiment 2. First, the 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 orange line in Figure 6) tends to be excessively divergent from true human vision, so that information in a small sample is still better than none. Second, priors

Sample size	After trial 10 (dB)	After trial 50 (dB)
RMSE		
0	0	0
5	−3.30	−1.25
12	−4.06	−1.53
30	−4.56	−1.75
100	−4.72	−1.88
Precision		
0	1.80	0.88
5	1.24	0.62
12	0.74	0.48
30	0.46	0.38
100	0.35	0.31

Table 2. The RMSE of AULCSF and the precision (variability; standard deviation of RMSE from replicated measurements) 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.

constructed from larger samples 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.

Discussion

In vision research, we are often faced with the dilemma that increasing the quality of measurement incurs 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 measurement by incorporating previous data as informative priors in the test. In the present study of CSF measurement, both experiments and simulations showed that HADO yields 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. The advantages of HADO manifest 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 present case study of CSF measurement provides a validation of HADO by showing 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 to achieve the desired improvement in measurement without incorporating unwanted bias.

A major goal of visual assessment in clinics is the diagnosis of patients with visual deficits. As demonstrated in our study, the distributions of the CSFs in different luminance conditions are heterogeneous, which is also the case for people with normal and abnormal CSFs (Hou et al., 2010). One could therefore use the estimated CSFs for diagnostic purposes. For instance, logistic regression has been employed to classify CSFs into two clinical populations (Hou et al., 2010; Kim et al., 2014). Going forward, we could also apply other classification tools such as support vector machines (Cristianini & Shawe-Taylor, 2000) to optimize such clinical diagnosis based on the parameter estimates from HADO or any other adaptive testing procedures.

It is important to note that as a general-purpose adaptive experimental method, HADO can also be applied in a straightforward way to other visual psychophysical functions—such as threshold-versus-external-noise contrast functions (Lesmes et al., 2006), sensory memory decay (Baek, Lesmes, & Lu, 2014; Lu, Neuse, Madigan, & Dosher, 2005; Sperling, 1960), and color-matching ellipses (Wyszecki & Stiles, 1982)—which all face the same problem of efficient measurement. While we note that the standard Bayesian testing procedure has already been adopted for estimating these functions (Baek et al., 2014; Kujala & Lukka, 2006; Lesmes et al., 2010), the HADO procedure is expected to provide additional gains in efficiency in these testing scenarios.

There are several ways in which HADO can be extended to further improve the measurement efficiency and accuracy that are demonstrated in the present study. To illustrate, any covariates relevant to the functional vision characteristics to be measured can be incorporated into the informative prior as part of a more sophisticated hierarchical model. For example, if CSF is assumed to vary linearly with a covariate cov, the second line in Equation 3 can be formulated as $\theta|a, b, \sigma \sim N(a + b\text{cov}, \sigma^2)$, a normal distribution with its mean modeled by a linear regression with cov as the regressor. Plausibly, the cost of 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 covary with the CSF characteristics and are very easy to obtain. In that way, even more efficiency could be gained when these variables are included as covariates in a group-level model to obtain a more informative specific prior for new measurements.

As demonstrated in Experiment 1, a misspecified 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 by changing the weights of the components according to the known or cheaply measurable covariates.

The use of kernel density estimation (KDE) for prior construction (described in the baseline experiment) may be considered a crude form of empirical Bayesian 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 accrue with many instances of Bayesian computations in real and simulated experiments in the current study. Technically, the distribution resulting from KDE is not precisely the same entity as the prior distribution shown in Equation 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 building a mixture prior is to fit all data sets across populations with Bayesian nonparametric methods (Bush & MacEachern, 1996; Rodriguez, Dunson, & Gelfand, 2008; Teh, Jordan, Beal, & Blei, 2006) 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 component 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 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 inform 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.

In conclusion, the present empirical validation study of HADO demonstrates a statistically justified way to incorporate information from previously collected data into a new test rather than relying on a noninformative 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. Through its combination of the advantages of hierarchical Bayesian modeling and the Bayesian adaptive testing method, HADO is a powerful and flexible statistical tool that can be applied for more realistic modeling and more robust and efficient measurement.

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

Acknowledgments

This research is supported by NIH grant R01-MH093838 to JIM and MAP and NIH grant R01-EY021553 to Z-LL. The authors wish to thank the reviewers for their valuable feedback and comments on an earlier version of this article.

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

Commercial relationships: none.
Corresponding author: Hairong Gu.
Email: gu.124@osu.edu.
Address: Department of Psychology, The Ohio State University, Columbus, OH, USA.

Footnote

[1] The executable MATLAB programs of the simulations are available for download from http://faculty.psy.ohio-state.edu/myung/personal. Readers can also gain access to the source code after they sign a software license agreement with The Ohio State University.

References

Alcalá-Quintana, R., & García-Pérez, M. A. (2004). The role of parametric assumptions in adaptive Bayesian estimation. *Psychological Methods, 9*(2), 250–271.

Alexander, K. R., Xie, W., & Derlacki, D. J. (1994). Spatial frequency characteristics of letter identification. *Journal of the Optical Society of America A, 11*, 2375–2382.

Applegate, R. A., Howland, H. C., Sharp, R. P., Cottingham, A. J., & Yee, R. W. (1997). Corneal aberrations and visual performance after radial keratotomy. *Journal of Refractive Surgery, 14*(4), 397–407.

Baek, J., Lesmes, L., & Lu, Z. L. (2014), Bayesian adaptive estimation of the sensory memory decay function: The quick partial report method. *Journal of Vision, 14*(10): 157, doi:10.1167/14.10.157. [Abstract]

Bernardo, J. M., & Smith, A. F. (1994). *Bayesian theory*. Chichester, UK: Wiley.

Bush, C. A., & MacEachern, S. N. (1996). A semi-

parametric Bayesian model for randomized block designs. *Biometrika, 83*, 275–285.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician, 39*(2), 83–87.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation, 22*(4), 887–905.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press.

DiMattina, C. (2015). Fast adaptive estimation of multi-dimensional psychometric functions. *Journal of Vision, 15*(9):5, 1–20, doi:10.1167/15.9.5. [PubMed] [Article]

DiMattina, C., & Zhang, K. (2008). How optimal stimuli for sensory neurons are constrained by network architecture. *Neural Computation, 20*(3), 668–708.

DiMattina, C., & Zhang, K. (2011). Active data collection for efficient estimation and comparison of nonlinear neural models. *Neural Computation, 23*(9), 2242–2288.

Dixon, W. J., & Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association, 43*(241), 109–126.

Dorr, M., Lesmes, L. A., Lu, Z. L., & Bex, P. J. (2013). Rapid and reliable assessment of the contrast sensitivity function on an iPad. *Investigative Ophthalmology & Visual Science, 54*(12), 7266–7273.

Garcia-Perez, M. A. (2001). Yes-no staircases with fixed step size: Psychometric properties and optimal setup. *Optometry and Vision Science, 78*(1), 56–64. [PubMed]

Ginsburg, A. P. (1981). Spatial filtering and vision: Implications for normal and abnormal vision. In L. Proenz, J. Enoch, & A. Jampolsky (Eds.), *Clinical applications of visual psychophysics* (pp. 70–106). Cambridge, UK: Cambridge University Press.

Ginsburg, A. P. (2003). Contrast sensitivity and functional vision. *International Ophthalmology Clinics, 43*(2), 5–15.

Hess, R. F. (1978). Contrast sensitivity assessment of functional amblyopia in humans. *Transactions of the Ophthalmological Societies of the United Kingdom, 99*(3), 391–397.

Hess, R. F. (1981). Application of contrast-sensitivity techniques to the study of functional amblyopia. In L. Proenz, J. Enoch, & A. Jampolsky (Eds.), *Clinical applications of visual psychophysics* (pp. 11–41). Cambridge, UK: Cambridge University Press.

Hou, F., Huang, C. B., Lesmes, L., Feng, L. X., Tao, L., Zhou, Y. F., & Lu, Z.-L. (2010). qCSF in clinical application: Efficient characterization and classification of contrast sensitivity functions in amblyopia. *Investigative Ophthalmology & Visual Science, 51*(10), 5365–5377. [PubMed] [Article]

Hou, F., Lesmes, L., Bex, P., Dorr, M., & Lu, Z.-L. (2015). Using 10AFC to further improve the efficiency of quick CSF method. *Journal of Vision, 15*(9):2, 1–18, doi:10.1167/15.9.2. [PubMed] [Article]

Hou, F., Lesmes, L., Kim, W., Gu, H., Pitt, M., Myung, J., & Lu, Z.-L. (2016). The usefulness of the quick CSF method: A large sample study. Manuscript submitted for publication.

Hou, F., Lu, Z.-L., & Huang, C. B. (2014). The external noise normalized gain profile of spatial vision. *Journal of Vision, 14*(13):9, 1–14, doi:10.1167/14.13.9. [PubMed] [Article]

Huang, C., Tao, L., Zhou, Y., & Lu, Z. L. (2007). Treated amblyopes remain deficient in spatial vision: A contrast sensitivity and external noise study. *Vision Research, 47*(1), 22–34.

Jindra, L. F., & Zemon, V. (1989). Contrast sensitivity testing: A more complete assessment of vision. *Journal of Cataract & Refractive Surgery, 15*(2), 141–148.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics, 49*, 227–229.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics, 29,* 41–59.

Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation, 26*, 2463–2492.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research, 34*(7), 885–912.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1–16.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research, 39*(16), 2729–2737.

Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology, 50*(4), 369–389.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science, 30*(3), 1–26.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

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*(19), 3160–3176.

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

Lesmes, L. A., Lu, Z.-L., Baek, J., Tran, N., Dosher, B. A., & Albright, T. D. (2015). Developing Bayesian adaptive methods for estimating sensitivity thresholds (d′) in Yes-No and forced-choice tasks. *Frontiers in Psychology, 6,* 1070, http://doi.org/10.3389/fpsyg.2015.01070.

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

Lu, Z.-L., & Dosher, B. A. (2013). *Visual psychophysics: From laboratory to theory*. Cambridge, MA, MIT Press.

Lu, Z.-L., Neuse, J., Madigan, S., & Dosher, B. A. (2005). Fast decay of iconic memory in observers with mild cognitive impairments. *Proceedings of the National Academy of Sciences, USA, 102*(5), 1797–1802.

Marmor, M. F. (1981). Contrast sensitivity and retinal disease. *Annals of Ophthalmology, 13*(9), 1069–1071.

McAnany, J. J., & Alexander, K. R. (2006). Contrast sensitivity for letter optotypes vs. gratings under conditions biased toward parvocellular and magnocellular pathways. *Vision Research, 46*(10), 1574–1584.

Oshika, T., Okamoto, C., Samejima, T., Tokunaga, T., & Miyata, K. (2006). Contrast sensitivity function and ocular higher-order wavefront aberrations in normal human eyes. *Ophthalmology, 113*(10), 1807–1812.

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Attention, Perception, & Psychophysics, 28*(4), 377–379.

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association, 103*, 1131–1154.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs, 74,* 1–29.

Taylor, M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America, 41,* 782–787.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*, 1566–1581.

Tyler, C. W. (1997). Colour bit-stealing to enhance the luminance resolution of digital displays on a single pixel basis. *Spatial Vision, 10*(4), 369–377.

Wackerly, D., Mendenhall, W., & Scheaffer, R. (2007). *Mathematical statistics with applications*. Belmont, CA: Cengage Learning.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses*. (pp. 181–207). New York: Springer.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision, 5*(9):6, 717–740, doi:10.1167/5.9.6. [PubMed] [Article]

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120.

Wolkstein, M., Atkin, A., & Bodis-Wollner, I. (1980). Contrast sensitivity in retinal disease. *Ophthalmology, 87*(11), 1140–1149.

Wyszecki, G., & Stiles, W. S. (1982). *Color science: Concepts and methods, quantitative data and formulae* (2nd ed.). New York: Wiley-Interscience.

Zhou, Y., Huang, C., Xu, P., Tao, L., Qiu, Z., Li, X., & Lu, Z. L. (2006). Perceptual learning improves contrast sensitivity and visual acuity in adults with anisometropic amblyopia. *Vision Research, 46*(5), 739–750.