A Hierarchical Adaptive Approach to Optimal Experimental Design

Woojae Kim\textsuperscript{1}, Mark A. Pitt\textsuperscript{1}, Zhong-Lin Lu\textsuperscript{1}, Mark Steyvers\textsuperscript{2}, and Jay I. Myung\textsuperscript{1}

\textsuperscript{1}Department of Psychology, Ohio State University, Columbus, OH 43210
\textsuperscript{2}Department of Cognitive Sciences, University of California, Irvine, CA 92697

\textbf{Keywords:} optimal experimental design, hierarchical Bayes, mutual information, visual spatial processing

\textbf{Abstract}

Experimentation is at the core of research in the behavioral and neural sciences, yet observations can be expensive and time-consuming to acquire. A major interest of researchers is designing experiments that lead to maximal accumulation of information about the phenomenon under study with the fewest possible number of observations. In addressing this challenge, statisticians have developed adaptive design optimization methods. This paper introduces a hierarchical Bayes extension of adaptive design optimization that provides a judicious way to exploit two complementary schemes of inference (with past and future data) to achieve even greater accuracy and efficiency in information gain. We demonstrate the method in a simulation experiment in the field of visual perception.

\section{Introduction}

Accurate measurement is essential in the behavioral and neural sciences to ensure proper model inference. Efficient measurement in experimentation can also be critical when observations are costly (e.g., MRI scan fees) or time-consuming, such as requiring hundreds of observations from an individual to measure sensory (e.g., eyes, ears) abilities or weeks of training (e.g., mice). The field of design optimization (Atkinson & Donev, 1992) pursues methods of improving both, with \textit{adaptive} optimization (e.g., DiMattina & Zhang, 2008, 2011) being one of the most promising approaches to date. These adaptive design optimization (ADO) methods capitalize on the sequential nature of experimentation by seeking to gain as much information as possible from data across the testing session. Each new measurement is made using the information learned from
previous measurements of an organism so as to achieve maximal gain of information about the processes and behavior under study.

Hierarchical Bayesian modeling (HBM) is another approach to increasing the efficiency and accuracy of inference (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Jordan, 1998; Koller & Friedman, 2009; Rouder & Lu, 2005). It seeks to identify structure in the data-generating population (e.g., the kind of groups to which an individual belongs) in order to infer properties of an individual given the measurements provided. It is motivated by the fact that data sets, even if not generated from an identical process, can contain information about each other. Hierarchical modeling provides a statistical framework for fully exploiting such mutual informativeness.

These two inference methods, ADO and HBM, seek to take full advantage of two different types of information, future and and past data, respectively. Because both can be formulated in a Bayesian statistical framework, it is natural to combine them to achieve even greater information gain than either alone can provide. Suppose, for instance, that one has already collected data sets from a group of participants in an experiment measuring risk tolerance, and data are about to be collected from another person. A combination of HBM and ADO allows the researcher to take into account the knowledge gained about the population in choosing optimal designs. The procedure should propose designs more efficiently for the new person than ADO alone, even when no data for that person have been observed.

Despite the intuitive appeal of this dual approach, to the best of our knowledge, a general, fully Bayesian framework integrating the two methods has not been published. In this letter, we provide one. In addition, we show how each method and their combination contribute to gaining the maximum possible information from limited data, in terms of Shannon entropy, in a simulation experiment in the field of visual psychophysics.

2 Paradigm of Adaptive Design Optimization (ADO)

The literature on optimal experimental design goes back to the pioneering work in the 1950s and 1960s in statistics (Lindley, 1956; Kiefer, 1959; Box & Hill, 1967). The recent surge of interest in this field (e.g., Atkinson & Donev, 1992; Chaloner & Verdinelli, 1995; Amzal, Bois, Parent, & Robert, 2006) can be attributed largely to the advent of statistical computing, which has made it possible to solve more complex and a wider range of optimization problems, and in some cases do so in real-time experiments. ADO is gaining traction in neuroscience (Paninski, 2003, 2005; Lewi, Butera, & Paninski, 2009; DiMattina & Zhang, 2008, 2011), and a growing number of labs are applying it in various areas of psychology and cognitive science, including retention memory (Cavagnaro, Pitt, & Myung, 2009; Cavagnaro, Myung, Pitt, & Kujala, 2010), decision making (Cavagnaro, Gonzalez, Myung, & Pitt, 2013; Cavagnaro, Pitt, Gonzalez, & Myung, in press), psychophysics (Kujala & Lukka, 2006; Lesmes, Jeon, Lu, & Dosher, 2006), and the development of numerical representation (Tang, Young, Myung, Pitt, & Opfer, 2010). In what follows, we provide a brief overview of the ADO framework.

ADO is formulated as a Bayesian sequential optimization algorithm that is executed over the course of an experiment. Specifically, on each trial of the experiment, on the basis of the present state of knowledge (prior) about the phenomenon under study,
Figure 1: Schematic illustration of the steps involved in adaptive design optimization (ADO; shaded region only) and hierarchical ADO (HADO; whole diagram). See text for further details.

which is represented by a statistical model of data, the optimal design with the highest expected value of a utility function (defined below) is identified. The experiment is then carried out with the optimal design, and measured outcomes are observed and recorded. The observations are subsequently used to update the prior to the posterior using Bayes’ theorem. The posterior in turn is used to identify the optimal design for the next trial of the experiment. As depicted in the shaded region of Figure 1, these alternating steps of design optimization, measurement, and updating of the individual-level data model are repeated in the experiment until a suitable stopping criterion is met.

In formal statistical language, the first, design optimization step of ADO entails finding the experimental design (e.g., stimulus) that maximizes a utility function of the following form (Chaloner & Verdinelli, 1995; Nelson, McKenzie, Cottrell, & Sejnowski, 2011; Myung, Cavagnaro, & Pitt, 2013):

\[
U(d_t) = \int \int \left[ \log \frac{p(\theta|y^{(1:t)}, d_t)}{p(\theta|y^{(1:t-1)})} \right] p(y^{(t)}|\theta, d_t) p(\theta|y^{(1:t-1)}) \, dy^{(t)} \, d\theta,
\]

(1)

where \( \theta \) is the parameter of a data model (or measurement model) that predicts observed data given the parameter, and \( y^{(1:t)} \) is the collection of past measurements made from the first to \( (t-1) \)-th trials, denoted by \( y^{(1:t-1)} \), plus an outcome, \( y^{(t)} \), to be observed in the current, \( t \)-th trial conducted with a candidate design, \( d_t \). In this equation, note
that the function \( p(y(t)|\theta, d_t) \) specifies the model’s probabilistic prediction of \( y(t) \) given the parameter \( \theta \) and the design \( d_t \), and \( p(\theta|y^{(1:t-1)}) \) is the posterior distribution of the parameter given past observations, which has become the prior for the current trial. Finally, \( \log \frac{p(\theta|y^{(1:t)}, d_t)}{p(\theta|y^{(1:t-1)})} \), referred to as the \textit{sample} utility function, measures the utility of design \( d_t \), assuming an outcome, \( y(t) \), and a parameter value (often a vector), \( \theta \).

\( U(d_t) \) in Eq. (1) is referred to as the \textit{expected} utility function, and is defined as the expectation of the sample utility function with respect to the data distribution \( p(y(t)|\theta, d_t) \) and the parameter prior \( p(\theta|y^{(1:t-1)}) \). Under the above particular choice of the sample utility function, the expected utility \( U(d_t) \) admits an information theoretic interpretation. Specifically, the quantity becomes the mutual information between the parameter variable \( \Theta \) and the outcome variable \( Y(t) \) conditional upon design \( d_t \), i.e.,

\[
U(d_t) = I(\Theta; Y(t)|d_t) \quad \text{(Cover & Thomas, 1991)}.
\]

Accordingly, the optimal design that maximizes \( U(d_t) \), or \( d^*_t = \arg \max U(d_t) \), is the one that yields the largest information gain about the model parameter(s) upon the observation of a measurement outcome.\(^1\)

The second, measurement step of ADO involves administering the optimal design \( d^*_t \) and observing the measurement outcome \( y(t) \), as illustrated in Figure 1. The final, third step of the ADO application is updating the prior \( p(\theta|y^{(1:t-1)}) \) to the posterior \( p(\theta|y^{(1:t)}) \) by Bayes’ theorem on the basis of the newly observed outcome \( y(t) \).

In implementing ADO, a major computational burden is finding the optimal design \( d^* \), which involves evaluating the multiple integrals in both the sample and the expected utility functions in Eq. (1) (integral is implicit in the sample utility). The integrals generally have no closed-form solutions and need to be calculated many times with candidate designs substituted during optimization. Further, online data collection requires that the integration and optimization be solved numerically on computer in real time. Advances in parallel computing have made it possible to solve some problems using grid-based algorithms. In situations in which grid-based methods are not suitable, several promising Markov chain Monte Carlo (MCMC) methods have been developed to perform the required computation (Müller, Sanso, & De Iorio, 2004; Amzal et al., 2006; Cavagnaro et al., 2010; Myung et al., 2013).

\section{Hierarchical Adaptive Design Optimization (HADO)}

As currently used, ADO is tuned to optimizing a measurement process at the individual participant level, without taking advantage of information available from data collected from previous testing sessions. Hierarchical Bayesian modeling (HBM; for

\(^1\)In defining the mutual information here, we assume that the goal of ADO is to maximize the information about all parameter elements of a model jointly, rather than some of them. In another situation, for example, the model may be a mixture model whose parameter \( \theta \) contains an indicator to a sub-model, and the goal of ADO may be to maximize the information about the indicator variable (i.e., the problem of model discrimination; e.g., Cavagnaro et al., 2010). In this case, the required change is to redefine the sample utility function in Eq. (1) by integrating out the parameters of no interest (e.g., sub-model parameters) from each of the distributions inside the logarithm.
theory, Good, 1965; de Finetti, 1974; Bernado & Smith, 1994; for application ex-
amples, Jordan, 1998; Rouder, Speckman, Sun, & Jiang, 2005; Rouder & Lu, 2005; Lee,
2006) not only provides a flexible framework for incorporating this kind of prior infor-
mation but is also well suited for being integrated within the existing Bayesian ADO
paradigm to achieve even greater efficiency of measurement.

The basic idea behind HBM is to improve the precision of inference (e.g., power
of a test) by taking advantage of statistical dependencies present in the data-generating
structure. For example, suppose that there are previous measurements taken from dif-
ferent individuals who are considered a random sample from a certain population. It
is highly likely that measurements taken from a new individual drawn from the same
population will share similarities with others.

In this situation, adaptive inference will enjoy greater benefit when taking the spe-
cific data structure into account rather than starting with no prior information. That is,
data sets, as a collection, contain information about one another, lending themselves to
more precise inference. Since individual data sets require themselves to be modeled
(i.e., a measurement model), the statistical relationship among them needs to be mod-
eled on a separate level, hence the model being hierarchical (for more examples of the
upper-level structure in a hierarchical model, see Koller & Friedman, 2009; Gelman et
al., 2004).

3.1 Formulation

To integrate HBM into ADO, let us first specify a common form of a hierarchical Bayes
model. Suppose that an individual-level measurement model has been given as a prob-
ability density or mass function,

\[ p(y_i | \theta_i) \]

given the parameter (vector), \( \theta_i \), for indi-
vidual \( i \), and the relationship among individuals is described by an upper-level model,

\[ p(\theta_{1:n} | \eta) \]

(e.g., a regression model with \( \eta \) as coefficients), where \( \theta_{1:n} = (\theta_1, \ldots, \theta_n) \)
is the collection of model parameters for all \( n \) individuals.\(^2\) Also commonly assumed
is conditional independence between individuals such that

\[ p(y_i | \theta_{1:n}, y_{-i}) = p(y_i | \theta_i) \]

where \( y_{-i} \) denotes the collection of data from all individuals except individual \( i \) (i.e.,
\( y_{1:n} = (y_1, \ldots, y_n) \) minus \( y_i \)). Then, the joint posterior distribution of the hierarchical
model given all observed data is expressed as

\[
p(\theta_{1:n}, \eta | y_{1:n}) = \frac{1}{p(y_{1:n})} p(y_{1:n} | \theta_{1:n}) p(\theta_{1:n} | \eta) p(\eta)
\]

\[ = \frac{1}{p(y_{1:n})} \prod_{i=1}^{n} p(y_i | \theta_i) p(\theta_{1:n} | \eta) p(\eta), \quad (2)\]

where \( p(\eta) \) is the prior distribution for the upper-level model’s parameter, \( \eta \), and the
marginal distribution \( p(y_{1:n}) \) is obtained by integrating the subsequent expression over
\( \theta_{1:n} \) and \( \eta \).

\(^2\)The upper-level model itself can be hierarchical (e.g., to model statistical depen-
dence among groups of individuals). Also, in addition to the individual-specific pa-
rameter, \( \theta_i \), the model can also have a commonly shared parameter (e.g., \( \lambda \)) such that
individual-level data are governed by \( p(y_i | \theta_i, \lambda) \). Extending our treatment of HADO to
such settings is straightforward, only with increased complexity of the notation.
The model also needs to be expressed in terms of an entity about which the measurement seeks to gain maximal information. In most measurement situations, it is sensible to assume that the goal is to estimate the traits of a newly measured individual most accurately. Suppose that a measurement session is currently underway on the $n$-th individual, and data from previous measurement sessions, $y_{1:n-1}$, are available. Then, the posterior distribution of $\theta_n$ for this particular individual given all available data is derived from (2) as

$$p(\theta_n | y_{1:n}) = \frac{1}{p(y_{1:n})} \int \int \left[ \prod_{i=1}^{n} p(y_i | \theta_i) \right] p(\theta_{1:n} | \eta) p(\eta) \, d\eta \, d\theta_{1:n-1} \quad (3)$$

where the marginal distribution $p(y_{1:n})$ is obtained by integrating the integrand further over $\theta_n$. From a computational standpoint, it is advantageous to turn the above posterior distribution into a sequentially predictive form. Under the assumption of conditional independence, Eq. (3) can be rewritten as

$$p(\theta_n | y_{1:n}) = \frac{p(y_n | \theta_n) p(\theta_n | y_{1:n-1})}{\int p(y_n | \theta_n) p(\theta_n | y_{1:n-1}) \, d\theta_n}, \quad (4)$$

where

$$p(\theta_n | y_{1:n-1}) = \frac{1}{p(y_{1:n-1})} \int \int \left[ \prod_{i=1}^{n-1} p(y_i | \theta_i) \right] p(\theta_{1:n} | \eta) p(\eta) \, d\eta \, d\theta_{1:n-1} \quad (5)$$

is the posterior predictive distribution of $\theta_n$ given the data from previous measurement sessions, $y_{1:n-1}$ (assuming that $y_n$ is yet to be observed). An interpretation of this form is that, as far as $\theta_n$ is concerned, the predictive distribution in Eq. (5) fully preserves information in the previous data $y_{1:n-1}$ and, in turn, serves as an informative prior for the current, $n$-th individual, which is updated upon actually observing $y_n$.

Having established the basic building blocks of hierarchical adaptive design optimization (HADO), we now describe how measurement within the HADO framework is carried out. Suppose that a measurement has been taken in trial $t-1$ for the $n$-th individual, and the session is in need of an optimal design to make the next observation, $y^{(t)}_n$, in trial $t$. Then, the optimal design, $d^*_t$, is the one that maximizes the following mutual-information utility:

$$U(d_t) = \int \int \left[ \log \frac{p(\theta_n, y_{1:n-1}, y^{(1:t)}_n, d_t)}{p(\theta_n | y_{1:n-1}, y^{(1:t-1)}_n)} \right] p(y^{(t)}_n | \theta_n, d_t) p(\theta_n | y_{1:n-1}, y^{(1:t-1)}_n) \, dy^{(t)}_n \, d\theta_n, \quad (6)$$

where $y_{1:n-1}$ denotes the data from previous $n-1$ measurement sessions, and $y^{(1:t)}_n$ contains the $n$-th individual’s measurements from past $t-1$ trials (i.e., $y^{(1:t-1)}_n$) plus an

---

3 Although the term *predictive distribution* is usually associated with a Bayesian model’s prediction of a future observation, it may also be used to mean the prediction of a future, latent variable in a hierarchical model, such as $\theta_n$ in the present context.
observation, \( y_n^{(t)} \), that is to be made in trial \( t \) using a candidate design, \( d_t \). Note that this utility function of HADO, similar as it may seem in its form to that of ADO in Eq. (1), takes all previously observed data into account through the hierarchical model, not just that from the current measurement session.

For HADO to be adaptive, Bayesian updating for posterior distributions inside the above utility function is performed recursively on two different levels (Figure 1). First, on the individual level (shaded region), updating is repeated over each measurement trial (i.e., to find the optimal design \( d_{t+1}^* \) after observing \( y_n^{(t)} \)) using Eq. (4) (i.e., Bayes’ theorem). Note that what is modified in Eq. (4) is only the individual data model (i.e., \( p(y_n|\theta_n) \)) with \( y_n = y_n^{(1:t-1)} \) augmented with a new measurement, \( y_n^{(t)} \). Next, when the session ends and a new one is to begin for the next participant (outside the shaded region), the hierarchical model is updated, again using Bayes’ theorem, on the basis of all \( n \) sessions’ data, \( y_{1:n} \), and expressed in a posterior predictive form for \( \theta_{n+1} \) (Eq. (5) with \( n + 1 \)). The session counter \( n \) now shifts to \( n + 1 \), the trial counter \( t \) is reset to 1, and the posterior predictive distribution becomes the prior for the new session to start with (i.e., \( p(\theta_{n+1}|y_{n+1}^{(1:0)}) = p(\theta_{n+1}|y_{1:n}) \)). This two-stage adaptation is a defining characteristic of HADO, hence the term “hierarchical adaptive.”

### 3.2 Implementation

In typical applications of hierarchical Bayes, posterior inference is conducted mainly to incorporate the data that have already been collected, and all the parameters of interest are updated jointly in a simulation-based method (e.g., via MCMC). This approach, however, is not well suited to HADO. Many applications of adaptive measurement require the search for an optimal design between trials to terminate in less than a second. To circumvent this computational burden, we formulated HADO, as described in the preceding section, in a natural way that suits its domain of application (experimentation), allowing the required hierarchical Bayes inference to be performed in two stages. Below we describe specific considerations for implementing these steps.

Once a numerical form of the predictive distribution (Eq. (5)) is available, updating the posterior distribution (Eq. (4)) within each HADO measurement session concerns only the current individual’s parameter and data just as in the conventional ADO. Accordingly, the recursive updating on the individual level will be no more demanding than the corresponding operation in conventional ADO since they involve essentially the same computation. Beyond the individual level, an additional step is required to revise the posterior predictive distribution of \( \theta_n \) given all previous data upon the termination of each measurement session, which is shown outside the shaded area in Figure 1. The result becomes a prior for the next session, serving as an informative prior for the individual to be newly measured.

Critical, then, to the implementation of HADO is the method for obtaining a numerical form of the predictive distribution of \( \theta_n \) before a new measurement session begins for individual \( n \). One straightforward approach is to define and compute the predictive distribution on a grid by integrating out all model parameters numerically, except \( \theta_n \). In particular, if \( \theta_i \)'s (or groups of \( \theta_i \)'s) are by design conditionally independent in the upper-level model (\( p(\theta_{1:n}|\eta)p(\eta) \) in Eq. (5)), it is possible to phrase the integral as re-
peated integrals so that simple, grid-based integration may be applied. Note that the shape of the integrands becomes highly pointed with a large number of observations per session (i.e., large \( t \)) and the posterior predictive of \( \theta_n \) tends to be narrow as well with accumulation of data over many sessions (i.e., large \( n \)). When this property is exploited in defining the grid (e.g., use of a sparse grid; Gerstner & Griebel, 1998) and parallel computing is utilized at the same time (e.g., with graphics processing units), the grid-based method can achieve highly accurate results within a manageable grid size.

Alternatively, the predictive distribution may still be represented on a grid, but the high-dimensional integration required to evaluate masses on it can be performed in a Monte Carlo fashion using a sampling algorithm such as MCMC. In case the grid-based approaches are not suitable, we may consider another method in which the quality of approximation to the predictive distribution may be compromised, yet the benefit from using the prior information is retained. One such approach is to use the technique of parametric density estimation with a sample drawn from the distribution. That is, a chosen parametric family, or a finite mixture of parametric families, \( p(\theta_n | \varphi) \), is fitted to a sample from \( p(\theta_n | y_{1:n-1}) \) so that \( p(\theta_n | \hat{\varphi}) \approx p(\theta_n | y_{1:n-1}) \).

Note that, although an estimate of \( \theta_n \) (e.g., posterior mean) is obtained at the end of the measurement session, the main purpose of posterior updating for \( \theta_n \) within the session is to generate optimal designs. Thus, the resulting estimate of \( \theta_n \) may not necessarily be taken as a final estimate, especially when the employed posterior predictive approximation is not highly precise. If needed, additional Bayesian inference based on the joint posterior distribution in Eq. (2) may be conducted after each test session with added data (top right box in Figure 1). This step will be particularly suitable when the upper-level structure (i.e., \( \eta \)) needs to be analyzed, or precise estimates of all previously tested individuals’ parameters are required for a certain type of analysis (e.g., to build a classifier that categorizes individuals based on modeled traits in the parameters).

It is also notable, from the computational perspective, that the procedure inside the shaded area in Figure 1 requires online computation during the measurement session, whereas the posterior predictive calculation outside the area (i.e., computing its grid representation) is performed offline between sessions. In case multiple sessions need to be conducted continually without an interval sufficient for offline computation, the same predictive distribution may be used as a prior for these sessions; for example, offline computation is performed overnight to prepare for the next day’s measurement sessions. This approach, though not ideal, will provide the same benefit of hierarchical modeling as data accumulate.

In sum, ADO for gaining maximal information from sequential measurements has been extended to incorporate the hierarchical Bayes model to improve information gain further. Conceptually, HADO improves the estimation of an individual data model by taking advantage of the mutual informativeness among individuals tested in the past. While there may be alternative approaches to forming an informative prior from past data for a Bayesian analysis, hierarchical Bayes is the method that enables both the generations of individual-level data and the relationship among them to be modeled and inferred jointly in a theoretically justified manner. The formulation and implementation of HADO provided above exploit the benefits of both hierarchical Bayes and ADO by integrating them within a fully Bayesian framework.
4 Application Example

The benefits of HADO were demonstrated in a simulated experiment in the domain of visual perception. Visual spatial processing is most accurately measured using a contrast sensitivity test, in which sine-wave gratings are presented to participants at a range of spatial frequencies (i.e., widths) and luminance contrasts (i.e., relative intensities). The objective of the test is to measure participants’ contrast threshold (detectability) across a wide range of frequencies, which together create a participant’s contrast sensitivity function (CSF). The comprehensiveness of the test makes it useful for detecting visual pathologies. However, because the standard methodology can require many hundreds of stimulus presentations for accurate threshold measurements, it is a prime candidate for the application of ADO and HADO.

Using the Bayesian framework described in Section 2, Lesmes, Lu, Baek, and Albright (2010) introduced an adaptive version of the contrast sensitivity test called qCSF. Contrast sensitivity, \( S(f) \), against grating frequency, \( f \), was described using the truncated log-parabola with four parameters (Watson & Ahumada, 2005):

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

where \( \gamma_{\text{max}} \) is the peak sensitivity at the frequency \( f_{\text{max}} \), \( \beta \) denotes the bandwidth of the function (full width at half the peak sensitivity), \( \delta \) is the low-frequency truncation level, and all variables and parameters are on base-10 log scales. The optimal stimulus selection through ADO, along with the parametric modeling, was shown to reduce the number of trials (<100) required to obtain a reasonably accurate estimate of CSF at only a minimal cost in parameter estimation compared to non-adaptive methods.

To demonstrate the benefits of HADO, the current simulation study considered four conditions in which simulated subjects were tested for their CSFs by means of four different measurement methods. We begin by describing how these conditions were designed and implemented.

4.1 Simulation Design

The two most interesting conditions were the ones in which ADO and HADO were used for stimulus selection. In the first, ADO condition, the qCSF method of Lesmes et al. (2010) was applied and served as the existing, state-of-the-art technique against which, in the second, HADO condition, its hierarchical counterpart developed in the present study was compared. If the prior information captured in the upper-level structure of the hierarchical model can improve the accuracy and efficiency of model estimation, then performance in the HADO condition should be better than that in the ADO (qCSF) condition. Also included for completeness were two other conditions to better understand information gain achieved by each of the two components of HADO: hierarchical Bayes modeling (HBM) and ADO. To demonstrate the contribution of HBM alone to information gain, in the third, HBM condition, prior information was conveyed through
HBM but no optimal stimulus selection was performed during measurement (i.e., no ADO). In the fourth, non-adaptive condition, neither prior data nor stimulus selection was utilized, so as to provide a baseline performance level against which improvements of the other methods could be assessed.

The hierarchical model in the HADO condition comprised two layers. On the individual level, each subject’s CSF was modeled by the four-parameter, truncated log-parabola described above. The model provided a probabilistic prediction through a psychometric function so that the subject’s binary response to a presented stimulus (i.e., detection of a sinusoidal grating with chosen contrast and frequency) could be predicted as a Bernoulli outcome (for detail, see Lesmes et al., 2010). On the upper level, the generation of a subject’s CSF parameters was described by a two-component, four-variate Gaussian mixture distribution, along with the usual, normal-inverse-Wishart prior on each component and the beta prior on mixture weights. The choice of a two-component mixture was motivated by the nature of the data. In practice, this type of information (i.e., membership to distinct groups) would be available for individuals so the use of a mixture distribution would be a sensible approach to lowering the entropy of the entity under estimation. While a more refined structure might be plausible (e.g., CSFs covary with other observed variables), we did not further investigate the validity of alternative models since the current hypothesis (i.e., individuals are similar to each other in the sense that their CSFs are normally distributed) was simple and sufficient to show the benefits of HADO.

The procedure for adaptive measurement (i.e., shaded area in Figure 1) followed the implementation of qCSF (Hou et al., 2010) in which all required computations for design optimization and Bayesian updating were performed on a grid in a fully deterministic fashion (i.e., no Monte Carlo integration). The posterior inference of the upper-level model, or the formation of a predictive distribution given the prior data (i.e., outside the shaded region in Figure 1), also involved no sampling-based computation. This was possible because the upper-level model (i.e., Gaussian mixture) allowed for conditional independence between individuals so that the posterior predictive density (Eq. (5)) of a particular \( \theta_n \) value could be evaluated as repeated integrals over individual \( \theta_i \)'s.

The ADO (qCSF) condition shared the same individual data model as specified in the HADO condition, but the variability among individuals was not accounted for by an upper-level model. Instead, each individual’s parameters were given a diffuse, Gaussian prior. The HBM condition took the whole hierarchical model from HADO, but the measurement for each individual was made with stimuli randomly drawn from a prespecified set. Finally, the non-adaptive method was based on the non-hierarchical model in ADO (qCSF) and used random stimuli for measurement.

To increase the realism of the simulation, we used real data collected from adults who underwent CSF measurement. There were 147 data sets, 67 of which were from individuals whose tested eye was diagnosed as amblyopic (poor spatial acuity). The remaining 80 data sets were from tests on non-diseased eyes. Thirty-six of these individuals took the qCSF test (300 trials with optimal stimuli) and 111 were administered the non-adaptive test (700 to 900 trials with random stimuli). The number of measurements obtained from each subject was more than adequate to provide highly accurate estimates of their CSFs.

To compare the four methods, we used a leave-one-out paradigm, treating 146 sub-
jects as being previously tested and the remaining subject as a new individual to be measured subsequently. We further assumed that, in each simulated measurement session, artificial data are generated from the real, estimated CSF (taken from the left-out subject) with one of the four methods providing stimuli. This situation represents a particular state in the recursion of measurement sessions shown in Figure 1; that is, the session counter is changing from $n = 146$ to $n = 147$ to test a new, 147th subject.

Finally, since the observations from each simulated measurement session were random variates generated from a probabilistic model, to prevent the comparison of performance measures from being masked by idiosyncrasies, ten replications of the 147 leave-one-out sessions were run independently and the results were averaged over all individual sessions (i.e., $10 \times 147 = 1,470$ measurement sessions were conducted in total).

### 4.2 Results

Performance of the four methods of measurement and inference described in the preceding section was assessed in three ways: information gain, accuracy of parameter estimation, and accuracy of amblyopia classification. These evaluation measures were calculated across all trials in each simulated measurement session. For information gain, the degree of uncertainty about the current, $n$-th subject’s parameter(s) upon observing trial $t$’s outcome was measured by the differential entropy (extension of the Shannon entropy to the continuous case):

$$H_t(\Theta_n) = -\int p(\theta_n|y_{1:n-1}, y_n^{(1:t)}) \log p(\theta_n|y_{1:n-1}, y_n^{(1:t)}) d\theta_n.$$  

(8)

Use of the differential entropy, which is not bounded in either direction on the real line, is often justified by choosing a baseline state and defining the observed information gain as the difference between two states’ entropies. In the present context, it is

$$IG_t(\Theta_0, \Theta_n) = H_0(\Theta_0) - H_t(\Theta_n),$$  

(9)

where $H_0(\Theta_0)$ denotes the entropy of a baseline belief about $\theta$ in a prior distribution so that $IG_t(\Theta_0, \Theta_n)$ may be interpreted as the information gain achieved upon trial $t$ during the test of subject $n$ relative to the baseline state of knowledge. In the current simulation, we took the entropy of the non-informative prior used in the conditions with no hierarchical modeling (i.e., ADO and non-adaptive) as $H_0(\Theta_0)$. Note that the information gain defined here is a cumulative measure over the trials in a session in the sense that $IG_t(\Theta_0, \Theta_n) = H_0(\Theta_0) - H_1(\Theta_n) + \sum_{s=2}^{t} [H_{s-1}(\Theta_n) - H_s(\Theta_n)]$ where the quantity being summed is information gain upon trial $s$ relative to the state before that trial.

Shown in Figure 2 is the cumulative information gain observed in each simulation condition designed to evaluate the performance of the four different methods. Each of the four curves corresponds to information gain ($y$-axis) in each condition over 200 trials ($x$-axis) relative to the non-informative, baseline state (0 on the $y$-axis). The information gain measures were averaged over all 1,470 individual measurement sessions in each condition. Then, we further normalized the measures by dividing them by the average information gain at the 200th trial achieved by the crude, non-adaptive method in
order to take the value of 1 as a baseline level of performance against which to compare performance of the other methods.

First, the results demonstrate that the hierarchical adaptive methodology (HADO) achieves higher information gain than the conventional adaptive method (ADO). The contribution of hierarchical modeling is manifested at the start of each session as a considerable amount of information (0.4) in the HADO condition (solid curve) than no information (zero) in the ADO condition (dashed curve). As expected, this is because HADO benefits from the mutual informativeness between individual subjects, which is captured by the upper-level structure of the hierarchical model and makes it possible for the session to begin with significantly greater information. As the session continues, HADO needs 43 trials on average to reach the baseline gain level (dotted, horizontal line) whereas ADO (qCSF) requires 62 trials. The clear advantage diminishes as information accumulates further over the trials since the measure would eventually converge to a maximum as data accumulate.

The HBM condition (dash-dot curve), which employs the hierarchical modeling alone and no stimulus selection technique, enjoys the prior information provided by the hierarchical structure at the start of a session and exhibits greater information gain than the ADO method until it reaches trial 34. However, due to the lack of stimulus optimization, the speed of information gain is considerably slower, taking 152 trials to attain baseline performance. The non-adaptive approach (dotted curve), with neither prior information nor design optimization, shows the lowest level of performance.

Figure 2: Information gain over measurement trials achieved by each of the four measurement methods.
Information gain analyzed above may be viewed as a summary statistic, useful for evaluating the measurement methods under comparison. Not surprisingly, we were able to observe the same profile of performance differences in estimating the CSF parameters. The accuracy of a parameter estimate was assessed by the root mean squared error (RMSE) defined by

$$\text{RMSE} \left( \hat{\psi}^{(t)} \right) = 20 \cdot \sqrt{\mathbb{E} \left[ \left( \hat{\psi}^{(t)} - \psi_{\text{true}} \right)^2 \right]},$$

where $\hat{\psi}^{(t)}$ is the estimate of one of the four CSF parameters (e.g., $\gamma^{\text{max}}$) for a simulated subject, which was obtained as the posterior mean after observing trial $t$’s outcome, $\psi_{\text{true}}$ is the true, data-generating parameter value for that subject, and the factor of 20 is multiplied to read the measure on the decibel (dB) scale as the parameter values are base-10 logarithms. The expectation is assumed to be over all subjects and replications, and hence was replaced by the sample mean over 1,470 simulated sessions.

Results from the second analysis, comparing parameter estimation error for each of the four models, are shown in Figure 3. Error was quantified in terms of RMSE ($y$-axis; described above) over 200 trials ($x$-axis) for each of the four parameters. As with the case of information gain, HADO benefits from the informative prior through the hierarchical model as well as the optimal stimuli through design optimization, exhibiting the lowest RMSE of all methods’ from the start to the end of a session, and this holds for all four parameters. The benefit of the prior information is also apparent in the HBM condition, making the estimates more accurate than with the uninformed, ADO method for the initial 40 to 80 trials, but the advantage is eclipsed in further trials by the effect of design optimization in ADO.

Finally, since accurate CSF measurements are often useful for screening eyes for disease, we performed yet another test of each method’s performance, in which the estimated CSFs were put into a classifier for amblyopia. Despite various choices of a possible classifier (e.g., support vector machine, nearest neighbor, etc.), the logistic regression model built on selected CSF traits (Hou et al., 2010), which had been shown to be effective in screening amblyopia, sufficed for our demonstration. Performance of each measurement method in classifying amblyopia was assessed in the leave-one-out fashion as well, by first fitting the logistic regression model using the remaining 146 subjects’ CSFs and then adding the left-out, simulated subject’s CSF estimate to the classifier to generate a prediction. The given, actual label (i.e., amblyopic or normal eye) of the left-out subject, which had been provided by an actual clinical diagnosis, was taken as the true value against which the classification result in each simulation condition was scored.

Not surprisingly, classification accuracy increases with accumulation of measurement data in all methods. This is seen in Figure 4, which shows the percentage of correct amblyopia classifications out of all cases of amblyopic eyes over the first 100
Figure 3: Accuracy of parameter estimation over measurement trials achieved by each of the four measurement methods.

As was found with the preceding tests, HADO demonstrates superior performance, requiring only a small number of trials to produce highly accurate classification results. Most notably, it takes on average 30 trials for HADO to correctly classifies an amlyopic eye 90% of the time, whereas the non-

4Classification results for normal eyes are not shown since the prior of CSF parameters was specified in a way that the classifier with any of the methods would categorize a subject as being normal when there is little or no accumulation of data (i.e., a bias was built in to avoid false alarms). In addition, the results are shown only up to 100 trials to provide a better view of performance differences across methods.
Figure 4: Accuracy of amblyopia classification over measurement trials achieved by each of the four measurement methods.

The hierarchical adaptive method (ADO) requires 53 trials to achieve the same level of accuracy, otherwise reaching 82% accuracy with the same 30 trials.

Note that in the early trials of ADO and HADO, there can be considerable fluctuation in classification accuracy. This is not due to a small sample size (proportions out of 670 amblyopic eyes have sufficiently small standard errors), but rather to the adaptive method itself. Seeking the largest possible information gain, the algorithm is highly exploratory in choosing a stimulus that would yield a large change in the predicted state of the tested individual. This characteristic especially stands out in early trials of the classification task by causing some of the amblyopic eyes near the classifier’s decision bound to alternate between the two sides of the bound across one trial to another. This effect remains even after taking proportions out of the large sample (670) because, with little accumulation of observations, selecting optimal stimuli in early trials is systematic without many possible paths of the selection. Although this can lead to short-term drops in accuracy, the benefits of early exploration pay dividends immediately and over the long term.

5 Discussion

The present study demonstrates how hierarchical Bayes modeling can be integrated into adaptive design optimization to improve the efficiency and accuracy of measurement.
When applied to the problem of estimating a contrast sensitivity function (CSF) in visual psychophysics, HADO achieved an average decrease of 38% (from 4.9 dB to 3.1 dB) in error of CSF parameter estimation and an increase of 10% (from 82% to 90%) in accuracy of eye disease screening over conventional ADO, under the scenario that a new session could afford to make only 30 measurement trials. In addition, efficiency of testing improved by an average of 43% in the sense that the required number of trials to reach a criterion of 90% screening accuracy decreased from 53 to 30 trials.

Although the simulation study served the purpose of demonstrating the benefit of the hierarchical adaptive methodology, the full potential of HADO should be greater than that demonstrated in our particular example. The level of improvement possible with HADO depends on the sophistication of the hierarchical model itself. In our case, the model was based on a simple hypothesis that a newly tested individual belongs to the population from which all other individuals have been drawn. Although the model has flexibility in defining the population as a mixture distribution, it conveys no further specific information about the likely state of a new individual (e.g., his or her membership to a mixture component is unknown).

There are various situations in which hierarchical modeling can take better advantage of the data-generating structure. For example, although modeled behavioral traits vary across individuals, they may covary with other variables that can be easily observed, such as demographic information (e.g., age, gender, occupation, etc.) or other measurement data (e.g., contrast sensitivity correlates with measures of visual acuity - eye chart test). In this case, a general, multivariate regression or ANOVA model may be employed as the upper-level structure to utilize such auxiliary information to define a more detailed relationship between individuals. This greater detail in the hierarchical model should promote efficient measurement by providing more precise information about the state of future individuals.

In many areas of behavioral science, there is more than one test that measures the same condition or phenomenon (e.g., memory, depression, attitudes). Often times, these tests are related to each other and modeled within a similar theoretical framework. In such situations, a hierarchical model provides a well-justified way to integrate those models in such a way that behavioral traits inferred under one model are informative about those estimated by another. Yet another situation in which hierarchical modeling would be beneficial is when a measurement is made after some treatment and it is sensible or even well known that the follow-up test has a particular direction of change in its outcome (i.e., increase or decrease). Taking this scenario one step further, a battery of tests may be assumed to exhibit profiles that are characteristic of certain groups of individuals. The upper-level structure can also be modeled (e.g., by an autoregressive model) to account for such transitional variability in terms of the parameters of the measurement model. With these kinds of structure built in the hierarchical model, HADO can be used to infer quickly the state of new individuals.

An assumption of the approaches to higher-level modeling discussed above is that the most suitable data-generating structure is already known. In fact, a sufficient amount of data is needed to determine which structure is best suited. To be more precise, the optimally complex structure for the best possible inference depends on the amount of information available; an arbitrarily complex model that is not validated by data will lead to sub-optimal inference. For this reason, HADO will perform best when the hi-
erarchical model evolves along with the accumulation of data. Larger data sets make it possible to evaluate better alternative modeling hypotheses, and analysis methods such as Bayesian model choice (Kass & Raftery, 1995) or cross validation can be performed to guide model revision. In effect, the upper-level model will evolve by incorporating increasingly richer structure (e.g., finer subgroup distinctions or better selected predictor variables in a regression model).

The notion of model evolution fits with recent advances in nonparametric Bayes methods that essentially seek to enable a statistical model to adapt itself to the amount of information in the data by adding more and more components with no preset limit (MacEachern, 2000; Rasmussen & Williams, 2006; Teh & Jordan, 2010). This methodology can further stretch the extent of model evolution and will be especially suited to HADO because most modern measurement processes are computer-based, so data collection and organization are effortless, allowing the method to quickly exploit a massive amount of data. We have performed a preliminary investigation by applying one of the nonparametric methods in a simulation experiment, and the results are promising.

To conclude, science and society benefit when data collection is efficient with no loss of accuracy. The proposed HADO framework, which judiciously integrates the best features of design optimization and hierarchical modeling, is an exciting new tool that can significantly improve upon the current state of the art in experimental design, enhancing both measurement and inference. This theoretically well-justified and widely applicable experimental tool should help accelerate the pace of scientific advancement in behavioral and neural sciences.

Acknowledgments

This research is supported by National Institute of Health Grant R01-MH093838 to J.I.M and M.A.P., as well as National Eye Institute Grant R01-EY021553-01 to Z.-L.L. We thank Fang Hou and Chang-Bing Huang for organizing the CSF measurement data. Correspondence concerning this article should be addressed to Woojae Kim, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210. Email: kim.1124@osu.edu.

References


