Real-time Adaptive Design Optimization within Functional MRI Experiments

Giwon Bahg

Department of Psychology, The Ohio State University

Per B. Sederberg

Department of Psychology, University of Virginia

Jay I. Myung, Xiangrui Li, Mark A. Pitt

Department of Psychology, The Ohio State University

Zhong-Lin Lu

Division of Arts and Sciences, NYU Shanghai; Center for Neural Science and Department of Psychology, New York University

Brandon M. Turner

Department of Psychology, The Ohio State University

Abstract

Efficient data collection is an important goal in cognitive neuroimaging studies because of the high cost of data acquisition. One method of improving efficiency is to maximize the informativeness of the data collected on each trial. We propose an Adaptive Design Optimization (Cavagnaro, Myung, Pitt, & Kujala, 2010; Myung, Cavagnaro, & Pitt, 2013) procedure to optimize the sequencing of stimuli for model-based functional neuroimaging studies. Our method uses the Joint Modeling Framework (B. M. Turner, Forstmann, & Steyvers, 2019; B. M. Turner, Forstmann, et al., 2013) to maximize the information learned about how the brain produces a behavior by integrating over neural and behavioral data simultaneously. We validate our method in simulation and real-world experiments by showing how Adaptive Design Optimization proposes the optimal stimulus sequence to reduce uncertainty and improve accuracy from a Bayesian perspective.

Introduction

Functional magnetic resonance imaging (fMRI) has become one of the most important tools in cognitive science to investigate human brain activity because of its noninvasive nature, reasonable temporal resolution, and precise spatial resolution (Poldrack, Mumford, & Nichols, 2011). However, the cost of data collection in neuroimaging studies using fMRI is exceptionally high due to high maintenance expenses of the scanner and low signal-to-noise ratio in the blood oxygenation level dependent (BOLD) response. Therefore, optimizing the experimental procedure and design is an important methodological issue for improving the efficiency of fMRI studies. To this point, many methods have been proposed to ameliorate certain limitations of fMRI measurements. For example, optimizing a stimulus sequence (for a recent review, see Holling, Maus, & van Breukelen, 2013) prior to performing an experiment, optimizing the scan acquisition sequence, or reducing the scanning area to specific brain regions (de Hollander, Keuken, van der Zwaag, Forstmann, & Trampel, 2017) can all improve the signal-to-noise ratio of fMRI measures.

However, most of the previous design optimization methods for fMRI experiments focus on detecting brain activations associated with a task or its across-condition contrast that heavily rely on general linear modeling (GLM), or estimating the shape of the hemodynamic response. Would these methods be helpful if your research interest is, for example, to study the mechanism of self-control? In particular, what if the population of interest is children, whose attention span is quite limited? Such a population is uniquely difficult to obtain extensive numbers of trials, and therefore efficient data collection is strongly required.

Although the previous approaches will be useful in increasing the signal-to-noise ratio of our experiment, they may not be optimally configured to study computational mechanisms associated with self-control (B. M. Turner et al., 2018). Moreover, there is no guarantee that focusing on the signal qualities of neural data when using methods relying on GLM will provide the optimally informative experimental design for understanding cognition across participants. In a computational cognitive model, different levels of cog-

nitive functioning are represented as different parameter values, which might affect the definition of the "optimal" set of stimuli for each individual. Also, there are often large differences in neural or behavioral responses to similar stimuli across participants, or even within a participant scanned at different points in time (Miller et al., 2002). Therefore, it is sometimes unclear which stimuli should be used for a given participant, which suggests the need for adaptive, rather than static, optimization of experimental design.

In the present study, we present a general-purpose methodology that overcomes many of the aforementioned limitations of fMRI measurement and optimization methods. The central feature of our algorithm is its optimally adaptive stimulus-proposal scheme as a way to maximize the information that is learned about how a brain produces a behavioral response. Specifically, the active-learning algorithm chooses a stimulus on each trial by making real-time statistical inferences, in this case about the participant's perceptual decision making process. The key advantages of our approach are three-fold. First, the data collection process in an fMRI experiment involves an optimization of the stimulus sequence in such a way to maximize information learned on each trial about the underlying decision process. Hence, the focus is on information about the decision, rather than number of trials or number of scans per trial. Second, the data collection process is completely adaptive: unlike static design optimization methods (Holling et al., 2013; Smucker, Krzywinski, & Altman, 2018), we analyze fMRI data from trial to trial in real time so that the stimulus search process is always conditional on the current state of knowledge about the participant's decision process. Third, our method incorporates both neural and behavioral data to optimize stimulus choice, a feature that is different from previous real-time design optimization methods such as QUEST (Watson & Pelli, 1983), Psi method (Kontsevich & Tyler, 1999), Dynamically Adaptive Imaging (Cusack, Veldsman, Naci, Mitchell, & Linke, 2012) and The Automatic Neuroscientist (Lorenz et al., 2016). All these advantages reduce overall scan time for a desired amount of information by automatically tailoring the experimental design to each individual participant. In simulated and empirical studies, we show how the method can be used to collect data that are more informative than what could otherwise be

obtained, despite neural variability and other complications that plague fMRI experiments (Greve et al., 2013).

Overview of the Methodology

Figure 1 provides a flowchart of the method we have developed to perform adaptive optimization of real-time fMRI experiments. Following typical structural scans and functional localizer tasks (Appendix A), fMRI data are collected during the task in real time and processed to determine activation of each region of interest with motion correction. The pattern of activation is then evaluated by a joint model (Palestro et al., 2018; B. M. Turner, Forstmann, et al., 2013; B. M. Turner, Van Maanen, & Forstmann, 2015), whose parameters convey the current knowledge of how brain activity best predicts the pattern of behavioral responses. To keep the knowledge of brain-behavior connections as current as possible, the parameters of the joint model are updated via Bayes rule on each trial using several techniques: (1) differential evolution Markov chain Monte Carlo (B. M. Turner, Sederberg, Brown, & Steyvers, 2013) to efficiently approximate the parameter posterior distributions (Section "Posterior Sampling via DE-MCMC"), (2) one-trial lag to prevent hemodynamic lag from adversely affecting the posterior estimates (Section "One-trial-lag Optimization"), and (3) dynamic gridding to adjust the grid used to approximate the joint posterior distribution (Section "Dynamic Gridding"). Finally, we rely on active learning through ADO (Cavagnaro et al., 2010; Myung et al., 2013) to guide selection of stimuli on a trial-by-trial basis. The advantage of using ADO is that it selects stimuli for the next trial based on the current parameter estimates in the joint model by integrating over all possible stimuli and all possible neural and behavioral responses. The stimulus design that is maximally informative about how the brain produces a behavior of interest is selected for the next trial, and the process repeats until a stopping criterion is reached. Supplementary code used in this study for implementing ADO is available on `https://github.com/MbCN-lab`.

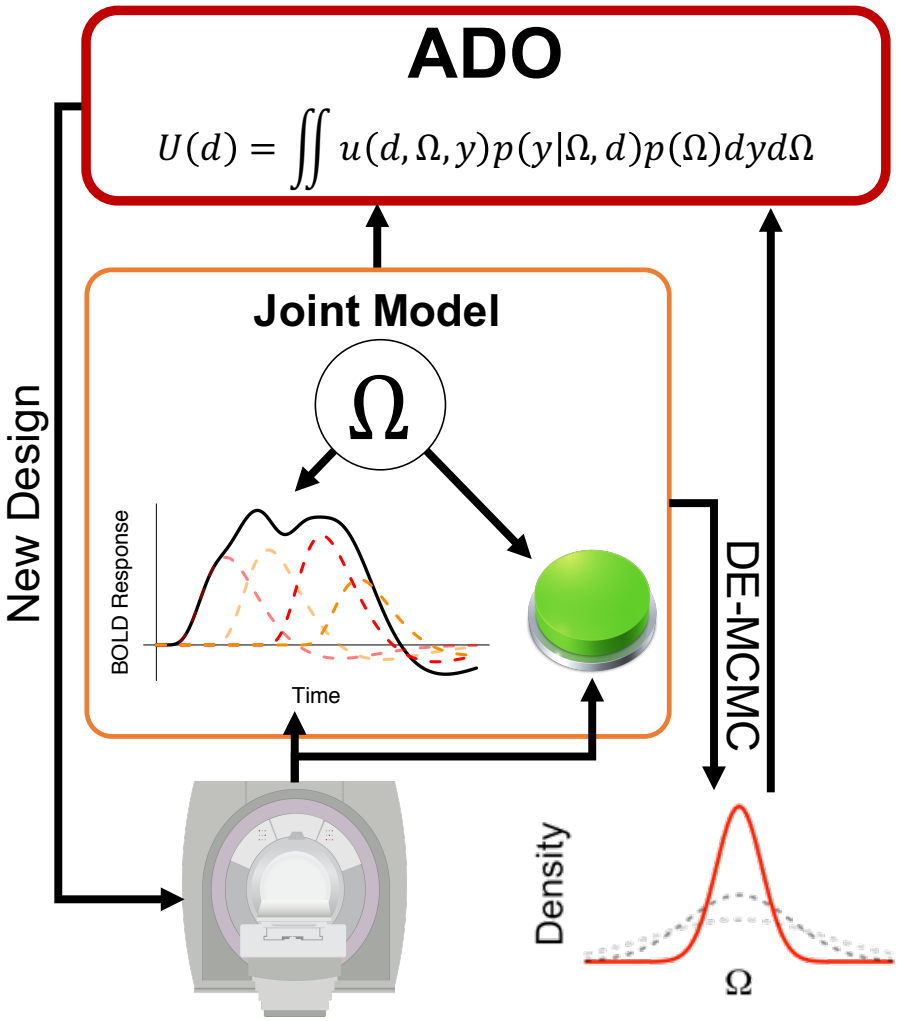

Figure 1. **Pipeline for the fMRI-ADO Framework**. The figure shows how Adaptive Design Optimization can be used to adaptively select stimuli from a set of potential candidates to maximize the information relating brain activity to a behavioral response. On each trial, a stimulus is selected based on the calculation of global utility (top), and the stimulus is shown to the participant (bottom left). A joint model analyzes the Blood Oxygenated Level Dependent (BOLD) response and the choice outcome from the new stimulus. Given the new information, we recalculate the global utility across the stimulus space to propose the next stimulus in the sequence. Occasionally (e.g., every four trials), a dynamic gridding process is used to effectively integrate over stimulus and parameter spaces.

Adaptive Design Optimization

ADO is a Bayesian and model-based method for optimal experimental design based on an information theoretical measure of design utility. ADO was originally proposed as

an online design optimization tool for model comparison in cognitive science experiments. However, when only considering one model, the method naturally reduces to an algorithm for optimizing parameter estimation. A cognitive process model, along with the history of a participant's responses, guides stimulus selection on each trial so that a selected stimulus is hypothesized to yield the greatest amount of information about model parameters. Although ADO is similar to many staircasing procedures used in psychophysical experiments, ADO is more general in that it can be applied naturally to different types of neural data (e.g., EEG, single-unit recordings, decision choices) or to any type of cognitive process model.

ADO proposes an optimal design for upcoming trials by solving an optimization problem. Given a candidate design of an experiment for the next trial $d \in D$, design proposals are made by selecting a design associated with the highest global utility $U(d)$. Here, $U(d)$ is defined with respect to the local utility $u(d, \theta, y)$, which is a function of the design d, the model parameter θ, and the anticipated (behavioral) response on the next trial y^*. A generic description of the design optimization is as follows:

$$
\begin{aligned}
d_{t+1} &= \underset{d}{\operatorname{argmax}}\, U(d) \\
U(d) &= \int_{y^* \in Y} \int_{\theta \in \Theta} u(d, \theta, y^*) p(y^* | \theta, d) p(\theta)\, d\theta\, dy^*.
\end{aligned} \tag{1}
$$

A local utility function $u(d, \theta, y)$ evaluates the utility or informativeness of a design d regarding a model parameter set θ when a design d is used and a response y is anticipated in a hypothetical experimental trial. The global utility $U(d)$ is computed as an "average" local utility by integrating the local utility over a parameter space Θ and a response space Y.

A posterior covariance matrix and the sum of squared errors are often used as utility functions (Ryan, Drovandi, McGree, & Pettitt, 2016). However, a standard implementation of ADO relies on mutual information to evaluate the utility of each design because mutual information performs well for both parameter estimation and model comparison.

In addition to d, θ, and y^*, assume $d_{1:t}$ and $y_{1:t}$ that represent a series of experimental designs and collected (behavioral) responses in the previous t trials, respectively. A global utility function based on mutual information is

$$U(d) = \int_{y^* \in Y} \int_{\theta \in \Theta} \log \frac{p(\theta|d_{1:t}, y_{1:t}, d, y^*)}{p(\theta|d_{1:t}, y_{1:t})} p(y^*|\theta, d_{1:t}) p(\theta) \; d\theta \, dy^*. \tag{2}$$

Note that by the definition of mutual information, a local utility function in Equation 1 is

$$u(d, \theta, y^*) = \log \frac{p(\theta|d_{1:t}, y_{1:t}, d, y^*)}{p(\theta|d_{1:t}, y_{1:t})} \tag{3}$$

(Myung et al., 2013).

Myung et al. (2013) suggested a simple integration strategy based on grid based methods. Myung et al.'s approach proceeds by first defining a number of grid points for each dimension of design, parameter, and response spaces. Once the grids are defined over an entire search space, ADO then evaluates local utilities (i.e., $u(d, \theta, y^*)$) and joint densities of θ and $y_{1:t}$ (i.e., $p(y^*|\theta, d_{1:t})p(\theta) = p(y^*, \theta|d_{1:t})$) for all grid points. When the grid is uniformly distributed, a global utility for a candidate design d is computed by taking a mean of weighted local utility values sharing a target design d :

$$U(d) \approx \frac{1}{n_d} \sum_{\{\theta, y^*\}} \log \frac{p(\theta|d_{1:t}, y_{1:t}, d, y^*)}{p(\theta|d_{1:t}, y_{1:t})} p(y^*|\theta, d_{1:t}) p(\theta) \tag{4}$$

where n_d is the total number of grid points assigned to a candidate design d. When a grid is defined by sampling from the parameter prior and a separate sampling distribution for the data, the term $p(y^*|\theta, d_{1:t})p(\theta)$ is not required (Myung et al., 2013).

When using any continuous measurements (e.g., neural activation level, reaction time), all probability measures must be introduced into the algorithm after normalization or considering grid-based partitioning. By normalization, we mean that all probability densities must be transformed so that they will sum up to one within each condition. The latter means that each probability value must be approximated by multiplying the

evaluated density with the region covered by the grid point, as in the Riemann sum. This normalization technique ensures that all quantities used within the algorithm are legitimate probability measures, not densities, to facilitate comparison among grid points.

ADO has been fruitfully applied in cognitive science (Cavagnaro, Aranovich, McClure, Pitt, & Myung, 2016; Cavagnaro, Pitt, Gonzalez, & Myung, 2013; Cavagnaro, Pitt, & Myung, 2011) and more recently to neuroscientific problems, but only in simulation work (DiMattina, 2016; Sanchez et al., 2014; Sanchez, Lecaignard, Otman, Maby, & Mattout, 2016). Although similar adaptive stimulus optimization methods have been applied on neurophysiology studies (for a recent review, see DiMattina & Zhang, 2013), there was no study directly connecting behavioral and neural data for optimizing stimuli. For example, DiMattina (2016) used adaptive stimulus generation, which has Bayesian adaptive mechanisms to compare contrast gain models in human vision. However, this application neither modeled neural activity nor used neural data directly. Instead, the researcher developed an encoding-decoding model to map contrast stimuli to hypothesized neural responses (encoding model) and then to behavioral responses (decoding model) such that ADO only operated on the behavioral response data. While this study involves a neurophysiological model, ADO has yet to be demonstrated as an effective tool in the online processing of neural data.

Joint Modeling Framework

Today, scientists interested in studying cognition are faced with many options for relating experimentally-derived neurophysiological variables to the dynamics underlying a cognitive process of interest. A recent trend in cognitive science is to blend the theoretical and mechanistic accounts provided by models in the field of mathematical psychology with the high-dimensional data brought forth by modern measures of cognition such as those collected in an fMRI experiment. One new approach for imposing a reciprocal link between brain measures and decision variables is the "joint modeling" approach. Unlike the traditional modeling approaches (for descriptions of uniqueness, see B. M. Turner et

al., 2019), joint models enforce a constraint on model parameters based on the random variation in the neural data. In other words, if one treats the neural data as a statistical covariate within the model, the estimates of the cognitive model parameters will be more constrained under mild conditions (B. M. Turner, 2015). The process of fitting the model to data procures estimates of neural activation parameters for each stimulus presentation. For the behavioral data, a cognitive model is developed, and similarly fit to behavioral data such as choice response time measures. To impose statistical reciprocity, a linking function specifies how the parameters of the neural data are related to the parameters of the cognitive model.

In a series of studies, joint models have been shown to outperform models that do not incorporate neural measures, suggesting that the information in neural measures can be used to make substantially better predictions for decisions (e.g., B. M. Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016). In addition, compared to approaches estimating single-trial neural and behavioral model parameters separately and correlating them (e.g., Forstmann et al., 2010, 2008), joint models can minimize the loss of information about statistical constraints. In the present investigation, we will optimize this framework to arrive at better representations of how the brain produces a behavior.

Adaptive Design Optimization: Extension to the Neural Data

Introducing neural data and its activation model does not change the definition of the global utility function and the searching process. However, the dimension of both parameter and response spaces increases because we have incorporated neural data and therefore need to consider the expected neural responses into ADO.

Ideally, a full joint model would allow ADO to use a raw BOLD time-series vector $\mathbf{N}$ as its neural input. Assuming a hierarchical joint model $\Omega = (\theta_{\text{hyper}}, \theta_{\text{neural}}, \theta_{\text{behavioral}})$, observed neural data during the previous t trials $\mathbf{N}_{1:end(t)}$, and anticipated neural observa-

tions $\mathbf{N}^*$, we can define global utility function as

$$U_{JM}(d) = \int\int\int u(d, \Omega, \mathbf{N}^*, y^*) p(\mathbf{N}^*, y^* | \Omega, d) p(\Omega)\ d\Omega\, d\mathbf{N}^*\, dy^*$$
$$= \int\int\int \log \frac{p(\Omega | d_{1:t}, \mathbf{N}_{1:end(t)}, y_{1:t}, \mathbf{N}^*, y^*)}{p(\Omega | d_{1:t}, \mathbf{N}_{1:end(t)}, y_{1:t})} p(\mathbf{N}^*, y^* | \Omega, d) p(\Omega)\ d\Omega\, d\mathbf{N}^*\, dy^* \quad (5)$$

Note that the subscript notation of the variables representing neural (i.e., $N_{1:end(t)}$) and behavioral (i.e., $y_{1:t}$) data are inconsistent due to the mismatch of temporal resolution between BOLD and behavioral responses. Here, $end(t)$ refers to the number of neural data samples (i.e., time points) until the end of the t-th trial.

However, using the raw neural data is practically impossible within ADO because of the interaction between ADO, the dimensionality of neural data increasing in real time, and the shape of the anticipated BOLD responses. Equation 5 suggests that all data points in the time-series vector $\mathbf{N}$ must be integrated over $\mathbb{R}^n$ where n is the length of the time-series vector. The problem in the real-time fMRI application is that new data are continuously added during the scan causing increases in the dimension of the neural data space, even when ADO is computing the global utility of candidate designs.

A more critical problem is that computation time required for ADO interacts with the data collection procedure. If ADO functions relying on the raw BOLD responses, it has to evaluate the expected neural responses for the next few time points. However, the number of time points to be considered is arbitrary here because computation time for ADO will delay the whole schedule of the next trial (e.g., stimulus presentation). Moreover, changes in the schedule of the next trial will conclude in changing the shape of predicted BOLD responses and essentially in the evalutation of the global utility. As these issues occur in real time while ADO computes the next optimal stimulus, ADO would not be able to handle this issue appropriately.

As an alternative, we can implement a global utility function based on a "limited" version of the joint model structure using trial-wise neural activation estimates. For example, we can make use of simple statistical models, such as a general linear model, to first

obtain estimates of the unknown stimulus- or trial-wise neural activations β, denoted $\hat{\beta}$ (e.g., Rissman, Gazzaley, & D'Esposito, 2004). Given these neural activation estimates for previous trials $\hat{\beta}_{1:t}$ and for the next hypothetical trial $\hat{\beta}^*$, a global utility is defined as

$$\begin{aligned} U_{LJM}(d) &= \int\int\int u(d, \Omega, \hat{\beta}^*, y^*) p(\hat{\beta}^*, y^* | \Omega, d) p(\Omega)\ d\Omega\, d\hat{\beta}^*\, dy^* \\ &= \int\int\int \log \frac{p(\Omega | d_{1:t}, \hat{\beta}_{1:t}, y_{1:t}, \hat{\beta}^*, y^*)}{p(\Omega | d_{1:t}, \hat{\beta}_{1:t}, y_{1:t})} p(\hat{\beta}^*, y^* | \Omega, d) p(\Omega)\ d\Omega\, d\hat{\beta}^*\, dy^*. \end{aligned}$$

When the limited joint model is used, single-trial neural activation estimates serve as the neural input into the ADO procedure, and this is an effective strategy because these estimates efficiently describe stimulus- or trial-wise brain activity, unlike the raw neural data as in Equation 5. By reducing the set of possible data points to single-trial activation parameters rather than a full BOLD time series, the computational burden of using ADO becomes manageable once again. However, this reduction does come at the cost of inflated uncertainty in the estimates of neural activation. Also, note that the response space for the continuous neural activation $\hat{\beta}$ must be discretized if one attempts to use a grid-based approximation as in Equation 4.

Introducing the Neural Data: Single-trial Neural Activation

The use of stimulus- or trial-wise neural activation estimates serves as a remedial strategy for the high dimensionality problem of raw BOLD responses. To actually use the single-trial activation estimates, fMRI-based ADO must include a component that estimates neural activation amplitude evoked by each stimulus or trial so that the neural estimates can be used for proposal generation.

The conventional approach to estimating single-trial activation is to perform a general linear model (GLM) analysis – an application of multiple linear regression to fMRI data. A GLM uses a design matrix consisting of vectors representing the onset times of events of interest (e.g., stimulus presentation, response production) convolved with a hemodynamic response function. A typical approach is to define condition-wise regressors for comparing

the mean activation estimates across conditions (for a more general introduction to this topic, see introductory textbooks for fMRI data analysis such as Poldrack et al., 2011).

However, when using ADO, GLM regressors must be defined at each stimulus- or trial-level because we need information of neural activity associated with each stimulus. Conceptually, stimulus-level regressors can be easily made by setting the onset vectors for each individual stimulus, not for each condition. A single-trial GLM can be implemented in a Bayesian framework (e.g., Palestro et al., 2018). However, full posterior estimation is time consuming in real-time fMRI experiments due to the large number of single-trial regressors or multiple BOLD response vectors. In our application, we used frequentist estimates to obtain trial-wise neural activation estimates efficiently. For example, ordinary least squares estimates can be derived as:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{N} \tag{6}$$

where $\mathbf{X}$ is a design matrix, a superscript T indicates the transpose operation, and $\mathbf{N}$ is a raw BOLD time-series vector. We relied on ordinary least square estimates for extracting the target region given the time constraint. However, during the task, we used estimates assuming the first-order temporal autocorrelation in the optimization routine for acquiring as accurate values as possible.

Issues in Estimation Methods. Although the idea of estimating stimulus-wise neural activation using GLMs seems straightforward, a few methodological issues can affect the quality of the estimates and computational burden imposted on ADO. The first issue is the shape of the HRF. As the observed neural data are assumed to be the product of convolving a sequence of experimental events and the HRF, how we define (or model) the HRF affects the estimates of the single-trial neural activation.

In this study, we used the canonical HRF model (also known as 'double-gamma HRF') with fixed shape parameters: $a_1 = 6$, $a_2 = 16$, $b_1 = 1$, $b_2 = 1$, and $c = 1/6$. Given the

time index t and fixed shape parameters, the double-gamma HRF is

$$h(t) = \frac{t^{a_1-1} b_1^{a_1} \exp(-b_1 t)}{\Gamma(a_1)} - c\frac{t^{a_2-1} b_2^{a_2} \exp(-b_2 t)}{\Gamma(a_2)}. \quad (7)$$

However, it is worth noting that if the HRF is misspecified, estimates of the trial-wise neural activity may be suboptimal, as is the case in nearly every model-fitting procedure (Lindquist, Loh, Atlas, & Wager, 2009).

Ideally, we could estimate the shape parameters of the HRF during the experiment. However, simultaneously estimating the shape parameters will quickly increase the computational complexity of the design optimization problem. As the main purpose of our study is proof of concept, we used the canonical HRF as a reasonable approximation.

Another statistical issue is that the single-trial GLM is vulnerable to multicollinearity, especially when an experiment uses rapid event-related designs (i.e., short interstimulus or intertrial intervals). This problem comes from the shape of the HRF, which has a temporally extended profile. If two experimental events are offset with a short time interval, the corresponding regressors will be similarly shaped to one another, making their correlation high. Although this problem might not apply to our study with better trial-by-trial separation, an appropriate methodological consideration is still needed.

Previous studies have discussed this issue and proposed alternative methods for better single-trial neural activation estimates (e.g., Abdulrahman & Henson, 2016; Mumford, Davis, & Poldrack, 2014; Mumford, Turner, Ashby, & Poldrack, 2012; B. O. Turner, Mumford, Poldrack, & Ashby, 2012). However, many of these alternatives use the strategy of fitting as many GLMs as the number of stimuli or trials to be analyzed, which could increase the computation time in the ADO pipeline. Also, selection of the estimation method must consider how one plans to update the single-trial neural estimates together (see "Incremental Estimation of Single-trial Neural Activation"). Hence, we decided to use a more traditional, single-GLM-based approach (Rissman et al., 2004) for this proof-of-concept study, while fully acknowledging its limitation.

Incremental Estimation of Single-trial Neural Activation. To update the neural activity from newly occurred events in the latest trial, estimation of single-trial neural activation is necessary at the end of every trial. However, using this incremental procedure implies that BOLD time-series will be continuously updated during an entire scanning session. For single-trial neural estimates that are already obtained, we cannot avoid slight changes in those estimates because newly updated data will change the likelihood (and therefore posterior density) of possible estimates. Hence, we have to determine how to deal with the variability of single-trial neural estimates during fMRI-based ADO experiments.

The first option to handle the variability of single-trial neural estimates is to block the updating of neural estimates included in ADO during previous trials. In this case, neural activation estimates of previous stimuli or trials will be fixed in further trials and new estimates for those trials will not be used in ADO. Only the estimates from a new trial will continue being added in the neural "data" – in this case, single-trial neural activation estimates – vector. This approach ensures the stability of ADO algorithm as the estimates of neural activity remain constant once they have been estimated on a given trial. Also, this approach can maximize computational efficiency of grid-based ADO. As long as the grid settings and previously obtained neural data do not change, we can store the posterior probability density of the current trial as the prior for the next trial, and simply call those values when evaluating the global utility.

The second option to handle the variability of single-trial neural estimates is to allow ADO to update the neural estimates every trial. From this perspective, ADO must use the best "data" – again, single-trial neural activation estimates – available at each trial. Hence, ADO must refer to new estimates as they become more accurate and less variable as the experiment moves on.

In the simulation experiments, we made an ideal assumption that we always obtain perfect estimates of stimulus-wise neural activations. Therefore, there is no need for considering the variability of neural estimates and updating the new parameters through the acquisition. In the fMRI experiments, however, we chose the second strategy that

updates neural estimates for every trial to make ADO use the best information available.

One-trial-lag Optimization. Ideally, we should use both neural and behavioral data from all previous trials. However, when we use typical lengths of interstimulus or intertrial intervals, obtaining neural estimates of the latest trial before computing global utility is almost impossible due to the temporal profile of hemodynamic responses.

In detail, the hemodynamic responses consist of an increasing period to a peak that takes 5-6 seconds, a decreasing period with an undershoot below a baseline activation, and a slow asymptotic recovery period. The total length of a hemodynamic response usually takes up to 30 seconds. As our main interest is the activation amplitude, we need to measure BOLD responses for a specific stimulus or trial for at least 5-6 seconds to characterize their peak intensity. However, a temporal lag of 5-6 seconds might be too long depending on stimulus presentation settings (i.e., stimulus duration, interstimulus/intertrial interval). In this case, we can collect a behavioral response but not a neural activation estimate at the end of the trial.

One possible solution for the loss of neural data is to use the neural and behavioral data obtained by the $(t-1)$-th trial to generate the optimal proposal for $(t+1)$-th trial, a strategy we refer to as 'one-trial-lag Adaptive Design Optimization (ADO)'. Figure 2 describes how one-trial-lag ADO works. For example, the first trial uses an ADO proposal that is derived by the prior distribution of model parameters, whereas the second trial uses randomly generated designs since the neural estimates from the first trial are not available at this point. During the second trial, the single-trial neural activation of the first trial is estimated and used together with behavioral data to compute the optimal design for the third trial. Similarly at the third trial, ADO uses the data obtained by the second trial (green blank rectangle) to generate the optimal proposal for the fourth trial.

The method described above was used in the simulation study as an 'ideal' schedule of imposing a lag because we can exploit ADO in as many trials as possible. However, we can also simplify the implementation of one-trial-lag ADO using randomly generated designs for the first few trials, which is the strategy used in the fMRI experiment. Compared to the

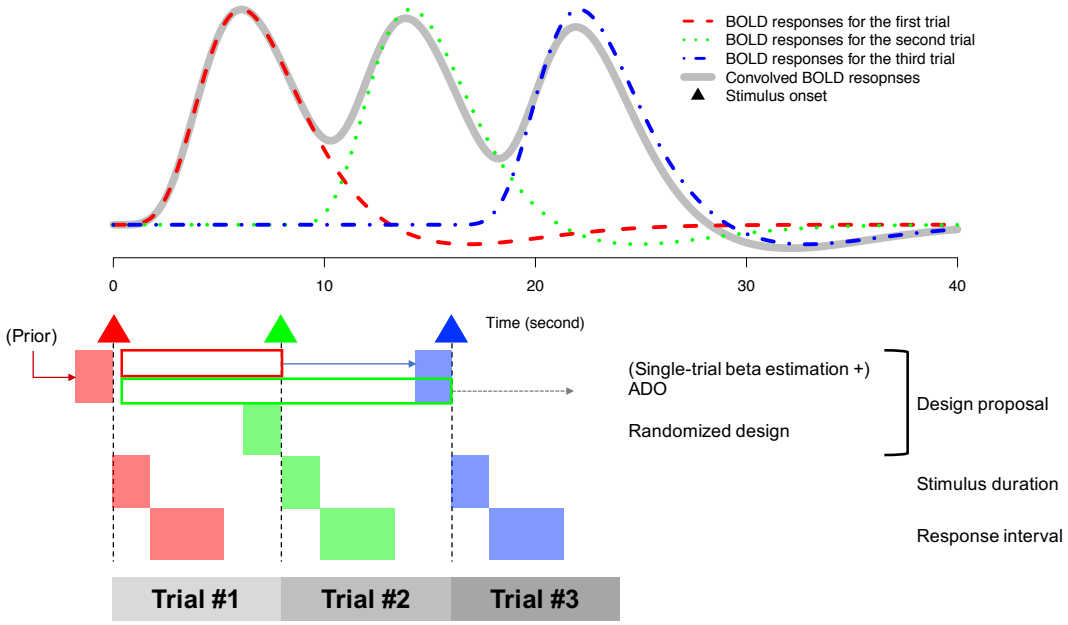

Figure 2. **Conceptual illustration of one-trial-lag ADO**. Dotted lines (red, blue, and green) refer to hypothetical hemodynamic responses evoked by a stimulus within each trial, and a straight line (gray) shows the expected value of convolved hemodynamic responses. The squares below the x-axis specifies the length of intervals required for each step.

method described above, the latter might be preferred from the perspective of controlling variability of neural estimates. As enough neural data have been collected in the first few trials, the neural estimates corresponding to the first few trials have already stabilized. As one-trial-lag ADO relieves us from burdensome computational time when acquiring single-trial beta estimates, we recommend using this procedure when single-trial beta estimates must be obtained to characterize the BOLD response.

Refining the Functionality of fMRI-based ADO

Posterior Sampling via DE-MCMC. In the practice of Adaptive Design Optimization, full posterior estimation of model parameters may be required in real-time for two reasons: evaluation of the performance of ADO and adaptive updating of the grid points. In this study, we used a Differential Evolution Markov chain Monte Carlo sampler (DE-MCMC; ter Braak, 2006; B. M. Turner, Sederberg, et al., 2013) for posterior updating.

DE-MCMC sampler uses information about the difference between chains to draw new posterior samples, enabling it to sample more efficiently from models with correlated dimensions. In addition, DE-MCMC sampler suffers less from autocorrelation in the sampling process than conventional Metropolis-Hastings algorithms.

To initialize the chains of the sampler, we used the grid points as a reference. In detail, initial chains were selected by multinomial sampling with a choice probability vector $\mathbf{p}^{(t)}$ constructed by normalized posterior densities of all grid points in the parameter space. Given the j-th grid point in the search space at trial t, $\theta_j^{(t)}$, and the total number of grids J, the i-th chain initialized after completing the t-th trial, $c_{i,t,1}$, is initialized by multinomial sampling:

$$\begin{aligned} c_{i,t,1} &\sim \text{Multinomial}(\mathbf{p}^{(t)}) \\ &\equiv \text{Multinomial}([p_1^{(t)}, p_2^{(t)}, \cdots, p_J^{(t)}]^T) \end{aligned}$$

Here, the probability that the j-th grid point is selected as an initial chain is

$$p_j^{(t)} = \frac{f(\theta_j^{(t)}|y_{1:t}, d_{1:t})}{\sum_{j=1}^{J} f(\theta_j^{(t)}|y_{1:t}, d_{1:t})}.$$

At the $(i-1)$-th iteration, given the chains from the previous iterations $c_{\cdot,t,i-1}$, DE-MCMC proposes a posterior sample with the following procedure. First, the sampler randomly selects two different chains, say $c_{m,t,i-1}$ and $c_{n,t,i-1}$, and take their difference: $\Delta c = c_{m,t,i-1} - c_{n,t,i-1}$. Second, a proposal based off on the third chain $c_{q,t,i-1}$ $(q \neq m, n)$ is generated by adding Δc scaled by a pre-specified factor γ and random perturbation ϵ to it. If this proposal passes the test by the Metropolis-Hastings probability, the new proposal is accepted as a posterior sample. If not, the previous sample is used again.

However, poor initialization can cause problems in the posterior due to “outlier" chains that deviate from the majority of the chains. Migration (Hu & Tsui, 2005) could be a reasonable remedy to solve this problem by swapping the location of outlier chains

during the first few trials with fixed probability. In addition, DE-MCMC can force the sampling procedure to focus more on the high-density region (this is called “burn-in" mode; B. M. Turner & Sederberg, 2012) so that we can center the posterior around its maximum a posteriori (MAP) estimate. For more details, we direct readers to publications investigating these ideas B. M. Turner and Sederberg (2012); B. M. Turner, Sederberg, et al. (2013).

Dynamic Gridding. The current implementation of fMRI-based Adaptive Design Optimization (ADO) relies on a grid-based method to approximate the global utility calculation. For efficient performance of ADO, we need to discretize both parameter and response spaces appropriately. Theoretically, an obvious first choice is to define a dense grid over a broad range of values in both parameter and response spaces. However, a tradeoff ensues between the number of grid points and computational efficiency due to multidimensionality of the grid space. Adding only one more grid point per dimension will result in an explosive increase of the number of grid points in the entire search space. Hence, simply specifying a very dense grid is not an appropriate solution.

Another disadvantage of the dense grid space is redundant grid points in low posterior density regions. Global utility based on mutual information relies on posterior densities obtained at each grid point. Joint posterior distributions of model parameters will be constrained as the experiment proceeds, and therefore the number of grid points with extremely small posterior density (i.e., $p(y|\theta, d)$) will increase. In the end, most of the grid points cannot contribute to generating new proposals due to small posterior densities, which makes computation and aggregation of global utility values inefficient.

One possible solution is to update the grid as the posterior distribution is updated. This approach allows ADO computation to be affordable with limited computing resources while achieving better efficiency. Implementation of this solution requires a method for automatically adjusting the distribution of grids to capture a region with high posterior density.

Here, we used a simple method based on eigendecomposition of a sample covariance

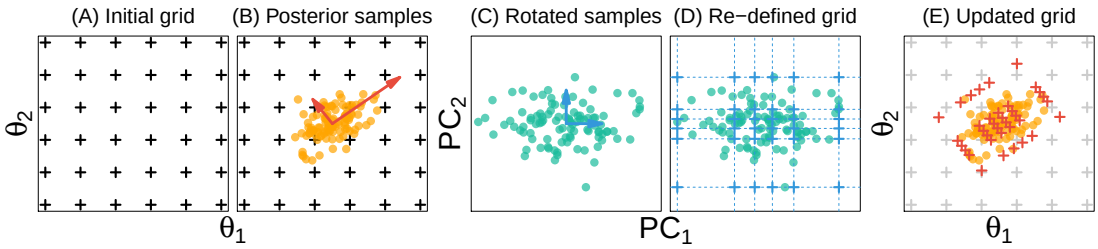

Figure 3. **Visual illustrations of eigendecomposition-based dynamic gridding**. ADO starts with an initial grid setting (A) and obtains posterior samples using an MCMC sampler (B). The covariance matrix from the posterior samples provides information about eigenvectors (red arrows in B), which enables rotation of the posterior samples to align them orthogonally (C). New grid points are defined for each dimension based on pre-specified percentiles (D). The eigenvectors of the covariance matrix allow rotation of the new grid back onto the original parameter space (E).

matrix motivated by principal component analysis (Johnson & Wichern, 2007). The main idea is that we can compute the sample covariance matrix S from the posterior samples obtained by MCMC procedures and decompose it into eigenvectors and associated eigenvalues. These eigenvectors provide an appropriate rotation scheme to orthogonalize the posterior samples. Figure 3 provides visual illustrations of the dynamic gridding procedure described here.

The result of eigendecomposition of S consists of two matrices – a square matrix R containing eigenvectors of S as its columns, and another diagonal matrix C whose diagonal elements are eigenvalues of S:

$$S = RCR^{-1}.$$

Because eigenvectors in R construct an orthogonal basis explaining the largest variance of the posterior samples, we can use R to map the original posterior samples, say A, onto an orthogonal principal component space without additional scaling: $\tilde{A} = AR$. Then, for each dimension, we can sample quantiles from an empirical marginal distribution given a set of pre-specified probabilities, which defines a new grid in the rotated space. As a last step, an inverse of the rotation matrix R maps the newly defined grid $\tilde{G}$ onto the original space: $G^* = \tilde{G}R^{-1}$. There are several software packages for statistical computing that offer the appropriate functions for implementing these operations (e.g., `eigen` and `quantile` in

R).

Note that this dynamic gridding method can sometimes generate invalid grid points according to assumptions on the model parameters. For example, the standard deviation of a normal distribution, say σ, is not allowed to have negative values by its definition. However, the SVD-based dynamic gridding might allow invalid grid points (i.e., $\sigma < 0$) by the shape of the joint posterior distribution and constraints imposed to other model parameters. These invalid grid points must be ignored in subsequent steps.

Simulation Study

In this section, we aim to provide the simulation-based verification of the performance of fMRI-based ADO. To this end, we first describe the contrast discrimination task that will be used in both the simulation study and the fMRI experiment, and then outline the joint model we used to explain both the neural and behavioral data.

Next, we report the result of one large simulation study we conducted to assess parameter recovery when using ADO-based experiments relative to a randomized design as a baseline. To investigate how well the parameter recovery results generalize, we performed parameter recovery analyses on 30 different parameter sets, each of which produce patterns of data that resemble human decision making in our task. The basic structure is to (1) choose a parameter value for the joint model from the 30-parameter set, (2) perform an ADO-based experiment with the data from each trial being produced by the joint model, (3) perform a Randomized Search based experiment by sampling a pair of contrasts on each trial at random, and (4) compare the parameter posterior estimates obtained in each experiment sequence. For (4), we compare the estimated parameter posteriors in terms of their accuracy (i.e., distance from the true parameter value) and precision (i.e., the variance in the estimated parameter posterior).

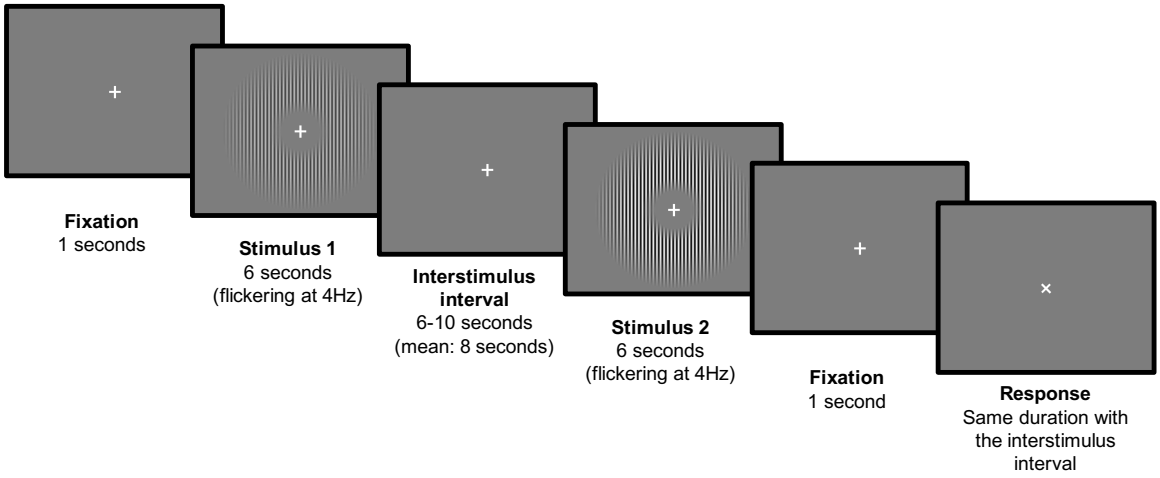

Figure 4. **The trial structure of the contrast discrimination task.** After the fixation for one second, a participant is presented two grating stimuli with different contrasts consecutively. Each stimuli is presented for six seconds flickering at 4Hz. The mean length of interstimulus and intertrial intervals is 8 seconds. The mean length of interstimulus and intertrial intervals in Randomized Search and Adaptive Design Optimization are 8 and 12 seconds, respectively.

Task

In the contrast discrimination task, a participant is presented two grating annuli consecutively, each having different contrast levels. Following the stimuli, a response cue is presented and the participant is instructed to respond by indicating which of the two stimuli were of higher contrast by pressing the corresponding button. Figure 4 illustrates the trial structure of the contrast discrimination task.

Contrast levels are defined in the interval $[0, 1]$. When the contrast level is 0, the stimulus is completely flattened and shown as a gray plane. When the contrast level is 1, the stimulus shows a fluctuating black-white stripe pattern. In the experiment, the contrast values are logarithmically spaced with 10 levels (i.e., 0.010, 0.017, 0.028, 0.046, 0.077, 0.129, 0.215, 0.359, 0.599, 1.000). We also restricted the experimental design such that no two stimuli had exactly the same contrast.

Developing a Joint Model of Contrast Discrimination

We have developed a joint model that relies on the Naka-Rushton equation and a Thurstonian decision model to describe how the neural response manifests in the discrimination experiment. In the model, neural activation (and therefore the amplitude of BOLD responses) is assumed to monotonically increase with increases in the contrast level, specifically, according to the well-known Naka-Rushton equation (DiMattina, 2016; Li, Lu, Tjan, Dosher, & Chu, 2008). From this prediction of stimulus-induced neural activations, a Thurstonian decision model is used to predict behavioral choice responses in the binary discrimination task (Thurstone, 1927). Both decisions about the functional form of the neural measures and the discrimination model were chosen because of their general applicability to other decision-making tasks, so that the results presented here could be readily adapted to other experimental task settings.

Neural Submodel. To describe the relationship between the contrast and activation of visual cortex, we use Naka-Rushton equation (DiMattina, 2016; Li et al., 2008). Given the two contrast levels c_1 and c_2, Naka-Rushton equation models predicted neural activation levels using three shape parameters (b, R_{max}, c_{50}):

$$\hat{\beta}_i = b + \frac{R_{max} c_i^2}{c_{50}^2 + c_i^2} \quad (i = 1, 2) \tag{8}$$

where b is baseline activation, R_{max} is the maximum amplitude above the baseline, and c_{50} is the contrast level that evokes half the maximum activation. We assume that the actually measured neural activation β_i is normally distributed with mean $\hat{\beta}_i$ and constant standard deviation $\delta/\sqrt{2}$:

$$\beta_i \sim N\Big(\hat{\beta}_i, (\delta/\sqrt{2})^2\Big). \tag{9}$$

Behavioral Submodel. On the behavioral side, we use a Thurstonian decision model (Thurstone, 1927) to model the discrimination process. Let us assume that the perceptual system represents the physical stimuli (i.e., the two grating stimuli) with inten-

sity ϕ_1 and ϕ_2 as ψ_1 and ψ_2 as a normally distributed random variable centered on the true physical state, but with some perceptual uncertainty s such that

$$\psi_i \sim N\Big(\phi_i, s^2\Big) \quad (i = 1, 2). \tag{10}$$

Then, we make a comparative judgment based on the difference between two mental representations, say $\psi_2 - \psi_1$. Hence, the difference of the two psychological variables can be written as

$$\psi_2 - \psi_1 \sim N\Big(\phi_2 - \phi_1, (\sqrt{2}s)^2\Big). \tag{11}$$

Given this difference distribution, we assume a behavioral response y is given according to a Bernoulli distribution

$$y \sim \text{Bernoulli}(p)$$

with probability p determined by the psychological mapping of the two physical intensities such that

$$p = 1 - \Phi^*\Big(0; \phi_2 - \phi_1, (\sqrt{2}s)^2\Big), \tag{12}$$

where $\Phi^*(\cdot; \mu, \sigma^2)$ is a cumulative density function of a Gaussian distribution with mean μ and standard deviation σ. Hence, $y = 1$ when our psychological experience suggests that $\phi_2 > \phi_1$.

A Linking Function. Any joint model requires a linking function that mathematically expresses the relationship between the neural and behavioral submodels. As a linking function, we simply assume that the neural encoding of the contrast stimuli works as a mental representation of the contrast level (i.e., $\phi_i \equiv \hat{\beta}_i$, $\psi_i \equiv \beta_i$). In addition, we assume that the uncertainty in behavioral responses δ is affected by the variability of neural activation as in Equation 9. Therefore, the complete joint model of contrast discrimination comprising of four parameters $(b, R_{max}, c_{50}, \delta)$ can be described as follows:

$$\beta_2 - \beta_1 \sim N(\hat{\beta}_2 - \hat{\beta}_1, \delta^2),$$
$$p = 1 - \Phi^*\left(0; \hat{\beta}_2 - \hat{\beta}_1, \delta^2\right) = \int_0^\infty N(x; \hat{\beta}_2 - \hat{\beta}_1, \delta^2)dx,$$
$$y \sim \text{Bernoulli}(p).$$

Methods

To perform grid-based Adaptive Design Optimization (ADO), we need to first specify environmental settings that include (1) prior distributions, (2) initial grid settings, (3) MCMC sampler parameters (e.g., the number of chains, burn-in steps and valid iterations), (4) dynamic gridding parameters. Tables 1 and 2 show the default settings and parameter sets used in the simulation study.

As for the levels of the contrast, we used ten logarithmically spaced points for each stimulus per trial. As we used two stimuli for each trial and excluded the designs where the first and second stimuli shared the same contrast, the design space consists of $10^2 - 10 = 90$ candidate designs. For a grid-based approximation of the parameter space, we decided to use five points per dimension. Therefore, the number of points in the parameter space is $5^4 = 625$.

If the response variable of interest relies on discrete measurements, we do not need further approximations for grid-based ADO because the response variable itself is already discretized. However, if the response variable is continuous, grid-based ADO requires discretization of the response space. In this simulation, we set ten levels of neural activation amplitudes for this approximation. As we used two neural measures per trial plus one binary choice, the discretized response space consists of $10^2 \times 2 = 200$ points.

When specifying the prior distributions, we could use non-uniform priors such as diffuse normal distributions for b and R_{max}, a truncated normal or beta distribution for c_{50}, and an inverse-gamma distribution for δ. However, we decided to use uniform priors to reduce computation time as much as possible, as we evaluated posterior densities every

trial with newly updated single-trial neural estimates (see Section for more details) or grid points.

In the simulation study, we defined measures of accuracy and precision of posterior estimates by root mean square deviation (RMSD) and standard deviation (PSD) of the posterior distribution. We considered mean values of the posterior samples as posterior estimates as in Equation 13, and then computed parameter-wise standard deviation ($PSD_{i,t}$) and pooled performance measures ($RMSD_t$ and PSD_t) at each trial t as follows: Given a set of "true" parameters assumed in each simulation $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \equiv (b, R_{max}, c_{50}, \delta)$, and x_{ijkt} representing a value of the j-th chain of the DE-MCMC sampler for the parameter θ_i at the k-th iteration ($j = 1, \cdots, 24$),

$$\bar{x}_{i\cdot\cdot t} = \frac{1}{24 \times 800} \sum_{k=201}^{1000} \sum_{j=1}^{24} x_{ijkt}, \tag{13}$$

$$RMSD_t = \sqrt{\sum_{i=1}^{4} (\bar{x}_{i\cdot\cdot t} - \theta_i)^2},$$

$$PSD_{i,t} = \sqrt{\frac{\sum_{k=201}^{1000} \sum_{j=1}^{24} (x_{ijkt} - \bar{x}_{i\cdot\cdot t})^2}{24 \times 800}},$$

$$PSD_t = \sqrt{\frac{1}{4} \sum_{i=1}^{4} PSD_{i,t}^2} \tag{14}$$

. The DE-MCMC sampler drew posterior samples for 1,000 iterations and discarded the first 200 iterations as burn-in.

Results

As our simulation involves randomness both within a given parameter set and between parameter sets, we present the results in two phases. Figures 5 and 6 illustrate the results for a single parameter combination within the set. First, Figure 5 compares design proposals from ADO (top row) and Randomized Search (RS; bottom row). Each dot represents a design candidate, and the relative intensity conveys the frequency of each stimulus

Variable		Details
The number of replicates		100 for each parameter set
The number of trials		20
Stimulus (Rounded to 3 decimal places)		{0.010, 0.017, 0.028, 0.046, 0.077, 0.129, 0.215, 0.359, 0.599, 1.000}
Prior	b	Uniform(-3, 5)
	R_{max}	Uniform(-3, 5)
	c_{50}	Uniform(0, 1)
	δ	Uniform(0.0001, 5)
Initial grid settings	b	{-2, -1, 0, 1, 2}
	R_{max}	{0.5, 1.125, 1.75, 2.375, 3}
	c_{50}	{0.05, 0.275, 0.5, 0.725, 0.95}
	δ	{0.001, 0.30075, 0.6005, 0.90025, 1.2}
	Neural response	{0, 0.22, 0.44, 0.67, 0.89, 1.11, 1.33, 1.56, 1.78, 2}
Grid size	Design space	$90 = 10^2 - 10$
	Parameter space	$625 = 5^4$
	Response space	$200 = 10^2 \times 2$
DE-MCMC	Chains	24
	Burn-in samples	200
	Valid posterior samples	800
	Migration probability	0.1
Dynamic Gridding	Method	Eigenvector-based rotation
	Schedule	After every trial
	Percentile	(20%, 35%, 50%, 65%, 80%)

Table 1
Default settings in Simulation Study

selection. Each column represents a different block of trials: 1-5 (left), 6-10 (middle), and 1-20 (all trials; right). As expected, the bottom row shows that RS selects design candidates (i.e., pairs of contrast values) with equal frequency. However, the ADO search selects design candidates with different frequencies over trials.

Figure 6 compares ADO (red) to RS (black) designs in terms of accuracy (left panel), precision (middle panel), and effective differences between the designs in terms of number of trials (right panel). For accuracy, we compared ADO to RS by computing the pooled root mean squared deviation (RMSD; left) between each estimated parameter posterior to the true parameter set. For precision, we compared ADO to RS by computing the pooled standard deviation (PSD) of each estimated posterior distribution. For both RMSD and PSD,

Set	Parameter values				Set	Parameter values			
	b	R_{max}	c_{50}	δ		b	R_{max}	c_{50}	δ
1	0.050	1.000	0.350	0.200	16	0.200	1.631	0.180	0.206
2	0.345	1.473	0.136	0.263	17	-0.009	2.026	0.156	0.297
3	0.371	1.544	0.203	0.203	18	0.454	1.678	0.194	0.356
4	0.378	1.750	0.114	0.390	19	0.269	1.220	0.122	0.368
5	0.233	1.340	0.391	0.303	20	0.134	1.173	0.107	0.421
6	0.206	2.078	0.374	0.257	21	0.018	1.123	0.165	0.373
7	0.210	2.199	0.177	0.463	22	0.423	1.351	0.208	0.432
8	0.302	1.287	0.248	0.345	23	0.480	1.706	0.147	0.402
9	0.012	1.480	0.239	0.310	24	0.402	1.835	0.232	0.261
10	0.025	1.620	0.262	0.409	25	0.204	1.999	0.314	0.242
11	0.277	1.809	0.395	0.462	26	0.030	1.527	0.284	0.206
12	0.136	1.321	0.179	0.457	27	0.057	1.048	0.126	0.317
13	0.393	1.937	0.118	0.282	28	0.086	2.152	0.357	0.430
14	0.362	1.823	0.352	0.374	29	0.176	1.813	0.343	0.421
15	0.235	2.186	0.357	0.466	30	0.083	2.054	0.267	0.493

Table 2
A list of 30 parameter sets used in Simulation Study. Parameter values are rounded to three decimal places.

smaller values are preferred. In both panels, we extrapolated the metrics corresponding to the RS design by extending the simulation by 10 trials. Across both panels, ADO clearly outperforms RS, attaining a smaller RMSD and PSD across all parameter sets. The right panel of Figure 6 extends the comparison illustrated in the left and middle panels; on each trial, we computed how many additional trials (y-axis) would be needed using RS to attain a similar RMSD (plus signs) and PSD (open circles) as a function of trial number (x-axis). This comparison shows compelling advantages for ADO search. For example, 10 trials worth of ADO search is roughly equivalent to 17-18 trials of RS, and 20 trials of ADO is roughly equivalent to 32 trials of RS.

Although the results of the within-parameter set analysis are encouraging, they lack generalizability across different brain-behavior relations. To this end, we can extend the analysis by aggregating the performance metrics shown in Figure 7 across 30 different parameter sets. Figure 7 shows scatter plots of the RMSD (left) and the PSD (right) to compare the performance of ADO (x-axis) to RS (y-axis). The gray shaded area indicates

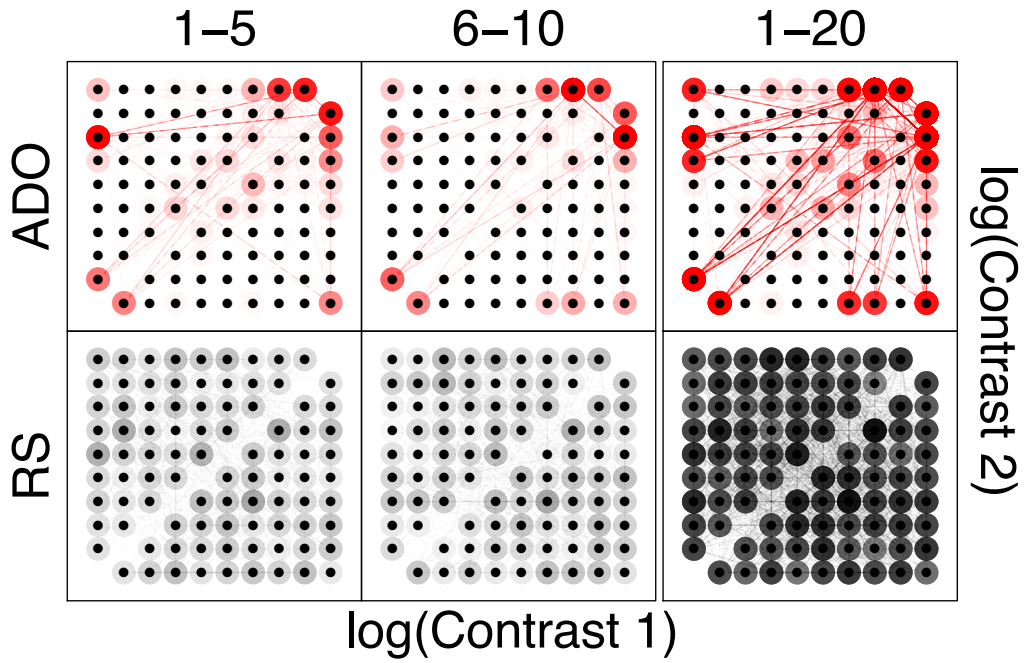

Figure 5. **Simulation Results from the Parameter Set 1**. The figure shows a path analysis comparing Adaptive Design Optimization (ADO; top panel) against Randomized Search (RS; bottom panel) separated by Trials 1-5 (left column), Trials 6-10 (middle column), and all trials (right column). Frequency of stimulus selection is indicated by intensity of the circles, where the first and second stimuli are shown on the $x-$ and $y-$axes, respectively. The labels for two axes were intentionally omitted for visual clarity.

regions of each metric space where the performance of ADO was superior to RS. In general, a significant proportion of the metrics ($\approx 71 - 75\%$ at maximum) are located above the identity line, and therefore we can conclude that ADO outperforms RS across these 30 parameter sets.

One feature of the aggregated results is that the performance metrics comparing ADO to RS tend to converge as the number of trials increase (e.g., Trial 20, purple contour in Figure 7). This is a well-established effect in design optimization: once enough data are collected, the benefits provided by ADO asymptote depending on the number of stimuli to choose from and the complexity of the cognitive model. In our case, as the experiment and model are both relatively simple, we should expect RS to eventually catch up to ADO beyond approximately 20 trials. However, substantially better ADO results would be

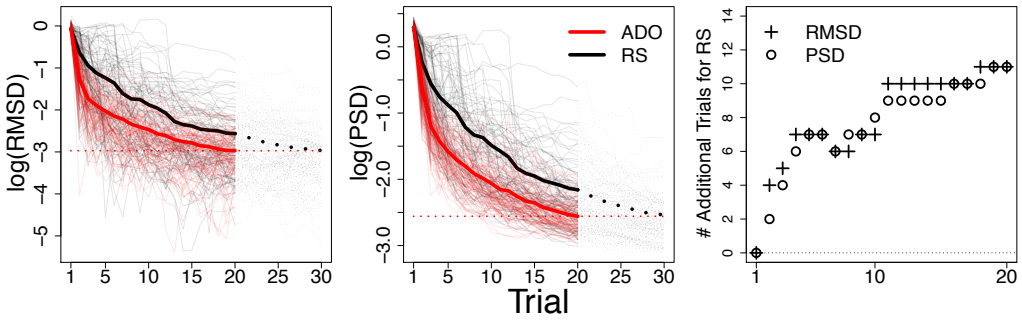

Figure 6. **Simulation Results from the Parameter Set 1**. The figure shows performance metrics comparing ADO (red) to RS (black) in one of the parameter sets tested in the simulation. The left and middle panel compares the experimental searches in terms of accuracy and precision by plotting the pooled root mean squared deviation (RMSD) and posterior standard deviation (PSD) of the estimated parameter posteriors, respecitvely. Semi-transparent lines represent individual results from 100 simulations for each method, whereas bold solid lines represent the average performance. Smaller values are preferred for both accuracy and precision. In RS experiments, results for additional 10 trials are shown to compare long-term mean performance of RS (black bold dotted lines) to the mean performance of ADO at the 20th trial (red dotted lines). The right panel shows the number of additional trials required for RS experiments to attain equivalent mean performance with the ADO algorithm in terms of RMSD (plus signs) and PSD (open circles).

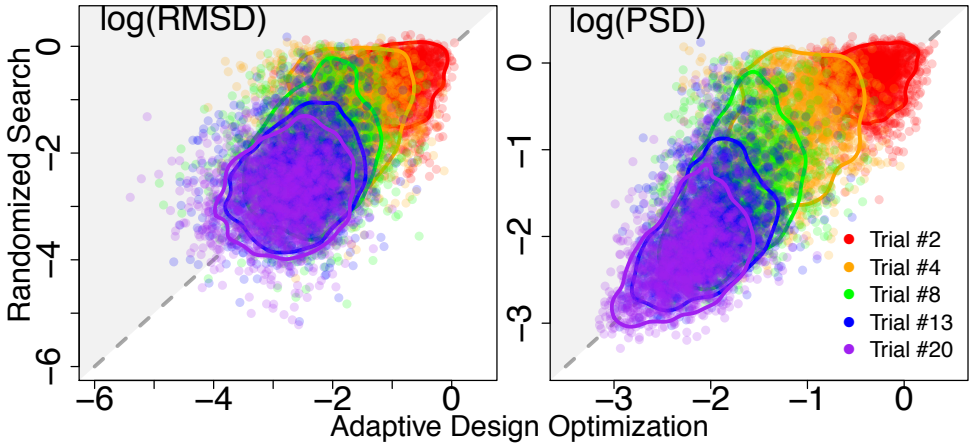

Figure 7. **Summary of the Simulation Results**. Scatter plots show the performance of ADO (x-axis) relative to RS(y-axis), based on RMSD (left panel) and PSD (right panel) aggregated across 30 parameter sets. In each panel, the distribution of each performance measure and its mean are shown as a contour plot and the "$\times$" marker, separated by blocks (see legend for details).

realized with either more candidate stimuli or a more detailed cognitive process model. Regardless, the main result is that the performance of ADO is better during the first few trials, suggesting that a stopping rule could be developed to facilitate more efficient data collection relative to RS.

fMRI Experiment

The result of the simulation study suggested that ADO supported by both neural and behavioral data can estimate model parameters more efficiently than a baseline RS procedure does. To validate the method in a real-world application, we compared the efficiency of ADO relative to RS in an fMRI experiment. Our goal was to establish the performance of ADO both across participants (i.e., between-participant), and within the same participant across different scanning sessions (i.e., within-participant).

Participants

Four participants completed the experiment. Each participant had three two-hour sessions including 90-minute functional MR scanning. Two among four participants were female, and the mean age of participants was 24.75. All participants were recruited from The Ohio State University and provided informed consent. The study was approved by the Institutional Review Board of The Ohio State University.

Stimuli and Task

All stimuli and instructions were generated by SMILE (State Machine Interface Library for Experiments; `http://smile-docs.readthedocs.io/en/latest/`), a Python library for programming psychological experiments, on a MacBook Pro 2016. Each participant laid on the scanner bed and viewed the stimuli presented onto a rear-projection screen through a mirror mounted in the coil. Stimuli were presented at eye level at a distance of 74cm.

Each grating stimulus was generated with spatial frequency of 3.06 cycles per degree, and formed as an annulus not to expose the grating patterns at fovea. The radii of the

external and internal circles were 14.52 degree and 3.48 degree in visual angle, respectively. In addition, a linear mask was applied to the annulus to allow gradual changes in stimulus intensity. The stimulus intensity increases from a distance of 1.74 degree reaches its maximum at a distance of 2.94 degree, and fades gradually from a distance of 4.34 degree from the center of screen.

A participant was presented two consecutive grating stimuli with different contrast levels and asked to keep fixation at a white "+" marker located at the center of a screen. When the fixation marker changed to a response cue (a white "$\times$" marker), the participant was asked to answer whether the first or the second stimulus was of higher contrast. The participant was given two 2-button response pads, one for each hand, and was instructed to use one button for each side to make a response. The response-button association rule altered every session. For example, a participant was asked to use the button in the left box to respond that the first stimulus had higher contrast level in one session, and to use the button in the right box to make the same response in the next session.

Each participant performed the same task over three separate scanning sessions, each lasting about 90 minutes. Within each of the three independent-replication sessions, participants completed two conditions: in one condition the stimulus sequence was generated based on RS, whereas in the other condition it was generated based on ADO. Due to participant dropout, the order between the ADO-based and RS-based runs was not counterbalanced. Participants 1 and 2 conducted the ADO-based runs first in the first and third replicate sessions, and the RS-based run first in the second session. Participant 3 conducted the RS-based run first in the first and third replicate sessions, and the ADO-based run first in the second session. Participant 4 conducted the ADO-based run first in the first replicate session, and the RS-based run first in the two remaining sessions.

The difference between the RS-based and ADO-based runs is the length of intertrial interval. ADO requires time to calculate an optimal design at the end of every trial, and for adjusting parameter grids after the 4th, 8th, 12th, and 16th trials. Specifically, fMRI-based ADO in this experiment requires 6-8 seconds for proposing the optimal design and

additional 4-5 seconds for full posterior estimation and grid adjustment. Therefore, 8 seconds of the mean intertrial interval used in the RS-based experiment was not enough in the ADO-based run. While the intertrial interval of the run without ADO was either 6, 8, or 10 seconds, that of the ADO-based run was extended for 4 seconds (i.e., 10, 12, or 14 seconds). The total length of the run without ADO was 624 seconds. The ADO-based run took approximately 15 minutes. Finally, we reduced the grid size for each dimension in the neural response from 10 to 7 grid points to reduce computational burden. Therefore, the size of the grid for the response space considering two neural responses and one behavioral response is now $7^2 \times 2 = 98$.

Protocol

Figure 8 provides a graphical summary of the scanning protocol and data flow in the ADO-based fMRI experiments. The experiment comprises of three stages: (1) acquisition of structural and functional localizer images, (2) inverse-registration of anatomical masks onto a standard space, and (3) data collection in the main task.

The first stage aims to collect information required for producing a task-specific mask in the subject-specific brain space. After completing set-up for online data transfer from an MR scanner to a terminal computer, an experimenter needs to collect structural images of a participant's brain and acquire a regional localizer based on an echo-planar imaging (EPI) sequence. The former constructs the basis of the subject space, whereas the latter limits the region to be scanned in the functional localizer and the main tasks. The functional localizer task is performed to detect task-relevant voxels as the last step. The functional localizer mask can be defined by performing a whole-brain GLM analysis with data from the localizer task and extracting voxels that have test statistics (e.g., t-statistics) greater than a specific threshold.

In the second stage, an experimenter extracts the task-relevant subject-specific mask using the data acquired from the first stage. We use a template structural image defined in a standard brain space such as MNI (Montreal Neurological Institute) atlas (Grabner et

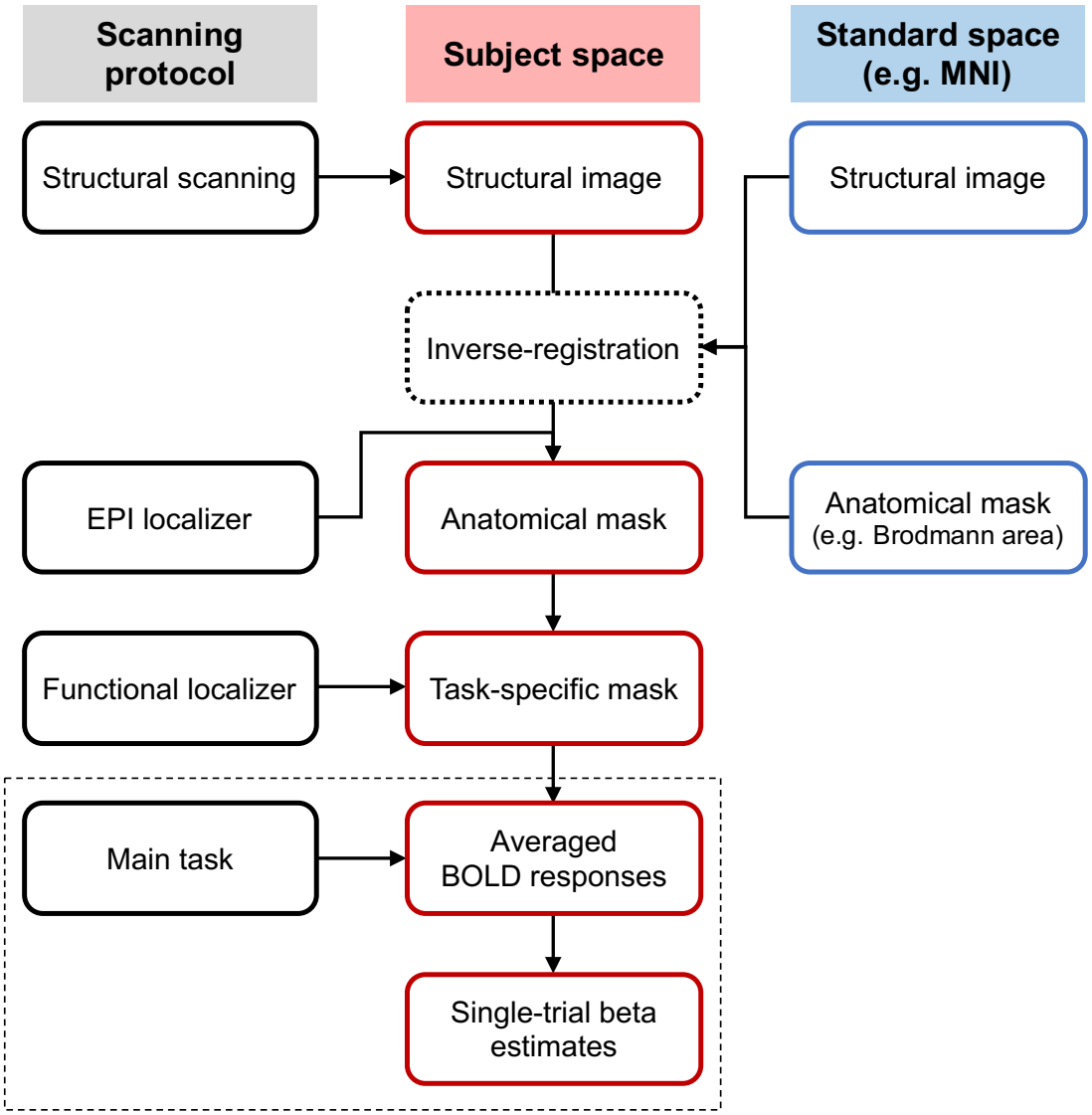

Figure 8. **The scanning protocol and data flow used in ADO-based real-time fMRI experiments**. The left column represents scanning protocols that should be set in the terminal computer that controls the MR scanner. The right column represents the template brain images that must be prepared before the experiment. The center column represents the data that we acquire from a participant using raw MR images, template brain images, and the appropriate computations on them.

al., 2006) as a reference. Once the experimenter collects the structural image in the subject space, it is registered to the standard brain template to obtain the transformation matrix that maps the subject space onto the standard space. The inverse-transformation matrix is derived by taking an inverse of the transformation matrix, and is used for mapping the

anatomical masks in the standard space to the subject space. When regions of interest (ROIs) must be constrained by masks provided by standard anatomical atlases (e.g., Jülich Histological Atlas; Eickhoff et al., 2005), we can transform the standard masks to subject-specific masks by using the inverse-transformation matrix. The conjunction between the inverse-transformed anatomical mask and the functional localizer mask defines the task-relevant mask in the subject space.

The task-specific mask enables one to obtain voxel-wise BOLD responses in real-time during the main task. When an experimenter is interested in a specific ROI defined by the task-relevant mask, a common approach is to average neural signals from all voxels in the mask for running the GLM analysis for stimulus-wise neural estimates. The stimulus-wise neural activation estimates are considered as neural inputs of ADO.

Our report will focus on the optimization during the main task. Regarding how we performed the functional localizer task and determined the voxels of interest, readers are referred to Appendix A.

Definition of the Benchmark and Distance Metrics

Unlike the simulation study, we don't have a "true" parameter that serves as a benchmark to compare the performances of ADO and RS, especially when the focus of our analysis is on accuracy. Therefore, we decided to use the posterior estimate obtained by using all the data from both ADO-based and RS-based runs within a session as a benchmark. We can justify this approach for two reasons: (1) the stimulus-wise neural activation estimates from ADO-based and randomized-design runs capture the neural activity of the same visual system, and (2) the uncertainty of model parameters will be most reduced by using all the available data. The variability of stimulus-wise neural activation estimates may raise questions about the first assumption because ADO might cause adaptation to repeatedly presented stimuli compared to randomized designs (Krekelberg, Boynton, & van Wezel, 2006). However, we suggest that using the combined data is the most reasonable way to establish a standard for performance evaluation given the constraints in our data

analysis.

Once the posterior samples from the ADO, RS, and benchmark settings were obtained, we computed the estimates used for performance evaluation. We originally intended to calculate a four-dimensional joint MAP estimate using multidimensional kernel density estimation. However, the currently available methods (e.g., Duong, 2007) either required substantial computation time or were very susceptible to slight differences in posterior samples. Therefore, we computed MAP estimates using an Epanechnikov kernel for each parameter, and used them in the offline analyses.

Again, we denote the parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \equiv (b, R_{max}, c_{50}, \delta)$. Given estimates obtained at trial t from ADO $\hat{\theta}_{ADO,t} = (\hat{\theta}_{1,ADO,t}, \hat{\theta}_{2,ADO,t}, \hat{\theta}_{3,ADO,t}, \hat{\theta}_{4,ADO,t})$, estimates from RS $\hat{\theta}_{RS,t} = (\hat{\theta}_{1,RS,t}, \hat{\theta}_{2,RS,t}, \hat{\theta}_{3,RS,t}, \hat{\theta}_{4,RS,t})$, and benchmark estimates $\hat{\theta}_{B,t} = (\hat{\theta}_{1,B,t}, \hat{\theta}_{2,B,t}, \hat{\theta}_{3,B,t}, \hat{\theta}_{4,B,t})$, we define the RMSD for each method $m \in \{ADO, RS\}$ as follows:

$$RMSD_{m,t} = \sqrt{\sum_{i=1}^{4} (\hat{\theta}_{i,m,t} - \hat{\theta}_{i,B,t})^2}.$$

The definition of the PSD follows Equation 14, except for the number of iterations in the DE-MCMC sampler. Due to the time concern, we sampled 500 iterations and discarded the first 200 samples as burn-in.

To strengthen our conclusion under the situation where there is no way to know the "true" parameter values, we compared the results in the data space as well as in the parameter space. As for the analysis in the data space, we focused on comparison of Naka-Rushton curves from ADO and RS due to the model structure that the behavioral process (i.e., Thurstonian decision model) depends on the neural encoding process (i.e., Naka-Rushton model).

For the comparison in the data space, we first recovered the shape of Naka-Rushton curves by plugging estimates of Naka-Rushton model parameters $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \equiv (b, R_{max}, c_{50}, \delta)$ into the Equation 8:

$$\hat{N}_{ADO}(c_i) = \hat{\theta}_{1,ADO} + \frac{\hat{\theta}_{2,ADO}{c_i}^2}{\hat{\theta}^2_{3,ADO} + {c_i}^2},$$
$$\hat{N}_{RS}(c_i) = \hat{\theta}_{1,RS} + \frac{\hat{\theta}_{2,RS}{c_i}^2}{\hat{\theta}^2_{3,RS} + {c_i}^2},$$
$$\hat{N}_{B}(c_i) = \hat{\theta}_{1,B} + \frac{\hat{\theta}_{2,B}{c_i}^2}{\hat{\theta}^2_{3,B} + {c_i}^2}$$

where $c = (0.010, 0.017, 0.028, 0.046, 0.077, 0.129, 0.215, 0.359, 0.599, 1.000)$ is the contrast used in the experiments, and $i = 1, \cdots, 10$.

The model fit metrics were defined by root mean squared error from the benchmark estimate: for estimated curves $\hat{N}_{ADO,t}$, $\hat{N}_{RS,t}$, and $\hat{N}_{B,t}$ for trial t,

$$DEVD_{ADO,t} = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left\{\hat{N}_{ADO}(c_i) - \hat{N}_B(c_i)\right\}},$$
$$DEVD_{RS,t} = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left\{\hat{N}_{RS}(c_i) - \hat{N}_B(c_i)\right\}}.$$

Results

The results from Participant 4 are not presented here because of the low quality of the neural data (i.e., the size of the region scanned in the experiment, and excessive head movement), but we refer the reader to Appendix B for equivalent analyses.

Proposed Designs. Figure 9 shows the designs proposed by ADO and RS in the fMRI experiment sessions. Compared to the results from the simulation, the pattern of proposals is not clearly discriminated between the two methods. However, we can see, for example, design combinations of extremely high and low contrasts (e.g., the four corners of each panel) are frequently sampled compared to RS. We can attribute this proposal pattern as an attempt to estimate the baseline parameter b and the maximum amplitude parameter R_{max}.

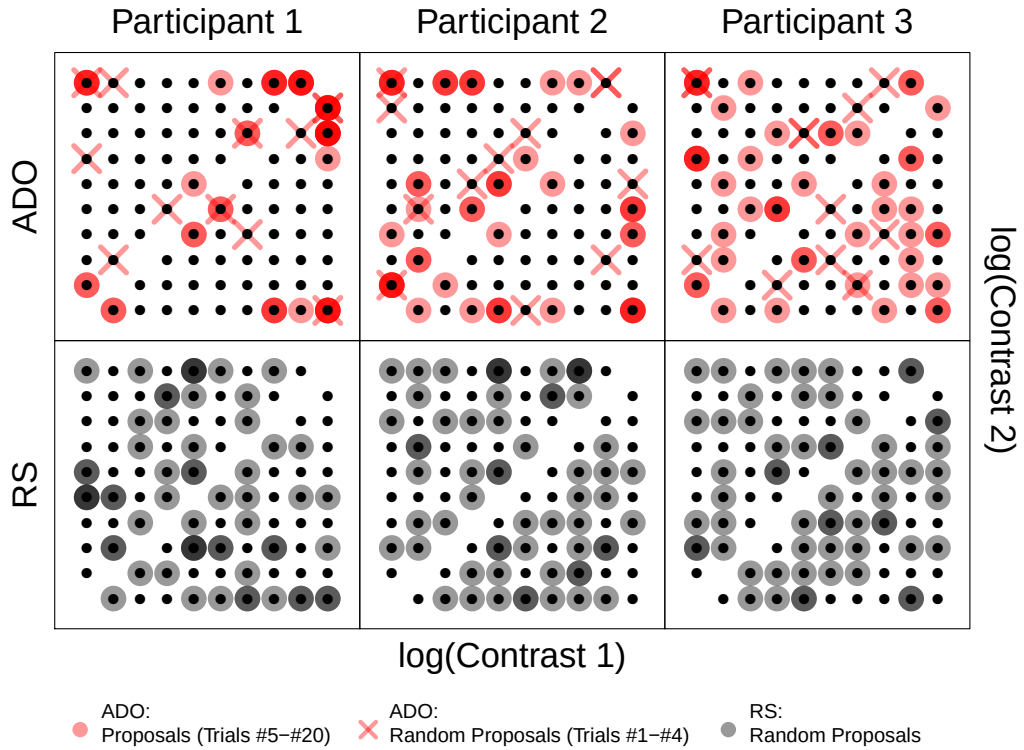

Figure 9. **Proposed Designs from the fMRI Experiment**. The figure shows a path analysis comparing Adaptive Design Optimization (ADO; top panel) against Randomized Search (RS; bottom panel), the column of which corresponds to each participant (left: Participant 1, center: Participant 2, right: Participant 3). Results from all three replicate sessions are collapsed for each participant. Frequency of stimulus selection is indicated by intensity of the circles, where the first and second stimuli are shown on the $x-$ and $y-$axes, respectively. The first four random trials in the ADO-runs are plotted with "$\times$" marks. The labels for two axes were intentionally omitted for visual clarity.

Accuracy and Precision of the Estimates. As in the simulation study, we compared the accuracy and precision of parameter estimates from each method (i.e., ADO, RS) using the RMSD and PSD, respectively. Note that the RMSD was defined with respect to the benchmark parameter.

Figure 10 shows that ADO tends to allow estimates that are closer to the benchmark estimates than RS does. At the 20th trial, the accuracy measures show that ADO outperforms RS in 8 out of 9 scanning sessions. Meanwhile, the result is more mixed in terms of precision and RS tends to perform better than ADO. We suspect that the selective sampling

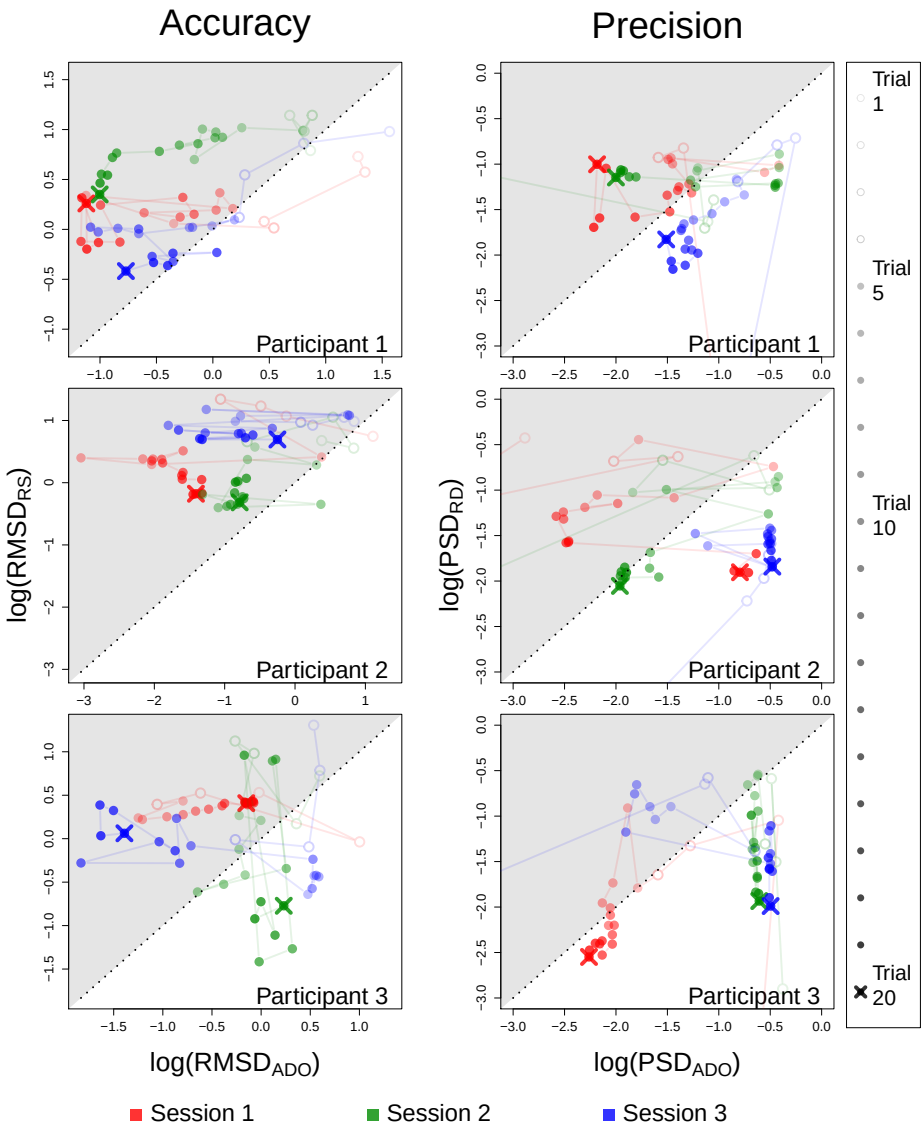

Figure 10. **Results of the fMRI Experiment: Parameter Space**. Performance of the two algorithms is compared in terms of the accuracy with respect to the benchmark estimate (left) and the posterior precision (right). Colored lines with circles and "×" marks represent the accuracy and precision changing over trials. Each row shows the results from diffrent participants (top: Participant 1, middle: Participant 2, bottom: Participant 3). Replicate sessions are color-coded (red: Session 1, green: Session 2, blue: Session 3). Empty dots represent the first four trials at which ADO had to use random proposals. The dot with "×" mark refers to the last trial of each session. The black dotted line represents the identity line. If a point is located in the gray area (i.e., above the identity line), it means that ADO shows higher accuracy or precision compared to RS on that trial. The ranges of both axes were truncated for visual clarity.

procedure of ADO and a low signal-to-noise ratio interacting with the model structure makes precise parameter estimation difficult.

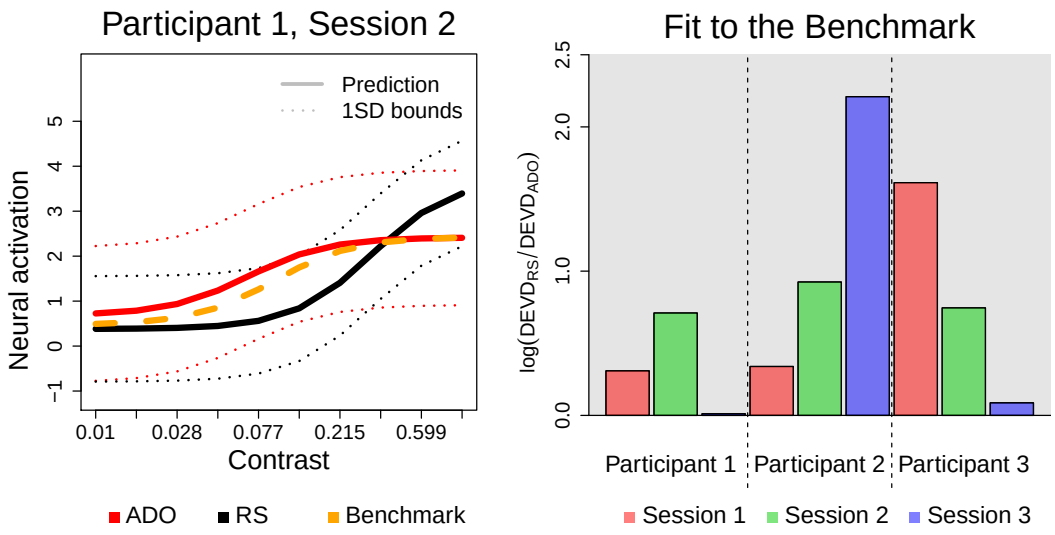

Figure 11. **Results of the fMRI Experiment: Data Space**. Performance of the two algorithms is compared in terms of the accuracy of predicted Naka-Rushton curves for one example participant session (left), and aggregated across all scanning sessions (right). The left plot compare the Naka-Rushton curves predicted from the MAP estimates obtained using Adaptive Design Optimization (ADO; red) and Randomized Search (black). Bold and dotted lines represent the mean prediction and associated standard deviation, respectively. Orange dashed lines represent the benchmark curve, pooled across both runs. The right panel shows the performance of ADO relative to RS in terms of model fit in the data space for each scanning session. To assess accuracy, we first calculated the distance between the model-wise prediction (bold lines) and the benchmark curve (orange dashed line), denoted $DEVD_{ADO}$ and $DEVD_{RS}$. The right plot shows their log-transformed ratio, where higher values support ADO in accuracy (gray area). Note that accuracy metrics are obtained by averaging $DEVD$ excluding the first 4 trials, where random stimuli were presented in both ADO and RS experiments.

Prediction Analysis in the Data Space. Based on the prediction accuracy, we defined the log-transformed ratio of PRED between ADO and RS run as a comparison metric. Note that comparison metrics greater than zero indicate superiority of ADO relative to RS. Figure 11 compares the performance of ADO and RS in terms of accuracy in the data space (right) with a representative example (left, center). In the right panel, each within-participant session is color-coded for clarity. Note that the performance metrics are defined

using the data excluding the first four trials, because the first ADO proposal was used on the 5th trial. Figure 11 shows that in terms of accuracy, ADO tends to outperform RS as all the metrics are greater than zero.

Post-hoc Analysis of Global Utility. A complementary analysis that demonstrates more clearly the superiority of ADO is to compare the amount of information extracted on each individual trial across the ADO and RS procedures. Such an analysis would reveal whether or not ADO was presenting the optimal stimulus on each trial within the run, and similarly, whether or not better stimuli could have been presented during each trial of the RS runs. To address this question, we computed the global utility (i.e., a measure of information) for each possible stimulus that could have been presented on each trial, conditional on the current state of knowledge about the brain-behavior relation (i.e., the joint model). We then normalized the global utility within each trial and compared the ADO and RS results. Figure 12 shows the distribution of global utility values on each trial, where each panel represents a separate participant. Further, each panel is divided into the three runs, where RS runs are illustrated in black and ADO runs are colored according to the run information. Finally, the right-hand side of each panel shows a violin plot of the distribution of global utility across all trials except the first four that used random stimuli.

Figure 12 shows that ADO performs substantially better in terms of trial-level global utility compared to RS. Namely, the utility obtained using ADO was larger than that of the RS in nearly all cases, indicating that ADO extracts better information about how the brain data predicts a behavioral response. Due to the post-hoc nature of this analysis, we could not perfectly account for all of the potential variables that occurred during data acquisition (e.g., variability in neural data, variance in the dynamic gridding process). However, to integrate out as much uncertainty in the data acquisition procedure as possible, we obtained Monte Carlo estimates of global utility by repeating the simulated data acquisition 50 times for each scanning session. Even after considering this additional uncertainty, the normalized global utilities shown in Figure 12 strongly support the performance of ADO relative to RS.

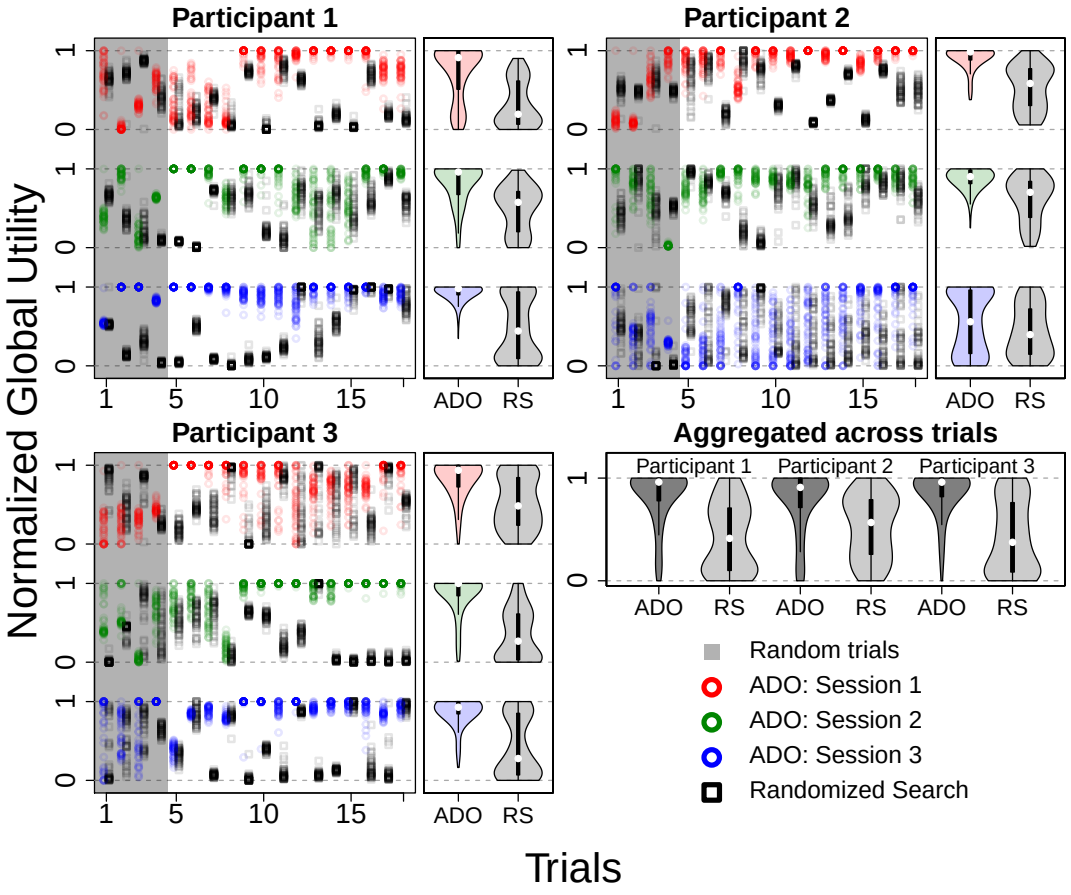

Figure 12. **Analysis of Global Utility Distributions**. Each panel shows the distribution of normalized global utility of all possible stimulus pairs generated by Adaptive Design Optimization (colored plots) and Randomized Search (RS; light gray plots). The lower right panel illustrates the same data after aggregating across trials and runs. For each participant, a scatter plot on the left panel shows how the distribution of normalized global utility changes over trials, whereas a violin plot on the right panel represents the same information aggregated across trials. Note that the first four trials were excluded from the violin plot because RS was used for both search procedures.

Discussion. The fMRI experiment showed mixed results compared to the simulation study. The global utility analysis suggested that ADO proposed stimulus sequences that maximized the expected amount of information. Focusing on the accuracy, the parameter estimates and the predicted Naka-Rushton equation from ADO outperformed those from RS. However, the precision of the parameter estimates of ADO was worse compared to that of RS.

We suspect the selective sampling procedure of ADO might cause inflated uncertainty of the estimates, together with a possibly low signal-to-noise ratio. This is atypical behavior given that the purpose of ADO is to reduce the uncertainty of parameter estimates. However, the unstable nature of the single-trial neural activation estimates might have interacted with the mechanism of ADO and prevented ADO from achieving its goal efficiently. In particular, one of the model parameters δ can be affected by the signal-to-noise ratio, which supports our suspicion. As δ is associated with the degree of deviation from the mean prediction of the Naka-Rushton equation, the low signal-to-noise ratio can propagate to not only δ, but also other shape parameters of the Naka-Rushton equation (i.e., b, R_{max}, c_{50}).

General Discussion

Limitations and Contributions

This study provides a proof of concept of online design optimization for model-based fMRI experiments seeking to exploit neural and behavioral data simultaneously. The results of the simulation studies and the fMRI experiment demonstrate that ADO can successfully incorporate both neural and behavioral data to maximize the acquisition of neurophysiological measures to explain behavioral responses. We have shown that these results are generalizable across between- and within-participant scanning sessions.

The impact of a few simplifications on the results deserve mention. One limitation of our method is the manner in which trial-wise brain activation is acquired. In our fMRI experiment, we simply estimated the single-trial activation parameters on each trial, and used them directly as input to the joint model. However, when using ADO in fMRI experiments, the unbalanced and interdependent nature of experimental designs generated by ADO can inflate variability of single-trial neural estimates. Because ADO is “greedy" in the way it maximizes global utility on the next trial, it can sometimes tend to over-select a particular stimulus pair. Because the stimulus pair is selected more frequently, extreme single-trial neural estimates become more likely, resulting in amplified variability.

In addition, task-irrelevant factors such as neural adaptation can potentially interact with unbalanced designs and affect the mean trend and variability of neural activation estimates. Although we found no conclusive evidence of neural adaptation in our experiment, we cannot rule out this possibility for future applications and list it as a way to potentially improve the algorithm.

Another potential shortcoming of the results presented here is our treatment of neural variability. It has been observed that the variability of neuronal firing rates increases according to the mean firing rate (Boynton, Demb, Glover, & Heeger, 1999), implying that the variability of the BOLD responses is a function of their amplitudes across time because neural firing rates are positively correlated with BOLD amplitudes (Heeger, Huk, Geisler, & Albrecht, 2000). To keep the model simple, we assumed that the variance in the BOLD responses were constant throughout the scanning session. However, if our assumption is violated, it is possible that our single-trial estimates would become inaccurate, thereby affecting the efficiency of the ADO procedure.

Lastly, the interstimulus interval (12 seconds on average) used in this study might not be desirable from the perspective of efficiency. Also, the performance of ADO might be partially due to better relaxation of BOLD responses with extended interstimulus interval. However, the interstimulus interval can easily be shortened by using high-performance computing resources and parallel computing to offload many of the ADO procedures. As the experiment we report in this article was more of a proof of concept, we didn't pursue these options here. Future work will incorporate more efficient computing so that more difficult optimization problems can be pursued.

Despite these limitations, we have shown that Adaptive Design Optimization can be applied to real-time fMRI experiments to successfully optimize the selection of stimuli for each individual. Our method has important improvements compared to previous design optimization methods in neuroimaging. Unlike many previous methods (Cusack et al., 2012; Holling et al., 2013; Lorenz et al., 2016), the model-based nature of ADO allows us to explore candidate designs that inform our understanding of the computations assumed to

underlie mental operations, pursuing more than localized activation of the brain. Moreover, our method not only incorporated both neural and behavioral data successfully for optimization, but does so in a formal and systematic way thanks to a joint model framework which provides common statistical constraints. Lastly, unlike adaptive procedures used in psychophysics (e.g. Leek, 2001) such as staircase procedures, ADO is a general-purpose design optimization algorithm, enabling it to be applied to any combination of neurocomputational and cognitive models, or data modality (e.g., EEG, fMRI, single-unit recording).

One may view the randomized search as an experimental design as a relatively low reference point by which to compare our ADO-based search. However, the randomized search is still the predominant design in cognitive neuroscience experiments. Previously developed online design optimization methods focused on slightly different optimization problems, making them inappropriate to compare against here. For example, many alternative optimization methods either ignore the neural data when performing optimization (e.g., DiMattina, 2016; Kontsevich & Tyler, 1999; Watson & Pelli, 1983), or they are not cognitive-model driven (e.g., Cusack et al., 2012; Lorenz et al., 2016). With our pipeline for fMRI-based ADO established, future work will systematically study the effect of different neural-behavioral modeling strategies and optimization methods.

Do Optimal Designs Guide Cognition Differently?

One general concern about using design optimization methods is that the proposed optimal designs could alter cognitive processes from what we would expect when using randomized or factorial experimental designs. Note that this problem is not unique to our proposed ADO method, in principle, because traditional design optimization methods for behavioral and fMRI experiments (e.g., de Hollander et al., 2017; Kontsevich & Tyler, 1999; Leek, 2001; Watson & Pelli, 1983) suffer from the same issue. However, traditional factorial experimental designs are also not immune to introducing possible distortions of cognitive biases.

A common assumption is that the use of design optimization techniques would be justified only when cognitive (and underlying neural) processes associated with the given task are equivalent, regardless of whether or not an optimization method is used. However, the relationship between ADO and the target cognitive process depends heavily on the nature of the task. In particular, the visual judgment task used in this study might not rely critically on higher-level processes such as (changes in) cognitive strategies. Also, ADO can implement strategies to avoid changes in the cognitive processes that are obviously problematic. For example, ADO can inadvertently create a more difficult or fatiguing task simply by proposing difficult trials consecutively. However, this problem can be avoided by intermittently inserting easy trials, although the standard for applying this correction and its effect must be tested formally. At this point, further investigations are required to make more conclusive statements about possible interactions among design optimization methods, experimental tasks, cognitive models, and participants' cognitive processes.

As for our ADO application to the fMRI data, we understand that the simultaneous use of neural and behavioral data in ADO makes this problem particularly non-trivial, as neural adaptation is an especially difficult hurdle. However, the application of the general-purpose design optimization methods in cognitive science is still in its infancy. Researchers should be aware of the possibility that the use of ADO could alter the underlying neural and cognitive processes from their standard, factorial design counterparts. At the same time, we should also remember that no design principle is problem-free, and the relationship between ADO and experimental tasks must be investigated further.

Quality Control of the Neural Data

Like other typical fMRI experiments, fMRI-based ADO needs neural data of high quality for obtaining clear results. Moreover, offline data preprocessing cannot be an option for fMRI-based ADO due to its nature as a real-time data collection method. Therefore, real-time quality control is one of the crucial factors in successful ADO experiments. Although we have applied only the minimum level of preprocessing methods (e.g., motion correction,

masking), one could take advantage of real-time filtering methods or even more integrative real-time fMRI frameworks such as OpenNFT (Koush et al., 2017).

As for head movement, real-time motion correction algorithms applied by the MR scanner might not be a perfect solution to the problem. Recent development of real-time monitoring software such as FIRMM (Framewise Integrated Real-time MRI Monitoring; Dosenbach et al., 2017) can help the experimenter detect any head motion anomalies, allowing them to correct the issue through participant instruction.

Multi-voxel Extension

In this study, we used a simple model connecting the average neural activation of V1 to behavioral decision processes. The use of activation amplitudes based on averaged signals came from a practical decision as the goal of this study is to provide a proof of concept of the fMRI-based ADO. However, many fMRI experiments focused on how distributed neural activations represent stimuli or underlying cognitive processes (for reviews, see Kriegeskorte & Diedrichsen, 2019; Norman, Polyn, Detre, & Haxby, 2006). The current fMRI-based ADO, in principle, can incorporate distributed neural representations with the same computational principle. However, most of the joint modeling approaches have connected cognitive model parameters with the average neural activation amplitude (e.g., Palestro et al., 2018; B. M. Turner, Forstmann, et al., 2013; B. M. Turner et al., 2016, 2015), in which distributed representation does not blend well. Therefore, the application of fMRI-based ADO must be accompanied by the development of joint models that are compatible with multi-voxel representations.

Practical Applications

Application of ADO to real-time neuroimaging experiments has great potential for both basic research and practical applications. Real-time comparison of computational cognitive models seems especially promising as neural data can sometimes provide discriminating evidence that could not be obtained on the basis of behavioral data alone

(e.g., Mack, Preston, & Love, 2013). Other domains where the need for adaptive and rapid assessment of brain-behavior relations occur is in cognitive psychometrics (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011) and computational psychiatry (Wiecki, Poland, & Frank, 2015). In these fields, obtaining high-quality data custom-tailored to each individual is of vital importance if we are to have confidence in our ability to assess and diagnose patients. With the groundwork of an adaptive, real-time methodology established, future refinements could automatically identify key brain regions for each task, allowing researchers to adjust scanning protocols to maximize the signal-to-noise ratio for each participant. We hope that the algorithm developed here will enable the field to look beyond problematic aggregation procedures and focus on custom-tailored experiments that optimize for our understanding of how the brain produces behavior.

Conclusion

In this study, we demonstrated the computational framework for optimizing experimental designs of cognitive-model-based fMRI experiments in real time. Using the joint modeling framework, fMRI-based ADO successfully incorporated neural and behavioral data simultaneously for proposing the sequence of experimental stimuli with the highest global utility. Simulation and actual fMRI experiments showed that fMRI-based ADO outperforms randomly proposed stimuli in accuracy and precision of parameter estimates. Given its model-based nature, fMRI-based ADO can help researchers investigate computational mechanisms of the human brain and mind with optimized experiments. Moreover, this method can assist experiments with special groups of interest (e.g., children, clinical populations) more efficiently.

Appendix A

Details of the fMRI Experiment

Functional Localizer

Before running the main task, we ran a functional localizer task to detect the voxels rigorously coactivating with the grating stimuli. The functional localizer task was based on a continuous carry-over design (Aguirre, 2007) that controls the order effect of the signal by considering all possible carry-over patterns from a stimulus pool. As we can expect that the order of stimuli affect the neural activation pattern, the continuous carry-over design can be used to detect voxels that share similar activation patterns and the carry-over effect.

The experiment using the continuous carry-over design uses a fixed stimulus presentation order that realizes all possible configurations of carry-over patterns. Here, we recommend making stimulus presentation settings as similar as possible to those of the main task. For example, we set the stimulus duration (6 seconds) and the mean interstimulus interval (8 seconds) as it was in the main task. However, generating all possible carry-over patterns from ten contrast levels made the task length excessive and therefore could have caused problematic issues such as participant fatigue and scanner drift. Hence, we decided to use only five logarithmically spaced contrast levels that could approximate contrast levels used in the main task (i.e., 0.01, 0.03, 0.1, 0.3, 1). The total length of the functional localizer task was 528 seconds.

In the task, the participant was instructed to press a button when the current stimulus was of the same contrast with the previous one while maintaining fixation at the center of the screen. However, the behavioral task served no function; it was required only to help participants concentrate on the stimulus presentation.

In-session Procedures 1: Preliminary Tasks

The participant went through a 30-minute briefing including informed consent, safety screening, and a brief introduction about the experimental task. MRI scanning was performed in the Center for Cognitive and Behavioral Brain Imaging at The Ohio State Univer-

sity. A Siemens MAGNETOM Prisma 3T Magnetic Resonance Imaging System was used with a 32-channel head coil.

First, the MPRAGE sequence was used for obtaining the anatomical structure of the brain ($1\times\ 1\ \times\ 1\ \ mm^3$ resolution, inversion time = 950 msec, repetition time = 1900 msec, echo time = 4.44 msec, flip angle = 12 degree, matrix size = 256 $\times$ 224 mm, 176 sagittal slices per slab; scan time = 6.5 minutes). As we hoped to constrain the ROI to the primary visual cortex (V1), the area to be scanned was then specified by covering the Brodmann area 17 and most of the occipital lobe with a T2*-weighted EPI sequence (repetition time = 2000 msec, echo time = 28 msec, flip angle = 72 degree, field of view = 200 $\times$ 200 mm, in-plane resolution = $2\times$ 2 mm, and 33 slices with 2-mm thickness), which is referred to as the EPI space henceforth for simplicity. All BOLD responses from the functional localizer task and the contrast discrimination task were obtained using the EPI sequence with the same setting.

We should mention that further analyses (i.e., detecting voxels of interest, real-time computation for Adaptive Design Optimization, offline data analysis) used brain images without preprocessing steps that are usually performed in offline analyses such as spatial and temporal filtering due to its time consumption. The only exception is motion correction: the MR scanner used in this experiment offers functionality for prospective motion correction – computational methods for reducing head motion artifacts during data acquisition (for a recent review of prospective motion correction, see Maclaren, Herbst, Speck, & Zaitsev, 2013).

In-session Procedures 2: Data preprocessing

We first carried out the functional localizer task to detect the voxels co-activating with the presented grating stimuli. After the functional localizer task was complete, we registered the anatomical images in the subject space to the standard MNI brain template with nonlinear warping using FLIRT and FNIRT (Andersson, Jenkinson, & Smith, 2007; Jenkinson, Bannister, Brady, & Smith, 2002; Jenkinson & Smith, 2001) in FSL (Smith et al.,

2004) . Next, we aligned the EPI localizer images to the anatomical images using FLIRT. By using the linear and nonlinear warping obtained from the previous steps, we converted the mask for Brodmann area 17 provided by Jülich histological atlas (Amunts, Malikovic, Mohlberg, Schormann, & Zilles, 2000; Eickhoff et al., 2005) to the EPI space. As these procedures usually take more than 7 minutes due to nonlinear registration, we asked the participant to practice the contrast discrimination task for (approximately) 6 minutes to learn the response-button mapping rule.

In-session Procedures 3: Determination of Voxels of Interest

The functional localizer task must detect voxels whose activation patterns are strongly associated with stimulus presentation in the task. For selecting target voxels in the main task, we performed a general linear model (GLM) analysis to all voxels in the EPI space using the data from the functional localizer task. The GLM design matrix used only one regressor representing the hemodynamic responses caused by all stimuli presented in the functional localizer task. This GLM analysis did not consider any temporally autocorrelated noise in the model structure because the analysis may be time-consuming.

Voxels in interest (VOIs) were determined by thresholding the t-statistic associated with the regression coefficient of the task-relevant regressor. The decision rule is as follows: If the number of voxels with $t \geq 5$ was equal to or greater than 200, we used the threshold as $t = 5$. However, when this criterion was not met, we adjusted the threshold to $t \geq 4$. If 100 or more voxels passed the adjusted threshold, we accepted the threshold $t = 4$. If this criterion was not met again, we ran the functional localizer task one more time and repeated the analysis. If the result did not allow 100 or more voxels even in the second attempt, we used the threshold allowing the greatest number of voxels among four options (i.e., $t \geq 5$ from the first run, $t \geq 4$ from the first run, $t \geq 5$ from the second run, and $t \geq 4$ from the second run).

Finally, we derived the subject-specific, task-relevant mask specifying VOIs in V1 by taking conjunction of the subject-specific V1 mask and the extracted task-relevant voxels.

Table A1 shows the number of voxels actually used in the mask. A Python library `nilearn` (Abraham et al., 2014) was used for formatting the final mask.

	Session 1	Session 2	Session 3
Participant 1	189	125	117
Participant 2	171	330	315
Participant 3	136	82	25
Participant 4	47	142	227

Table A1
The number of voxels used in the experiment and post-hoc analysis.

In-session Procedures: Contrast Discrimination Task

The contrast discrimination task was carried out after the processing of the mask was finished. Two runs were done separately based on Adaptive Design Optimization (ADO) and Randomized Search (RS) within a scanning session so that we could consider between-session variability of the neural signal.

In the ADO-based run, the first four trials are randomly proposed because of the hemodynamic lag that prevents immediate estimation of stimulus-wise neural activation estimates. From the fourth trial, ADO computed the global utility of candidate designs and proposed an optimal stimulus pair by the following procedure. First, we extracted the BOLD time series from the VOIs and averaged them. Then we estimated single-trial neural activation for each grating stimulus by fitting a general linear model (GLM) with the first-order temporal autocorrelation (AR(1)) model for the noise in the data using a Python library `statsmodel` (Seabold & Perktold, 2010). Here, the AR(1) model assumes that the measurement noise at time t is correlated with measurement noise at time $t-1$. Once we obtained the stimulus-wise estimates of neural activation, they were put into ADO together with behavioral responses for computing the optimal design of the next trial. After the 4th, 8th, 12th, and 16th trials, we sampled the joint posterior distribution using the DE-MCMC sampler (B. M. Turner, Sederberg, et al., 2013) for 500 iterations, and used the last 300 samples for dynamic gridding.

The total length of both ADO-based and RS-based experiments is 20 trials. In other

words, ADO used a simple stopping rule based on a fixed number of trials (20 trials), as we need to control the amount of data for parameter estimation.

Preliminary Analysis of the Neural and Behavioral Data

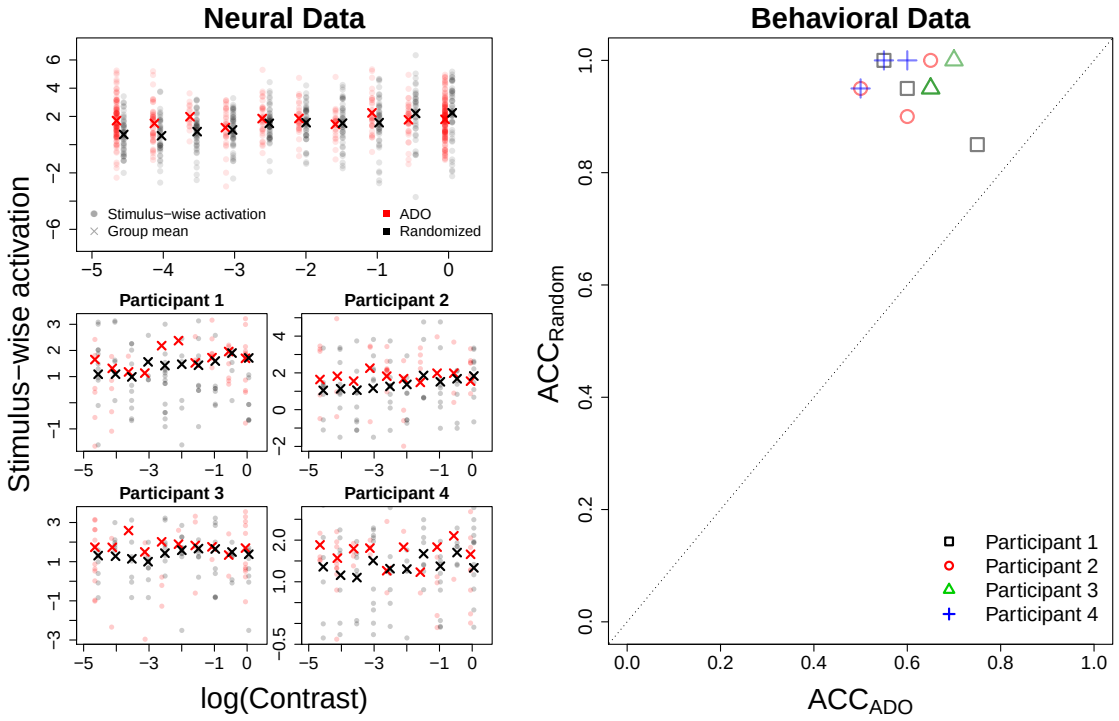

Figure A1. **Summary of the Neural and Behavioral Data**. The left panels show scatter plots of log-transformed contrast levels versus stimulus-wise neural activation levels in Adaptive Design Optimization (red) and Randomized Search (black) experiments. The upper left panel presents the stimulus-wise activation estimates aggregated across participants, while the four lower left panels illustrate the same data but separately for each participant. The right panel shows a scatter plot of accuracy of behavioral responses of four participants. The x-axis represent behavioral accuracy in Adaptive Design Optimization experiments, whereas the y-axis represent behavioral accuracy in Randomized Search experiments.

Figure A1 summarizes the neural (i.e., single-trial neural estimates; left panel) and the behavioral data (accuracy of the behavioral responses; right panel). The upper left panel shows distributions of stimulus-wise neural activation estimates for each contrast, collapsed across participants. The four lower left panels present the same data for each participant. Theoretically, the single-trial neural estimates are expected to monotonically

escalates according to the increase of the contrast level (Boynton et al., 1999). However, single-trial neural estimates are broadly distributed due to their high variability (Abdulrahman & Henson, 2016; Mumford et al., 2012) and unbalanced designs. When their group means were compared, Randomized Search (RS; black) experiments tend to allow a monotonically increasing pattern, whereas the expected pattern is not clearly observed in Adaptive Design Optimization (ADO; red) experiments.

The right panel of Figure A1 shows the accuracy of behavioral responses in ADO (x-axis) and RS (y-axis) experiments. If a dot is located below the identity line (dotted line), we consider that the performance in ADO experiments is better than in RS experiments. The result consistently shows that participants made more accurate responses in RS experiments than in ADO experiments. This tendency is partially explained by that ADO in this experiment frequently focuses on small contrast values to obtain information about the baseline parameter of Naka-Rushton Equation (See Figure 5 in the main text for an example of the proposal trace in Simulation Study).

Posterior Sampling

For offline analyses to compare the performance of ADO to RS, we estimated parameters with a complete data set. We first estimated stimulus-wise neural activation levels from ADO and RS experiments. After averaging the extracted BOLD time-series from all voxels in the mask, we fitted a general linear model with the first-order temporal autocorrelation in noise to estimate stimulus-wise neural activation parameters. Once the single-trial neural estimates were acquired, the joint model parameters were finally estimated by the DE-MCMC sampler with the stimulus-wise neural activation and behavioral responses as the data.

Compared to the simulation study, we had to modify the DE-MCMC sampler settings due to the quality of neural data associated with the mechanism of ADO. ADO tends to generate the same design repeatedly until it gets enough information about the specific parameter, and then proposes distinct patterns of the design to explore different model

parameters. As mentioned in Discussion, we found that the unbalanced design of ADO adds a significant amount of variability of stimulus-wise neural activation estimates and may induce difficulties in getting well-constrained posterior distributions.

Therefore, we decided to use a “burn-in mode" of the DE-MCMC sampler that concentrates posterior samples to the high-density regions compared to the regular “sampling mode" (B. M. Turner & Sederberg, 2012), in addition to high migration probability. Specifically, the DE-MCMC sampler was run with the “burn-in mode" for 3,000 iterations in total: the sampler used the first 2,000 iterations as a burn-in phase while applying migration at every iteration, and generated the valid posterior samples for the last 1,000 iterations.

Note that brain images from the ADO-based and randomized-design runs shared the same data preprocessing procedures to make the stimulus-wise activation estimates from both experiments comparable. We used the motion-corrected images exported directly from the MR scanner, and did not apply spatial and temporal filtering. The neural signal was extracted from the same VOI mask defined for ADO.

Appendix B

Performance of ADO: Participant 4

In the case of Participant 4, which is not reported in the main text, ADO failed to show better performance in two out of three scanning sessions. Figures B1, B2, and B3 provide summary plots of the performance of ADO and RS in the data set of Participant 4. In Figure B1, the design proposals made by ADO seem more distributed compared to other participants described in Figure 9. It is not easy to say any decisive conclusion only with this plot because of factors that affect the actual fMRI experiment (e.g., session-by-session variability, head motion). However, the lack of specificity toward the combinations of extremely low and high contrasts, which are useful for estimating b and R_{max}, suggests that the performance of ADO was suboptimal. Figure B2 shows the accuracy with respect to the benchmark estimate and the precision of the parameter estimates. Unlike other participants' sessions where ADO performed better in accuracy, the results from Participant

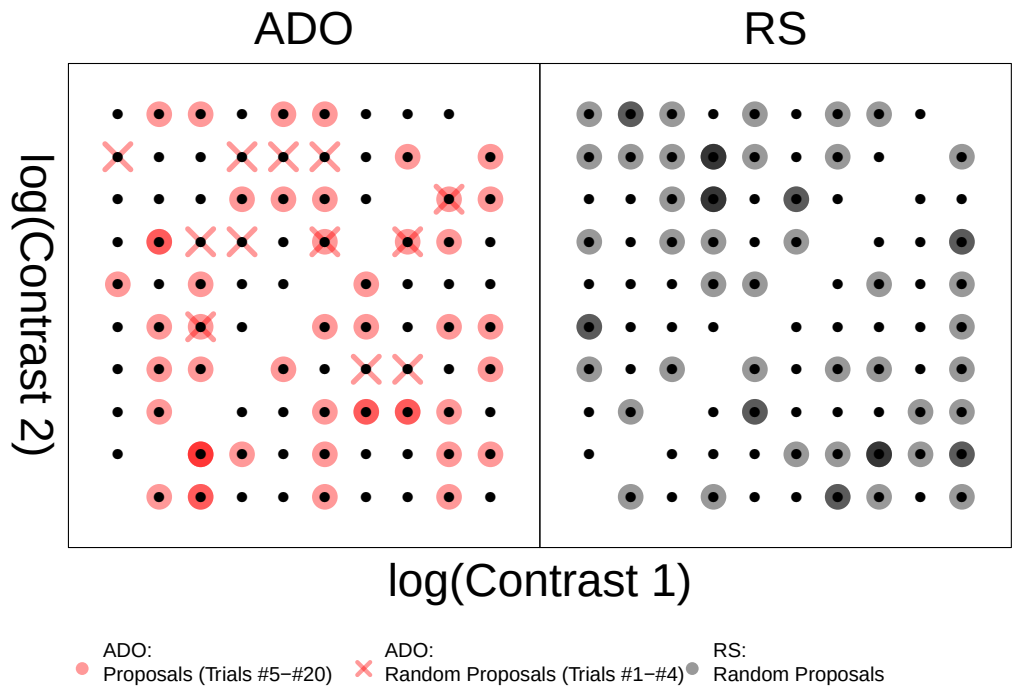

Figure B1. **Participant 4: Proposed Designs from the fMRI Experiment**. The figure shows a path analysis comparing Adaptive Design Optimization (ADO; left) against Randomized Search (RS; right). Results from all three replicate sessions are collapsed for each participant. Frequency of stimulus selection is indicated by intensity of the circles, where the first and second stimuli are shown on the $x-$ and $y-$axes, respectively. The first four random trials in the ADO-runs are plotted with "$\times$" marks. The labels for two axes were intentionally omitted for visual clarity.

4 are mixed. Although the results from the second (green) and third (blue) sessions claims that the estimates were more precise, the third session (blue) loses this advantage due to inaccurate estimates. In Figure B3, the bar plot on the left side shows the performance comparison metrics acquired across three scanning sessions. The value of the performance metric at the first and third sessions are negative, which means that estimated Naka-Rushton curves in RS runs showed a better fit to the benchmark curve than in ADO runs. The latter two plots show the distribution of normalized global utility recovered by post-hoc analyses to test whether ADO appropriately presented the optimal sequence of stimuli. The result of the first session (red scatter and violin plots) reveals that the ADO might not have been successful in proposing optimal stimuli because the normalized global utility

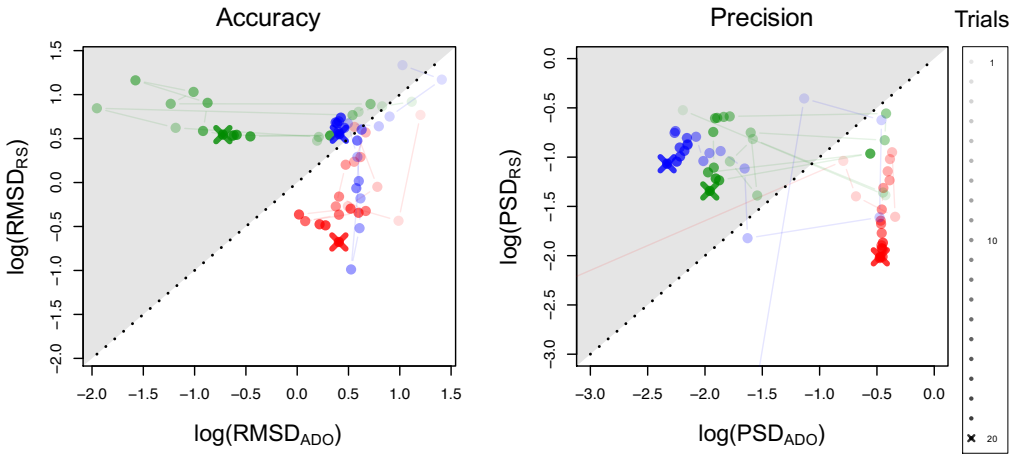

Figure B2. **Participant 4: Accuracy and Precision of Parameter Estimates**. Performance of the two algorithms is compared in terms of the accuracy with respect to the benchmark estimate (left) and the posterior precision (right). Colored lines with circles and "×" marks represent the accuracy and precision changing over trials. Each row shows the results from diffrent participants (top: Participant 1, middle: Participant 2, bottom: Participant 3). Replicate sessions are color-coded (red: Session 1, green: Session 2, blue: Session 3). Empty dots represent the first four trials at which ADO had to use random proposals. The dot with "×" mark refers to the last trial of each session. The black dotted line represents the identity line. If a point locates in the gray area (i.e., above the identity line), it means that ADO shows higher accuracy or precision compared to RS at that trial. The ranges of both axes were truncated for visual clarity.

distributions do not show differences between the two methods.

To investigate why ADO performed worse in these sessions, Figure B4 provides summary statistics of the neural data. Figure B4a plots the performance comparison metric (i.e., $\log(DEVD_{RS}/DEVD_{ADO})$) against head movement measures (left) and the number of voxels (right). In both panels, white areas designate regions of the statistical space where ADO performs worse than RS. In the left plot, the log-transformed ratio of mean absolute displacement between RS and ADO is shown on the x-axis, where greater values are preferred. Both plots reveal that the performance of ADO tends to be better under conditions in which there is less head movement (left) and the size of the region of interest consists of a greater number of voxels (right). Figure B4b shows the region of interest extracted from our functional localizer task, color coordinated by session for Subject 4. Here, the figure shows that the mask identified in Session 1 (red) deviated considerably in

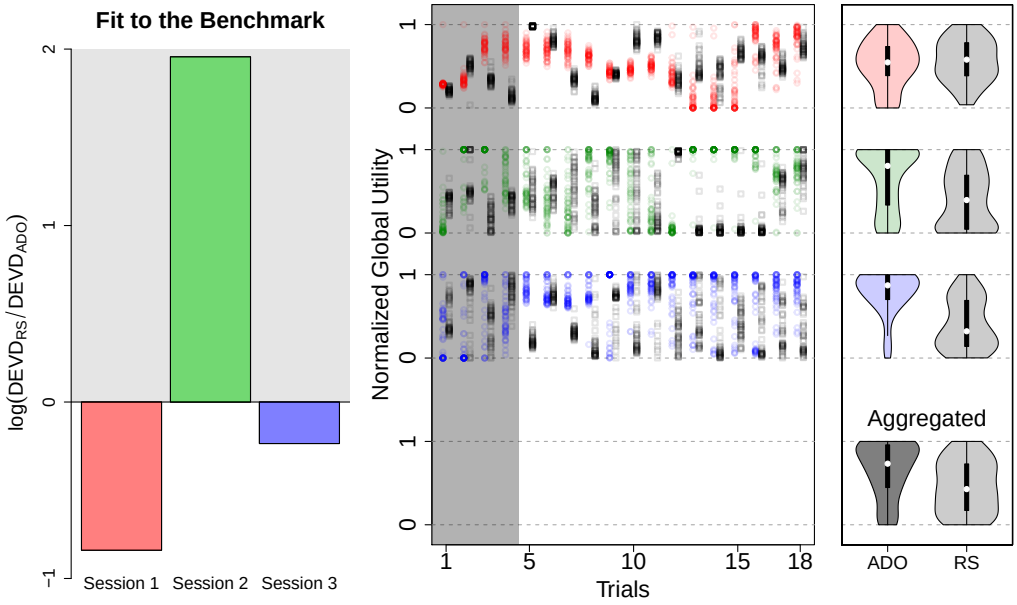

Figure B3. **Participant 4: Prediction Analyses and Global Utility Distributions**. The left plot shows the log ratio of averaged fit measures of RS to ADO compared to the benchmark, as illustrated in Figure 11 in the main text. Higher values support ADO in accuracy (gray area). The latter two plots show the distribution of normalized global utility of all possible stimulus pairs generated by ADO (colored plots) and RS (gray plots). A scatter plot in the center shows how the distribution of normalized global utility changes over trials, whereas a violin plot on the right panel represents the same information aggregated across trials. For all plots, note that the first four trials were excluded from the violin plot because RS was used for both search procedures.

both size and location from Sessions 2 and 3. Finally, B4c shows the displacement from all three sessions of ADO (colored lines) and RS (black lines) as a function of time. ADO Session 3 in particular showed considerably more movement relative to the corresponding RS run. Hence, these analyses reveal that ADO performs worse than RS only when the quality of the neural data are poor, which we encountered in the first and third sessions for Subject 4.

In summary, our post-hoc analyses revealed why ADO performed worse than RS in the two scanning sessions of Subject 4. Specifically, the mask defined in Session 1 following our functional localizer consisted of a small number of voxels that were not representative of the key visual areas. In Session 3, we observed much larger head movements in the

ADO condition relative to the RS condition.

References

Abdulrahman, H., & Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, *125*, 756–766.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. Retrieved from https://www.frontiersin.org/article/10.3389/fninf.2014.00014 doi: 10.3389/fninf.2014.00014

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *NeuroImage*, *35*(4), 1480–1494.

Amunts, K., Malikovic, A., Mohlberg, H., Schormann, T., & Zilles, K. (2000). Brodmann's areas 17 and 18 brought into stereotaxic space - where and how variable? *NeuroImage*, *11*(1), 66–84.

Andersson, J. L. R., Jenkinson, M., & Smith, S. (2007). Non-linear registration aka Spatial normalisation (FMRIB Technical Report TR07JA2). Retrieved from https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf

Boynton, G. M., Demb, J. B., Glover, G. H., & Heeger, D. J. (1999). Neuronal basis of contrast discrimination. *Vision research*, *39*(2), 257–269.

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, *52*, 233–254.

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*, 887–905.

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, *47*, 255–289.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin and Review*, *18*, 204–210.

Cusack, R., Veldsman, M., Naci, L., Mitchell, D. J., & Linke, A. C. (2012). Seeing different objects in different ways: measuring ventral visual tuning to sensory and semantic features with dynamically adaptive imaging. *Human brain mapping*, *33*(2), 387–397.

de Hollander, G., Keuken, M. C., van der Zwaag, W., Forstmann, B. U., & Trampel, R. (2017). Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7T and 3T. *Human Brain Mapping*, *38*(6), 3226–3248.

DiMattina, C. (2016). Comparing models of contrast gain using psychophysical experiments. *Journal of Vision*, *16*, 1–18.

DiMattina, C., & Zhang, K. (2013). Adaptive stimulus optimization for sensory systems neuroscience. *Frontiers in neural circuits*, *7*, 101.

Dosenbach, N. U., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., ... others (2017). Real-time motion analytics during brain MRI improve data quality and reduce costs. *NeuroImage*, *161*, 80–93.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, *21*(7), 1–16.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*, 1325–1335.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*(36), 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., & Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 58–66).

Greve, D. N., Brown, G. G., Mueller, B. A., Glover, G., Liu, T. T., et al. (2013). A survey of the sources of noise in fMRI. *Psychometrika*, *78*(3), 396–416.

Heeger, D. J., Huk, A. C., Geisler, W. S., & Albrecht, D. G. (2000). Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? *Nature Neuroscience*, *3*(7), 631.

Holling, H., Maus, B., & van Breukelen, G. J. P. (2013). Optimal design for functional magnetic resonance imaging experiments. *Zeitschrift für Psychologie*, *221*, 174–189.

Hu, B., & Tsui, K.-W. (2005). Distributed evolutionary Monte Carlo with applications to Bayesian

analysis (Technical Report Number 1112). Retrieved from http://www.stat.wisc.edu/techreports/tr1112.pdf

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Johnson, R. A., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

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

Koush, Y., Ashburner, J., Prilepin, E., Sladky, R., Zeidman, P., Bibikov, S., . . . Van De Ville, D. (2017). OpenNFT: An open-source Python/Matlab framework for real-time fMRI neurofeedback training based on activity, connectivity and multivariate pattern analysis. *NeuroImage*, *156*, 489–503.

Krekelberg, B., Boynton, G. M., & van Wezel, R. J. (2006). Adaptation: from single cells to BOLD signals. *Trends in Neurosciences*, *29*(5), 250–256.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*, 407–432.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.

Li, X., Lu, Z.-L., Tjan, B. S., Dosher, B. A., & Chu, W. (2008). Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. *Proceedings of the National Academy of Sciences of the United States*, *105*, 6202–6207. Retrieved from https://doi.org/10.1073/pnas.0801390105

Lindquist, M. A., Loh, J. M., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *NeuroImage*, *45*(1), S187–S198.

Lorenz, R., Monti, R. P., Violante, I. R., Anagnostopoulos, C., Faisal, A. A., Montana, G., & Leech, R. (2016). The Automatic Neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI. *NeuroImage*, *129*, 320–334. Retrieved from https://doi.org/10.1016/j.neuroimage.2016.01.032

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization

from its neural implementation. *Current Biology*, *23*(20), 2023–2027.

Maclaren, J., Herbst, M., Speck, O., & Zaitsev, M. (2013). Prospective motion correction in brain imaging: A review. *Magnetic Resonance in Medicine*, *69*(3), 621–636.

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., ... Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*, *14*(8), 1200–1214.

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*, 2636–2643.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*, 53–67.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, *10*, 424–430.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. New York: New York: Cambridge University Press.

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, *23*, 752–763.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*, 128–154. Retrieved from `https://doi.org/10.1111/insr.12107`

Sanchez, G., Daunizeau, J., Maby, E., Bertrand, O., Bompas, A., & Mattout, J. (2014). Toward a new application of real-time electrophysiology: online optimization of cognitive neurosciences hypothesis testing. *Brain Sciences*, *4*(1), 49–72. Retrieved from `https://doi.org/10.3390/brainsci4010049`

Sanchez, G., Lecaignard, F., Otman, A., Maby, E., & Mattout, J. (2016). Active SAmpling Protocol (ASAP) to optimize individual neurocognitive hypothesis testing: A BCI-inspired dynamic experimental design. *Frontiers in Human Neuroscience*, *10*, 347. Retrieved from `https://www`

.frontiersin.org/article/10.3389/fnhum.2016.00347 doi: 10.3389/fnhum.2016.00347

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th python in science conference* (Vol. 57, p. 61).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219.

Smucker, B., Krzywinski, N., & Altman, N. (2018). Optimal experimental design. *Nature Methods*, *15*(8), 559-560.

ter Braak, C. J. F. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*, 239–249.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 278–286. Retrieved from https://doi.org/10.1037/h0070288

Turner, B. M. (2015). Constraining cognitive abstractions through Bayesian modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 199–220). New York: Springer.

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint models of neural and behavioral data*. New York, NY: Springer.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Turner, B. M., Rodriguez, C. A., Liu, Q., Molloy, M. F., Hoogendijk, M., & McClure, S. M. (2018). On the neural and mechanistic bases of self-control. *Cerebral Cortex*, *29*(2), 732–750.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: A method for simultaneously modeling EEG, fMRI, and Behavior. *NeuroImage*, *128*, 96–115.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375–385.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*, 368–384.

Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Combining Cognitive Abstractions with Neurophysiology: The Neural Drift Diffusion Model. *Psychological Review*, *122*, 312–336.

1248 Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. G. (2012). Spatiotemporal activity
1249 estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, *62*,
1250 1429–1438.
1251 van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive
1252 psychology meets psychometric theory: On the relation between process models for decision
1253 making and latent variable models for individual differences. *Psychological Review*, *118*(2),
1254 339.
1255 Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception*
1256 *& Psychophysics*, *33*(2), 113–120.
1257 Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-based cognitive neuroscience approaches
1258 to computational psychiatry: clustering and classification. *Clinical Psychological Science*, *3*(3),
1259 378–399.

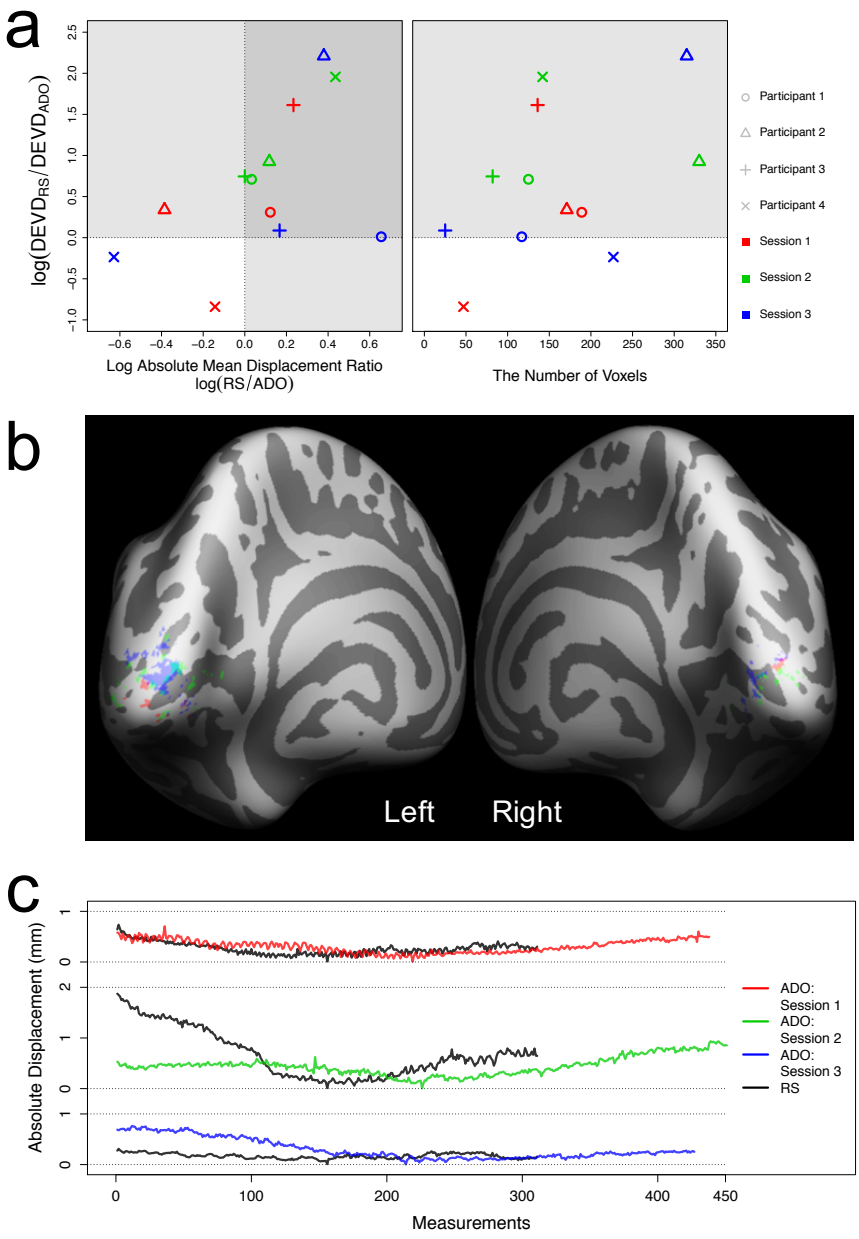

Figure B4. **Quality of the Neural Data and Performance of ADO**. Plots in panel a show scatter plots for comparing the performance comparison metric (i.e., $\log(DEVD_{RS}/DEVD_{ADO})$ and quality assurance metrics (left: log absolute mean displacement ratio of RS to ADO, right: the number of voxels used in the offline data analyses). Here, the performance comparison metric is the same as what used in the right plot of Figure 11 in the main text. The value greater than zero supports ADO in accuracy. The x-axis of the left plot is the log-transformed ratio of absolute mean displacement between RS and ADO, where absolute mean displacement is a summary metric revealing the degree of displacement from a single reference brain volume. Higher values of this ratio mean that head position was more stable in ADO than RS, and therefore are more preferred. In panels b and c, we describe the mask used for offline analyses (b) and time-series of absolute displacement (c) of Participant 4, who showed bad performance in ADO runs. In panel b, red, green, blue dots represent the mask used in the first, second, and third scanning session, respectively. In panel c, we used the same color-coding rule to represent absolute displacement in ADO runs, while black lines represent absolute displacement metric in RS runs.