ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 21 (2004) 401–413

Efficiency, power, and entropy in event-related fMRI with multiple trial types
Part II: design of experiments

Thomas T. Liu*

Center for Functional Magnetic Resonance Imaging, University of California San Diego, La Jolla, CA, USA

Received 27 June 2003; revised 29 August 2003; accepted 10 September 2003

The performance of an experimental design for functional magnetic resonance imaging (fMRI) can be characterized by its estimation efficiency, which is the ability to make an estimate of the hemodynamic response, its detection power, which is the ability to detect an activation, and its conditional entropy, which is a measure of the randomness of the design. In Liu and Frank [Neuroimage 21 (2004) 387–400], it is shown that there is a fundamental theoretical trade-off between estimation efficiency and detection power for experiments with multiple trial types and that there is an empirical relation between estimation efficiency and conditional entropy. This paper provides an intuitive interpretation of the theoretical results and examines the practical implications of these results for the optimal design of fMRI experiments with multiple trial types. The properties of block designs, permuted block designs, m-sequence designs, clustered m-sequence designs, and mixed designs are explored. It is shown that these designs nearly achieve the theoretically predicted performance and can be used in practice to obtain advantageous trade-offs among efficiency, power, and entropy.

Keywords: Functional magnetic resonance; Efficiency; Entropy

Introduction

Event-related designs for functional magnetic resonance imaging (fMRI) in which different types of stimuli are presented in rapid succession have proven to be useful for a wide range of cognitive experiments, especially in cases where the psychological confounds, such as habituation and anticipation, associated with more traditional block designs interfere with the cognitive task under study. In a companion paper (Liu and Frank, 2004), we examined the relation between three metrics that are useful for characterizing the performance of event-related fMRI experiments with multiple trial types. These are estimation efficiency, which is a measure of the ability to estimate the hemodynamic response function (HRF) in the absence of a priori assumptions about its shape; detection power, which is a measure of the ability to detect a functional activation assuming a known HRF; and conditional entropy, which is a measure of the randomness of a design. It has been previously shown that there is a fundamental theoretical trade-off between estimation efficiency and detection power for experiments with a single trial type (Liu et al., 2001). The finding in Liu and Frank (2004) is that this trade-off also exists in experiments with multiple trial types, and in fact the form of the trade-off is identical to that seen for single trial type experiments. In addition, an empirical relation between estimation efficiency and conditional entropy was found, with entropy increasing with the logarithm of estimation efficiency. Numerical simulations using a variety of experimental designs were used in Liu and Frank (2004) to verify the form of the theoretically predicted trade-offs.

The purpose of this paper is twofold. First, we provide an intuitive presentation of the theoretical results from Liu and Frank (2004) so as to make them more accessible to the nontechnical reader. Second, we explore the design of experiments that can be used to achieve the predicted trade-off in practice. These include block and permuted block designs, m-sequence and clustered m-sequence designs, and mixed designs.

Block designs have been previously described in the literature as designs that can achieve maximal detection power at the price of low estimation efficiency (Friston et al., 1999; Birn et al., 2002). We examine the sensitivity of block designs to low frequency nuisance terms and show that a design consisting of two blocks of each trial type provides high detection power and good robustness against nuisance terms. Designs based on m-sequences were introduced by Buracas and Boynton (2002) and were shown to offer significantly higher estimation efficiency than designs obtained via random search, especially as the number of trial types increases. We review the properties of m-sequence designs and show that they nearly achieve the theoretical upper bound on estimation efficiency over a wide range of experiments. Permuted block designs were introduced in Buxton et al. (2000) as designs that can offer a wide range of intermediate trade-offs between estimation efficiency and detection power for experiments with a single trial type. Here we show that they are also useful for attaining theoretically predicted trade-offs for experiments with

---

* Center for Functional Magnetic Resonance Imaging, University of California San Diego, 9500 Gilman Drive, MC 0677, La Jolla, CA 92093-0677. Fax: +1-858-822-0605.
E-mail address: ttliu@ucsd.edu.

1053-8119/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.09.031

multiple trial types. Clustered m-sequence designs are introduced here as novel designs that can offer additional trade-offs. These designs take advantage of the optimal estimation efficiency exhibited by m-sequence designs. Finally, mixed designs can be formed by combining block designs and m-sequence designs. By varying the relative lengths of the block and m-sequence designs, variable trade-offs between efficiency and power can be obtained.

To provide the necessary background for evaluating the various experimental designs, we begin with an intuitive interpretation of the results presented in Liu and Frank (2004). The motivated reader is referred to that article for proofs and additional technical details. The review of the theoretical results is followed by a detailed examination of the various experimental designs.

Design theory

Experimental assumptions

In this paper, we consider fMRI experiments with Q trial types plus an optional null or control condition so that there are up to $Q + 1$ experimental conditions. The trial types are constrained to be nonoverlapping in time so that at each time point in the design, only one trial type may have a stimulus present. For example, in an experiment with $Q = 2$ trial types, A and B, plus a null condition N, the experimental designs are of the form ABBNABN. The stimulus for each trial type is binary, either on or off, so that the stimulus may be represented as a binary sequence where 1 denotes the presence of the trial type and 0 denotes its absence. We assume that the HRF associated with each trial type can be described by a vector with k points. As an example, if the HRF is assumed to last for 20 s and the experimental repetition time (TR) is 1 s, then $k = 20$. The temporal duration of the HRF can be longer than the spacing between stimuli from different trial types so that the responses to different trial types can overlap even though the stimuli do not. The length of the experiment is equal to the number of temporal data points acquired and is denoted as N. The measured signal $y[n]$ is assumed to be equal to the sum of (1) the time series obtained by convolving the HRF for each trial type $h_i[n]$ with its stimulus pattern $x_i[n]$, (2) nuisance terms $s_i[n]$ to model the low frequency drifts commonly observed in fMRI experiments, and (3) a Gaussian noise term $e[n]$, where n denotes the discrete time index. In mathematical terms, the signal is written as $y[n] = \sum_{i=1}^{Q} x_i[n]*h_i[n] + \sum_{i=1}^{l} b_i s_i[n] + e[n]$, where * denotes convolution and b_i is the coefficient for the ith nuisance term. The matrix equivalent of this expression is given in Eq. (1) of Liu and Frank (2004).

Statistical efficiency: detection power and estimation efficiency

The statistical efficiency of an experimental design is inversely proportional to the variance in the estimates of the parameters of interest. Increasing the efficiency of a design is equivalent to either decreasing the time required to obtain an estimate with a desired level of variance or decreasing the variance of an estimate given a fixed amount of measurement time. In fMRI experiments, the types of estimates that are of interest can be roughly divided into two main categories: estimates of the amplitudes of functional activation and estimates of the HRFs associated with functional activation.

Estimates of the activation amplitudes are important in experiments that attempt to detect activation to map areas of functional activity or to compare levels of activity between brain regions, trial types, or subjects. We refer to the efficiency of the amplitude estimates as detection power. This nomenclature is consistent with the formal definition of statistical power as the probability of detection when an activation is present and the fact that power decreases monotonically with increases in variance. For experiments with detection as a focus, the shape of the HRF is not typically of interest and a prototypical HRF with unknown amplitude is used in the statistical analysis. For each trial type, a statistical test is performed on the estimate of the unknown amplitude with the power of the test given by the inverse of the variance of the amplitude estimate. Similar tests may be performed on all pairwise contrasts. The overall detection power is then defined as

$$R_{\text{tot}} = \frac{K}{(\text{average variance of HRF amplitude estimates for all trial types and pairwise contrasts})} \tag{1}$$

where $K = (\bar{h}_0^T \bar{h}_0)^{-1}$ is a normalization constant and $\bar{h}_0$ is the shape of the HRF, which is assumed to be the same across trial types. As an example, the detection power for an experiment with two trial types, A and B, is given by $R_{\text{tot}} = K[(\text{var}(\hat{\mu}_{\text{A}}) + \text{var}(\hat{\mu}_{\text{B}}) + \text{var}(\hat{\mu}_{\text{A}} - \hat{\mu}_{\text{B}}))/3]^{-1}$, where $\hat{\mu}_{\text{A}}$ and $\hat{\mu}_{\text{B}}$ are the amplitude estimates.

Estimates of the HRFs are important for experiments that attempt to characterize the functional response to patterns of brief stimuli. For example, measurements of the responses to brief stimuli can provide insights into the physiology of the fMRI response (Miller et al., 2001). We refer to the efficiency of the HRF estimates as estimation efficiency and define it as

$$\xi_{\text{tot}} = \frac{1}{(\text{average variance of HRF estimates for all trial types and pairwise contrasts})} \tag{2}$$

For the two-trial-type example, the efficiency is $\xi_{\text{tot}} = [(\text{var}(\hat{h}_{\text{A}}) + \text{var}(\hat{h}_{\text{B}}) + \text{var}(\hat{h}_{\text{A}} - \hat{h}_{\text{B}}))/3]^{-1}$, where $\hat{h}_{\text{A}}$ and $\hat{h}_{\text{B}}$ are the estimates, $\text{var}(\hat{h}_{\text{A}})$ is defined as the sum of the variances of the estimates of the k unknown parameters of the HRF for trial type A, and $\text{var}(\hat{h}_{\text{B}})$ and $\text{var}(\hat{h}_{\text{A}} - \hat{h}_{\text{B}})$ are similarly defined.

In examining estimation efficiency, it is important to make a distinction between unbiased and biased estimates of the HRF. An unbiased HRF estimate makes no a priori assumptions about the shape of the HRF except for its temporal length and results in the definition of estimation efficiency used in Liu and Frank (2004) and Liu et al. (2001). Since no assumptions are made about the HRF, the expressions for estimation efficiency depend only on the experimental design. Biased estimates take advantage of the fact that although the exact shape of the HRF can vary greatly across subjects (Aguirre et al., 1998), the basic form of the HRF is a smooth function with a full width half maximum of about 5 to 6 s. One method for using this a priori knowledge is to constrain the HRF to lie in a subspace spanned by a set of smooth basis functions, such as a set of gamma density functions plus their temporal derivatives (Friston et al., 1998). The use of basis functions leads to a generalized definition of estimation efficiency with two important limiting cases. In the first case, if we assume full knowledge of the shape of the HRF, then there is only one basis function equal to $\bar{h}_0$ and the generalized estimation efficiency is equal to detection power, as shown in Appendix A5 of Liu and Frank, 2004. In the second case, if the number of basis functions is equal to the number k of unknown parameters in the HRF, then the generalized estimation efficiency is equal to the estimation efficiency for an unbiased estimate. Thus, detection power and estimation efficiency for an unbiased estimate are two measures of statistical efficiency that differ only in the amount of a priori information that is assumed about the HRF. For the purposes of this paper, we use the term estimation efficiency to refer to the efficiency for an unbiased estimate.

Conditional entropy

The perceived randomness of an experimental design is an additional factor to consider in many fMRI experiments, especially those involving complex cognitive tasks. Randomness can be critical for minimizing confounds, such as anticipation and habituation, which can arise when a subject can too easily predict the evolution of the stimulus pattern. For example, a decision-making task in which a subject is choosing between several outcomes could be hampered if the subject can readily predict the most advantageous outcome based on the history of previous trials. Conditional entropy is a fundamental and widely used metric for the randomness of a sequence (Cover and Thomas, 1991). As discussed in Liu and Frank (2004), it is related to other measures of randomness, such as predictability, mutual information, and counterbalancing, that have been previously presented in the fMRI literature (Bischoff-Grethe et al., 2001; Liu et al., 2001; Wager and Nichols, 2003). The rth-order conditional entropy H_r is a measure of the uncertainty in the next trial type given knowledge of the previous r trial types. A formal definition is provided in Eq. (28) of Liu and Frank (2004) and in Appendix A. Conditional entropy is measured in bits and is equal to the average number of binary (e.g., yes or no) questions required to determine the next trial type given the r previous trials (Cover and Thomas, 1991). A conditional entropy of 0 bits corresponds to a sequence in which the r previous trial types completely determine the next trial type. For an experiment with Q trial types, the maximum conditional entropy is equal to $\log_2(Q + 1)$ bits. This corresponds to a completely random sequence where each of the $Q + 1$ experimental conditions (i.e., Q trial types plus null condition) is equally probable given knowledge of the r previous trial types. As an example, the maximum conditional entropy for an experiment with three trial types is 2 bits, corresponding to a series of two questions of the form—Is the next trial type either trial type A or B? If it is either A or B, is it trial type A, and if it is neither A nor B, is it trial type C?

In comparing the entropy of different designs, it is convenient to use the quantity 2^{H_r}, which is a measure of the average number of random outcomes. For example, an entropy of 1 bit corresponds to $2^1 = 2$ equally probable outcomes, while an entropy of 2 bits corresponds to $2^2 = 4$ equally probable outcomes. Thus, a design with an entropy of 2 bits is twice as random as a design with an entropy of 1 bit, and the quantity 2^{H_r} serves as a linear measure of randomness.

The entropy of a design can vary significantly over its time course. For example, in the Design section, mixed designs are shown to exhibit areas of both high and low entropy. To assess the performance of such designs, it is useful to define a metric for *local conditional entropy*. This is a running measure of conditional entropy where the local entropy at time point n in the design is the entropy computed over a temporal window of width W centered about n.

Relation among estimation efficiency, detection power, and conditional entropy

It was shown in Liu et al. (2001) that there is a fundamental trade-off between estimation efficiency and detection power for experiments with one trial type of interest. Experimental designs, such as random designs, that optimize estimation efficiency yield low detection power, while designs, such as block designs, that achieve high detection power provide low estimation efficiency. Semirandom and mixed designs lie in between block and random designs and can offer a range of intermediate trade-offs between efficiency and power. For example, a semirandom design can offer a 100% increase in detection power with only a 20% decrease in estimation efficiency as compared to a random design optimized for estimation efficiency (Liu et al., 2001).

In the companion paper Liu and Frank (2004), the theoretical framework is extended to show that the fundamental trade-off between estimation efficiency and detection power also holds for experiments with multiple trial types. In fact, the form of the relation between efficiency and power was found to be identical to that previously presented for experiments with a single trial type, with the shape of the trade-off curves depending only on k, the number of parameters in the HRF. The theoretical model predicts fairly well the trade-off observed for $Q = 2$ trial types, but as the number of trial types increases, it becomes increasingly difficult to find designs that achieve the predicted trade-off, as shown in Fig. 1 of Liu and Frank (2004). A discussion of novel designs that can come closer to achieving the theoretical trade-off is provided in the Design section.

Based on the results of numerical simulations, Liu and Frank (2004) found that conditional entropy and estimation efficiency exhibited the empirical relation $H_r \approx \log_2(1 + Q\xi_{\mathrm{tot,norm}})$, where $\xi_{\mathrm{tot,norm}}$ is the total estimation efficiency normalized by its theoretical upper bound. Examples of the empirical relation are shown in Figs. 3 and 4. The empirical relation works well for first- and second-order conditional entropies but does not hold in some instances for third-order conditional entropies. The reasons for this are discussed in Liu and Frank (2004) and below in m-Sequence-based designs. The empirical relation between conditional entropy and estimation efficiency states that a reduction in estimation efficiency is accompanied by a reduction in conditional entropy. Since there is a fundamental trade-off between estimation efficiency and detection power, this empirical relation implies a similar trade-off between conditional entropy and detection power.

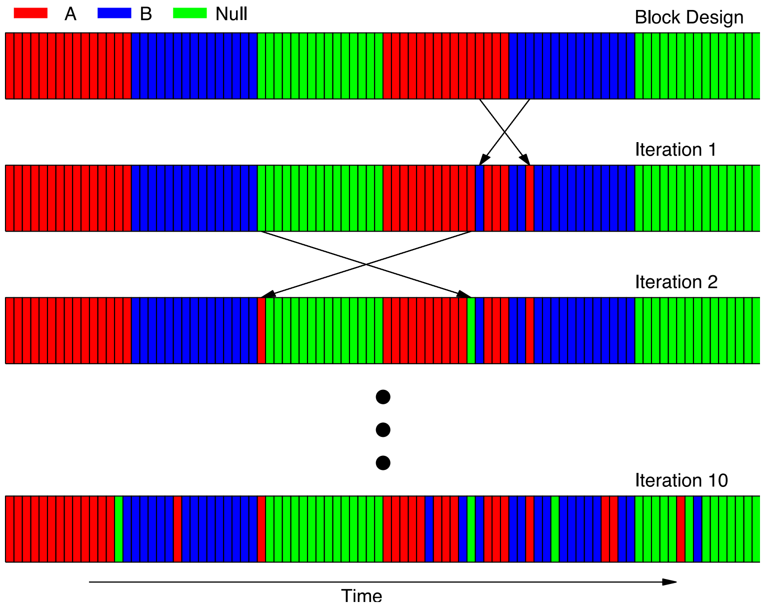


Fig. 1. Generation of permuted block designs.

The fundamental trade-off between estimation efficiency and detection power does not hold when a priori knowledge about the HRF in the form of basis function expansions is assumed. As shown in Appendix A6 of Liu and Frank (2004), the upper bound on the generalized definition of estimation efficiency with basis functions is greater than the bound on the estimation efficiency for an unbiased estimate by a factor of k^2/s^2, where s is the number of basis functions, which is assumed to be less than or equal to k. In addition, the design that maximizes estimation efficiency with basis functions is in general less random than the design that maximizes estimation efficiency and therefore exhibits decreased conditional entropy and increased detection power with respect to the random design. As a result, as we migrate from a random design to a semirandom design, both estimation efficiency with basis functions and detection power increase until a maximum in the estimation efficiency with basis functions is reached, after which the estimation efficiency with basis functions decreases as detection power increases (see for example Fig. 3 of Liu and Frank (2004)).

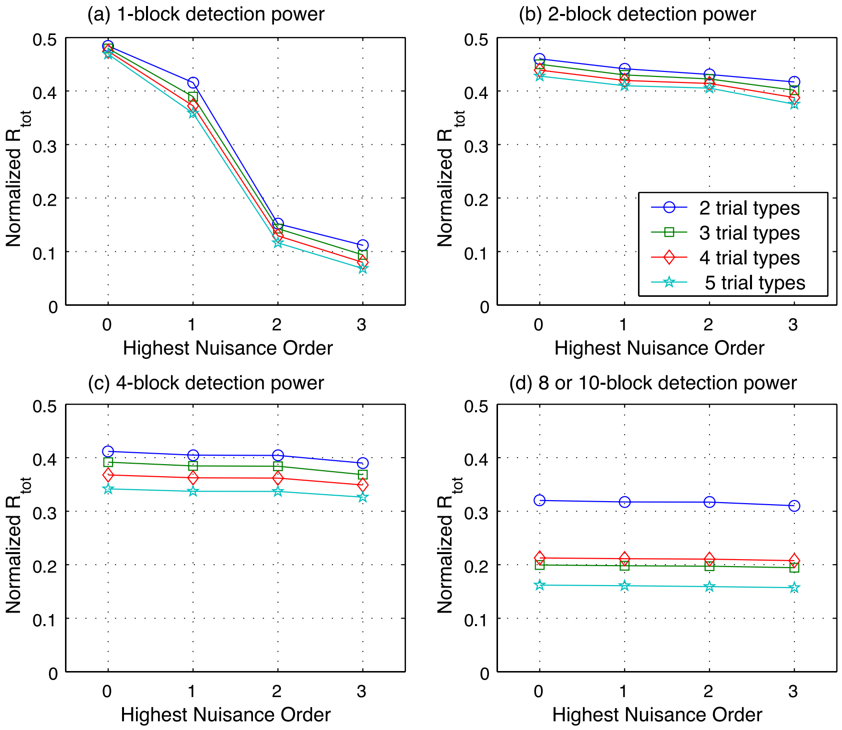


Fig. 2. Effect of nuisance functions on detection power for block designs with 1, 2, 4, and 8 (Q = 2, 4, and 5) or 10 (Q = 3) blocks. The horizontal axis is labeled by the highest order of Legendre polynomial used.

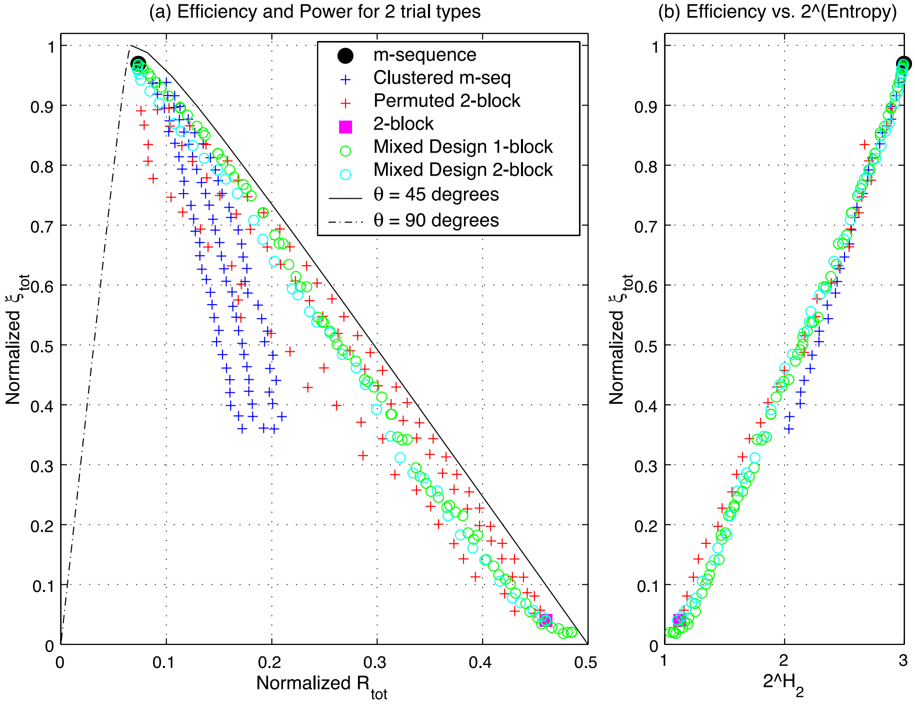


Fig. 3. Estimation efficiency, detection power, and second-order conditional entropy for a two-block design, permuted block designs, m-sequence design, clustered m-sequence designs, and mixed designs for experiments with two trial types. For permuted block and clustered m-sequence designs, the minimum, median, and maximum detection power of 1000 permutation paths are shown.

Optimal frequency of occurrence

A first step in the design of an experiment is to determine the optimum number of stimuli to use. The frequency of occurrence p is defined as the number of stimuli for a trial type divided by the total number of points in the experimental design. For a random design, it is the probability of occurrence of a stimulus. If we assume the same frequency of occurrence for all trial types, then it

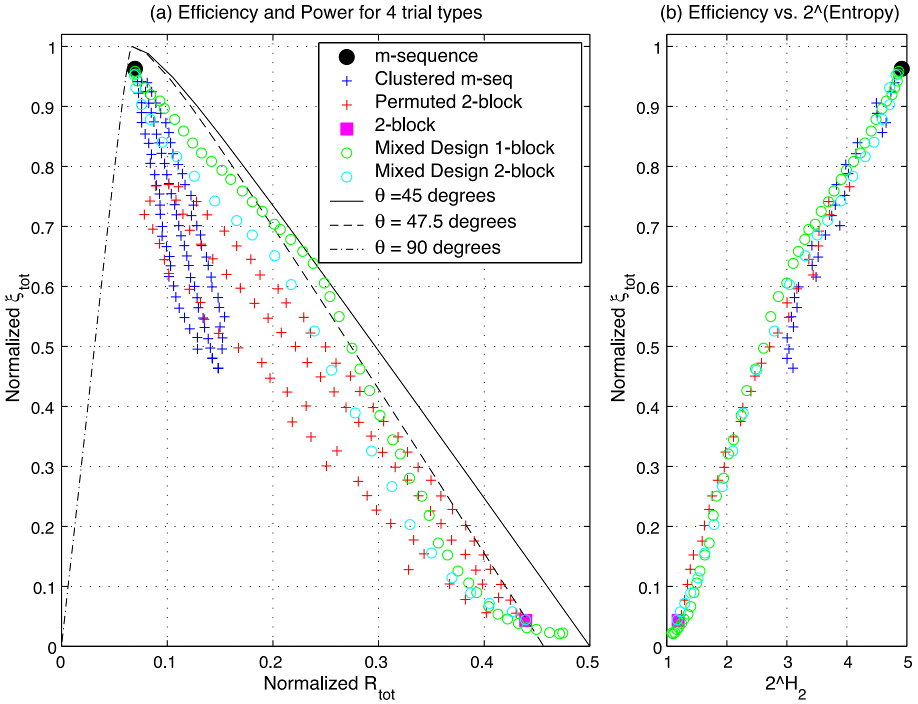
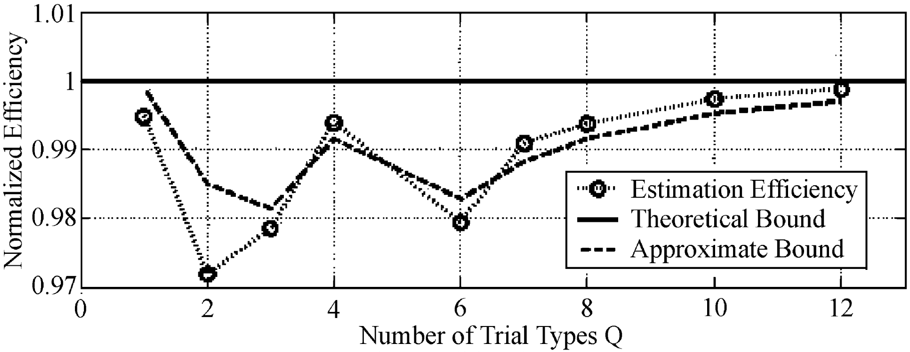


Fig. 4. Estimation efficiency, detection power, and second-order conditional entropy for a two-block design, permuted block designs, m-sequence design, clustered m-sequence designs, and a mixed designs for experiments with four trial types. For permuted block and clustered m-sequence designs, the minimum, median, and maximum detection power of 1000 permutation paths are shown.

(a) Estimation Efficiency



(b) Conditional Entropy

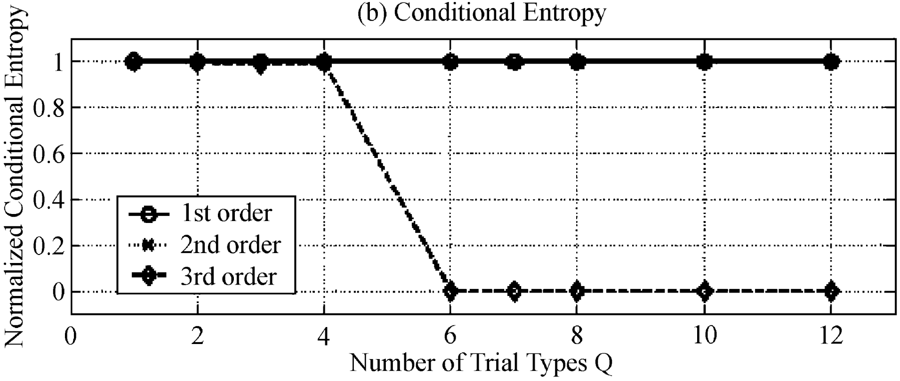


Fig. 5. Estimation efficiency and conditional entropy of m-sequence-based designs for 1, 2, 3, 4, 6, 7, 8, 10, and 12 trial types.

can be shown that the frequency of occurrence that maximizes both estimation efficiency and detection power is $p = 1/(Q + 1)$, where Q is the number of trial types (Liu and Frank, 2004). For example, in an experiment with two trial types, the optimal frequency of occurrence for each trial type is $p = 1/3$. The optimality of $p = 1/(Q + 1)$ has been previously stated using the concept of null events (Burock et al., 1998; Friston et al., 1999). In some experiments, it may be advantageous to modify the definitions of estimation efficiency and detection power to adjust the relative weights of individual trial types and pairwise contrasts. The optimal frequency of occurrence then depends on the relative weighting. For example, if we are interested only in estimating the HRFs or detecting activations for individual trial types, then the optimal frequency is $p = (Q - \sqrt{Q})/(Q^2 - Q)$ for $Q > 1$. On the other hand, if we are interested only in estimating the pairwise contrasts between HRFs or detecting pairwise contrasts of activations, then the optimal frequency is $p = 1/Q$. For an experiment with two trial types, the optimal frequency is 0.2929 when only individual trial types are of interest and 0.5 when only pairwise contrasts are of interest. For other relative weightings, the formula for the optimal frequency of occurrence is stated in Eq. (23) of Liu and Frank (2004). To simplify the presentation, we assume $p = 1/(Q + 1)$ for the remainder of the paper.

Bounds on estimation efficiency and detection power

In the search for optimal designs, knowledge of the upper bounds on performance is useful in deciding when to terminate the search. In Eqs. (26) and (27) of Liu and Frank (2004), it is shown that the upper bounds on estimation efficiency and detection power are

$$\xi_{\mathrm{tot}} \le \beta/k \text{ and } R_{\mathrm{tot}} \le \beta k \qquad (3)$$

where $\beta = N/(2(Q + 1))$. The upper bounds increase with the length N of the experiment, reflecting the improved statistical efficiency obtained by acquiring more data. The bounds are inversely proportional to the number of trial types Q, reflecting the reduction in data per trial type as the number of types increases. The bound on estimation efficiency decreases with the number of unknown parameters k in the HRF. This reflects the increase in the variance of each HRF estimate, which is the sum of the variances of the k parameter estimates. In contrast, the bound on detection power increases with k, reflecting the fact that the maximum obtainable energy of the time series obtained by convolving the stimulus convolved with the HRF increases with the temporal width of the HRF (Liu et al., 2001). This is most easily understood by considering a block design convolved with a fictitious HRF composed of all 1's. As the length of the HRF increases, the overlap between the responses to adjacent stimuli increases and the overall energy of the resulting signal also increases.

As shown below, it is possible in practice to come quite close to the theoretical upper bound on estimation efficiency using m-sequence-based designs. It does not, however, appear possible to obtain the upper bound on detection power. To understand why this is the case, we need to consider two questions: (1) For an assumed HRF, what is the experimental design that maximizes detection power; and (2) for a given design, what is the HRF that maximizes detection power? For the HRFs typically encountered in fMRI (e.g., a gamma density model), it has been shown that detection power is maximized by experimental designs that concentrate most of their energy at low frequencies (Birn et al., 2002). For example, in a one-trial-type ($Q = 1$) experiment, a block design with only one on or off period maximizes detection power when the only nuisance term is a constant term (Liu et al., 2001). To answer the second question, the intuition is that detection power is maximized by an HRF with most of its spectral energy centered about the fundamental frequency of the stimulus. In more technical terms, the HRF that maximizes detection power is equal to the dominant eigenvector associated with the design (Liu and Frank, 2004; Liu et al., 2001). For a low frequency block design, this eigenvector has the property that it attempts to concentrate its energy around the fundamental frequency of the stimulus. The gamma density HRF model has a much broader bandwidth than the dominant eigenvector and thus yields significantly lower detection power. This effect can be approximated with the bound

$$R_{\mathrm{tot}} \leq \beta k \cos^2 \theta_{\min} \tag{4}$$

where $\theta_{\min}$ is empirically defined as the angle between the assumed HRF (e.g., gamma density model) and the dominant eigenvector for a design with only one on-off period. In practice, we have found that $\theta_{\min} = 45°$ yields a reasonable empirical bound. Additional technical details of the above arguments are provided in Liu and Frank (2004) and Liu et al. (2001).

Experimental designs

In the following sections, we describe some approaches to designing experiments that attempt to achieve the theoretically predicted trade-off between estimation efficiency and detection power.

Block designs

Block designs in which stimuli from the same trial type are clustered into blocks are typically used in experiments where the detection of activation is the primary goal. As shown in Figs. 3 and 4, the estimation efficiency and conditional entropy of these designs are very low. For a given experiment length N, a block design can be characterized by the number of blocks B of each trial type that are present, with the number of events in each block equal to $N/B/(Q + 1)$. An example of a two-block design for an experiment with two trial types is shown in the top row of Fig. 1. The experiment length is $N = 90$ and each block contains 15 events.

In choosing the optimal number of blocks, an important consideration is the robustness of the detection power of the design to the presence of nuisance terms that model the low frequency drifts commonly observed in fMRI data. A design is robust when it is nearly orthogonal to the nuisance terms (Liu et al., 2001). Fig. 2 shows the detection powers for designs with experimental length $N = 240$, two to five trial types, and 1 to 10 blocks per trial type, plotted versus the number of nuisance terms used. Panel d shows the performances for a 10-block design for $Q = 3$ and 8-block designs for $Q = 2$, 4, and 5. Consistent with the simulations described in Liu and Frank (2004), the time scale of the simulations in this and following sections is $\Delta t = 1$ s and the HRF is a gamma density function of the form $h[j] = (\tau n!)^{-1}((j - 1)\Delta t/\tau)^n e^{-j\Delta t/\tau}$ for $j \geq 1$ and 0 otherwise, where $n = 3$ and $\tau = 1.2$. The low frequency nuisance terms are chosen to be Legendre polynomials of orders 0 to 3, with the increasing orders corresponding to a constant term, a linear term, a quadratic term, and a cubic term. The Legendre polynomials are the orthogonalized version of the regular polynomials commonly used in fMRI analyses (Cox, 1996) and thus span the same subspace as the regular polynomials. The horizontal axis is labeled by the order of the highest Legendre polynomial used so that when the highest order is 3, polynomials of orders 0 to 3 are used as nuisance terms. Another commonly used set of nuisance terms are the low frequency sine and cosine functions (Friston et al., 1995). The detection powers of the one-block designs decrease rapidly as the number of nuisance terms increases, while the detection powers of the two- and four-block designs decrease only slightly. Overall, the two-block designs offer higher detection power than the four-block designs and are thus preferred from the point of view of statistical efficiency. However, other factors, such as subject fatigue or habituation may factor into the selection of the optimal number of blocks. For example, with $N = 240$ and two trial types, the two-block design with 40 events per block may lead to habituation effects that are decreased by going to a four-block design with 20 events per block.

Permuted block designs

Because of their high detection power, block designs are a good starting point for generating designs that attempt to optimally decrease detection power in exchange for an increase in estimation efficiency. A straightforward approach is to begin with a block design and obtain new designs by exchanging the positions of two randomly chosen events (Buxton et al., 2000). The events are chosen without regards to the trial type of the event. With each iteration, the designs become increasingly random. A graphical description of this method is shown in Fig. 1 and an additional example is shown in Fig. 7.

The performance of permuted block designs for experiments of length $N = 240$ with two and four trial types is shown in Figs. 3 and 4, respectively. The initial starting point for the iteration process is a two-block design, which was chosen because of its robustness to nuisance terms, as compared to the one-block design. Each iteration path consisted of 100 iteration steps, and a total of 1000 unique iteration paths were generated. In other words, each of the 1000 iteration paths started with the same two-block design but followed a different trajectory of random permutations. To reduce the number of points plotted, the estimation efficiencies are divided into 150 equally spaced bins and the minimum, median, and maximum detection powers for a subset of these bins are plotted. The metrics are normalized by the theoretical upper bounds stated in Eq. (3). Curves showing the predicted trade-off between estimation efficiency and detection power are also shown. These are computed using the equations provided in Appendix A with the angle θ equal to 45° and 47.5° for two and four trial types, respectively. The choice of angle is empirical and meets two criteria: the maximum theoretical detection power is greater than the maximum observed detection power, and the slope of the predicted trade-off is approximately equal to the slope of the observed trade-off in the region near the maximum observed detection power.

For the experiment with two trial types shown in Fig. 3, the increase in estimation efficiency as the maximum detection power per bin decreases is very close to the predicted trade-off for normalized efficiencies up to 0.8. For higher normalized efficiencies, the gain in efficiency as detection power decreases is less than predicted. For the experiment with four trial types shown in Fig. 4, the predicted trade-off is closely approximated for efficiencies up to 0.4. This represents an increase in performance over the permuted two-block design shown in Fig. 1 of Liu and Frank (2004) where only one randomly chosen iteration path was

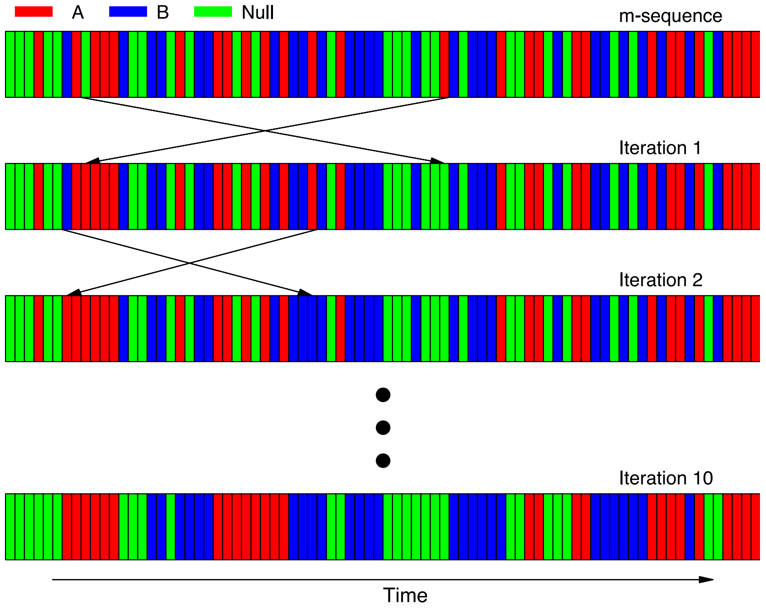


Fig. 6. Generation of clustered m-sequence designs.

displayed. Thus, searching over a large enough space of iteration paths is a critical factor in attaining designs that come closer to the predicted trade-off. For the two- and four-trial-type experiments, a search over 100 iteration paths (results not shown) provided only slightly less performance than the search over 1000 iteration paths.

As shown in Fig. 4, the maximum normalized estimation efficiency obtained by the permuted two-block designs with four trial types is about 0.8, which is close to the maximum efficiency obtained by a search over 1000 random designs shown in Fig. 1 of Liu and Frank (2004). This observation is consistent with the fact that as the number of iterations increases, the permuted block design approaches a random sequence. The implication for experimental design is that the maximum achievable efficiency obtainable by permuted block designs will be similar to that of random designs. As shown in

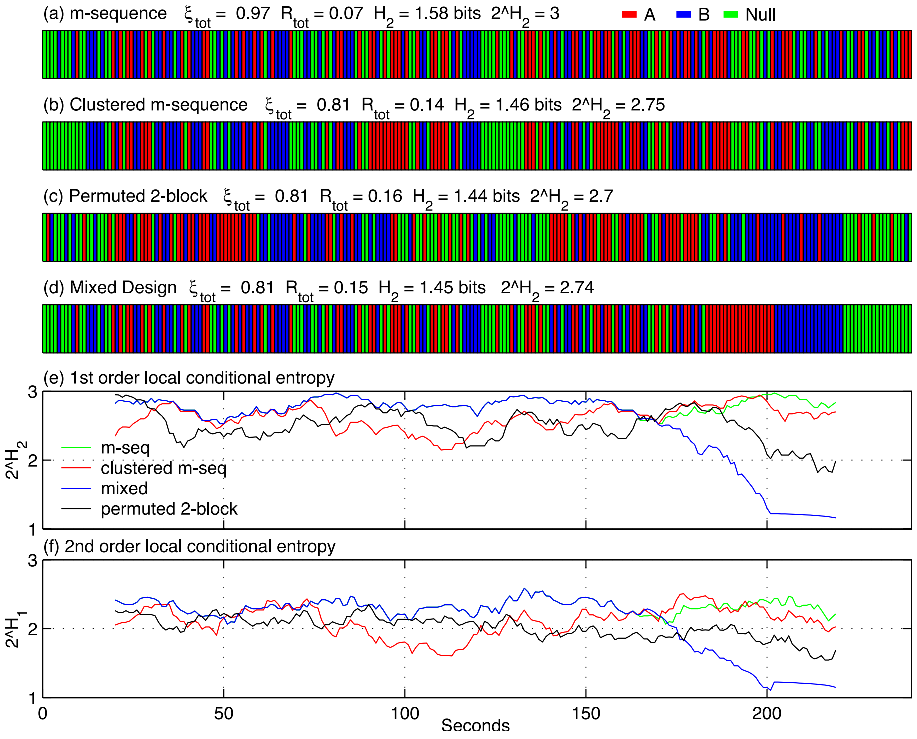


Fig. 7. Stimulus patterns of m-sequence, clustered m-sequence, permuted block, and mixed designs for two trial types and first- and second-order local conditional entropy.

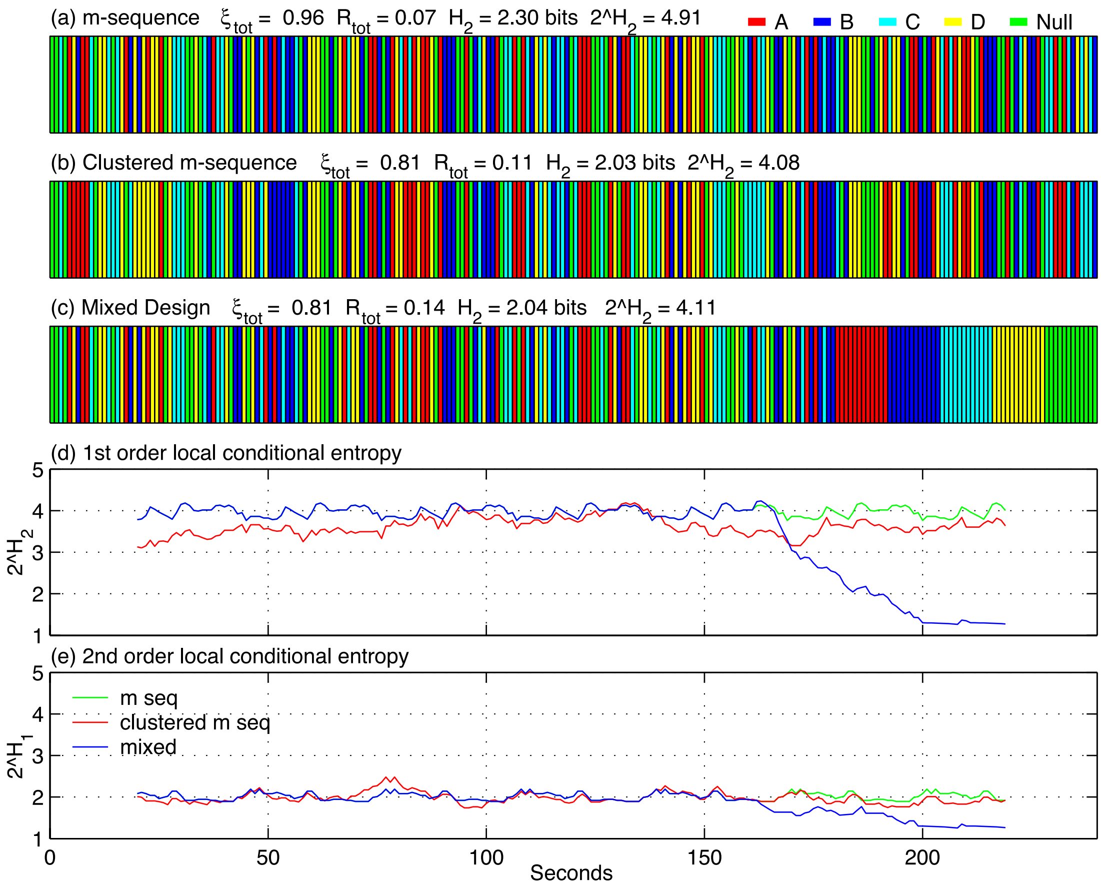


Fig. 8. Stimulus patterns of m-sequence, clustered m-sequence, and mixed designs for four trial types and first- and second-order local conditional entropy.

Fig. 1 of Liu and Frank (2004), the maximum normalized efficiency achievable with random designs drops as the number of trial types increases, thus limiting the ability of permuted block designs to achieve the predicted trade-off when the desired estimation efficiency is high. This limitation can be partially addressed with the m-sequence and clustered m-sequence designs discussed in the next two sections.

m-Sequence-based designs

Experimental designs based on maximal length sequences or m-sequences were introduced into the fMRI literature by (Buracas and Boynton, 2002). These designs were shown to exhibit significantly greater estimation efficiency than designs obtained via random search, especially as the number of trial types increased. The gain in efficiency is due to the fact the m-sequences have nearly ideal temporal autocorrelation properties so that a shifted m-sequence is nearly orthogonal to the original sequence. A design for Q trial types is generated from an m-sequence with $L = Q + 1$ levels. For example, a three trial type experiment would be generated from a four-level m-sequence of the form ...0, 1, 3, 2, 0, 3, 1, 2,...where 0 is assigned to the null condition and each of the three trial types is indexed by the positive numbers in the m-sequence.

m-Sequences are known to exist when the number of levels L is either a prime or a power of a prime (Godfrey, 1993). An L level m-sequence of length $L^n - 1$ is generated using an L-level shift register bank with n stages and appropriate feedback taps. For example, a three-level shift register (ternary logic) with five stages is used to generate a three-level m-sequence of length 242, while a four-level shift register with four stages is used to generate a four-level m-sequence of length 255.

An alternative method of generating m-sequence-based designs when L is a power of a prime is to use hybrid sequences constructed from the integer weighted sum of shifted prime level sequences (Buracas and Boynton, 2002). For example, a four-level hybrid sequence $m_{\text{hybrid}}[n]$ can be constructed from the weighted sum $m_{\text{hybrid}}[n] = 2m[n] + m[n + M]$ of shifted two-level m-sequences $m[n]$ and $m[n + M]$, where the appropriate shift M is chosen empirically. This method has the advantage of increasing the number of designs that are available for selection. The conditional entropy of the hybrid sequences appears to be slightly lower, however, than that of the power of prime m-sequences, and a more detailed comparison of the two approaches would be useful. Examples of m-sequence designs are shown in Figs. 6–8.

To demonstrate the performance of m-sequence-based designs, we constructed m-sequences with Q = 1, 2, 3, 4, 6, 7, 8, 10, 12 and corresponding lengths N = 255, 242, 255, 624, 342, 511, 728, 1330, and 2196. These were computed using MATLAB codes described in Buracas and Boynton (2002) for Q = 1, 2, and 4 and algorithms detailed in Godfrey (1993) for Q = 3, 6, 7, 8, 10, and 12. We are not aware of m-sequence-based designs for Q = 5, 9, and 11. Fig. 5a shows the estimation efficiencies of the designs. Also plotted are the theoretical upper bound on estimation efficiency β/k and an approximate bound on efficiency $M_A(p,N)\beta/k^2$ where $M_A(p,N) = \sum_{q=1}^{k} p(1-p(1-(q-1)/N))(N-(q-1))$

was derived in (Liu et al., 2001). The approximate upper bound accounts for the fact that some small fraction of events are shifted out of the design matrix during the convolution process, whereas the theoretical upper bound assumes that no events are shifted out. Note that all quantities are normalized by the upper bounds. The estimation efficiencies of the m-sequence-based designs attain at least 97% of the upper bound. The closeness to the bound depends in part on the length of the design used, since in longer designs fewer events are shifted out of the design matrix. This dependence on length is shown by the behavior of the approximate bound, which is a function of both the frequency of occurrence p and the length N.

Fig. 5b shows the first- through third-order conditional entropies of the designs normalized by the upper bound $\log_2(Q + 1)$. The first- and second-order conditional entropies are all at least 99.5% of the upper bound. The third-order conditional entropies are at least 99% of the upper bound for experiments with up to four trial types but are identically zero for designs with 6 to 12 trial types. This reflects the fact that the m-sequences for the zero entropy designs are all based on three-stage shift registers so that knowledge of the three previous trial types completely determines the next trial type. In addition, it is important to note that entropy is not a function of the number of trial types since a shorter four-trial-type design with length 124 based on a three-stage shift register would also have a third-order conditional entropy equal to zero, as shown in Fig. 2 of Liu and Frank (2004). In practice, we expect that the theoretical third-order entropy grossly underestimates the perceived entropy of the design. For example, the theoretical third-order entropy of the six trial design is based on perfect memory of which trial type followed each of $7^3 = 343$ combinations of three previous trial types, including the possibility of a null condition. Since a subject is unlikely to remember so many combinations, the perceived entropy will probably be higher than the theoretical entropy. A reasonable but untested conjecture is that the first- and second-order conditional entropies are sufficient to characterize most designs with four trial types or less (a maximum of 25 combinations), while for designs with more than four trial types, the first-order conditional entropy is probably sufficient.

Clustered m-sequences

Because m-sequences offer nearly ideal estimation efficiency, they serve as a good starting point for designs that attempt to optimally decrease estimation efficiency to increase detection power. In this section, we describe a method for randomly permuting an m-sequence to generate a new type of sequence that we call a clustered m-sequence. The basic idea behind the approach is to increase the clustering of events of the same trial type at each cycle of the permutation process.

The steps of the process are as follows:

Step 1. Find the smallest hole for this trial type. A hole is defined as a gap between successive events of the same trial type. For example, a sequence of the form ...BCAA CBAAA... has a hole of size 2 between the blocks of trial type A. If there is more than one hole with the smallest size, randomly pick one of these holes.

Step 2. Find a filler for the hole. First, look for all singletons of this trial type, where a singleton is defined as an event with no adjacent events of the same trial type. For example, in the sequence BBBACABCAA, the first two events of trial type A are considered to be singletons. Find the singleton that is farthest away from any other event of this trial type, randomly picking a singleton if there is a tie. The singleton that is picked is the filler for the hole. If no singletons exist, then find the smallest blocks of events with this trial type. For example, in the sequence BBBAACAAABCCBAAB, the first and third blocks of events with trial type A are of length 2 and have the smallest size. From the population of smallest blocks, pick the block that is farthest away from the closest adjacent block (which can be of any size), and then randomly pick an event within the block to be the filler. If there is more than one block with the smallest size, randomly pick one of these blocks.

Step 3. Fill in the hole. Exchange the trial type of the hole with the trial type of the filler. For example, in a sequence of the form BBCAABAACBCA, the hole for trial type A is at event number 6 and the filler is the last event at position 12. After the exchange, event number 6 has trial type A, while event number 12 has trial type B.

Step 4. Using the next trial type, go to Step 1.

A graphical description of the clustering process is shown in Fig. 6. The top row shows a two-trial-type design based on a three-level m-sequence of length 80. At the first iteration, a hole of trial type A is filled; and at the second iteration, a hole of trial type B is filled. As the number of iterations increases, the clustered m-sequences become more block-like. Additional examples of clustered m-sequence designs are shown in Figs. 7 and 8.

The performance of clustered m-sequence designs for experiments of length $N = 240$ with two and four trial types is shown in Figs. 3 and 4, respectively. For the two-trial-type designs, the starting point for the clustering process is a three-level m-sequence of length 242 that is truncated to 240 points. For the four-trial-type designs, the starting point is a five-level 124-point m-sequence that is repeated and then truncated to 240 points. The estimation efficiencies of the m-sequence designs are slightly lower than those for the designs shown in Fig. 5, which were not truncated to 240 points. Each clustering path consisted of 30 iterations, and a total of 1000 clustering paths were generated. Other details, such as the method used to reduce the number of displayed points and the generation of the theoretical curves, are described in the section on permuted block designs.

For the experiment with two trial types, the increase in the detection power as the estimation efficiency decreases approximates the predicted optimal trade-off down to a normalized efficiency of 0.9. Below that, the detection power does not increase as quickly as predicted. For most of the region where the clustered m-sequences do not perform well, the permuted block designs approximate the predicted trade-off so that the combination of clustered m-sequences and permuted block designs provides nearly optimal performance over the entire range of estimation efficiencies.

For the experiment with four trial types, the clustered m-sequences offer a range of estimation efficiencies and detection powers that are not achievable with the permuted block designs but fall short of attaining the optimal trade-off. For example, at a normalized estimation efficiency of 0.8, the highest detection power achieved by a clustered m-sequence is about 30% lower than the predicted maximum detection power.

Mixed designs

A mixed design is formed by concatenating a block design with a semirandom design, such as a permuted block design or a clustered m-sequence design (Liu et al., 2001). The simplest mixed design is obtained by concatenating a block design with an m-sequence design. The idea is that the block design contributes high detection power, while the m-sequence design contributes high estimation efficiency. By varying the relative lengths of the block and m-sequence designs, different trade-offs between estimation efficiency and detection power can be obtained.

To demonstrate the performance of mixed designs, a series of designs for the two- and four-trial-type experiments were generated. Each design consisted of an m-sequence design followed by either a one- or two-block design. The length L_B of the block design was varied from $(Q+1)$ to N in steps of $Q+1$ with $N = 240$ points. The corresponding length of the m-sequence design was $N - L_B$. Examples of mixed designs using a one-block design are shown in Figs. 7d and 8c for $Q = 2$ and 4 trial type experiments, with $L_B = 57$ and 60, respectively.

The performance of the mixed designs is shown in Figs. 3 and 4. For two-trial-type designs, the one-block mixed designs provide slightly better performance than the clustered m-sequence and permuted block designs for a normalized efficiency of about 0.85. For four-trial-type designs, the one-block mixed designs provide significantly better performance than both the clustered m-sequence and permuted block designs for normalized efficiencies ranging from 0.4 to 0.85, while the two-block mixed designs exhibit significantly smaller increases in performance for normalized efficiencies of 0.6 to 0.75. The excellent performance of the one-block mixed designs must be weighed, however, against their sensitivity to nuisance terms. As shown above, the detection power of one-block designs decreases rapidly as the number of nuisance terms increases, and one-block mixed designs inherit some of this sensitivity. When the highest order nuisance term is a linear term, the detection power of the one-block mixed designs decreases significantly from what is shown in Fig. 4 for normalized efficiencies below 0.3 (results not shown). With higher order terms (quadratic and cubic) included, the decrease in detection power occurs for normalized efficiencies below about 0.75. In contrast, the detection power of the two-block mixed designs is robust to nuisance terms over the entire range of performance.

With the use of the one-block mixed designs constrained to regions (normalized efficiencies greater than 0.75) where they are robust to nuisance terms, the addition of both the one- and two-block mixed designs to the clustered m-sequence and permuted block designs provides a selection of designs for four trial types that comes very close to providing the predicted trade-off obtained with $\theta = 47.5°$ (see Permuted block designs for choice of this angle) over the entire range of normalized estimation efficiencies. In other words, it appears that the theoretically predicted trade-off can serve as an accurate guide for what can be achieved in practice.

In considering the usefulness of mixed designs in attaining optimal trade-offs, one additional point must be considered. This is the fact that the local conditional entropy of mixed designs varies greatly over the time course of the design. Figs. 7 and 8 show examples of m-sequence designs, clustered m-sequence designs, and mixed designs for two and four trial types, respectively. In addition, a permuted block design is shown for the two trial types. The clustered m-sequence designs, mixed designs, and permuted block designs were all chosen to have a normalized estimation efficiency of about 0.8. For two trial types, the clustered m-sequence, permuted block design, and mixed design all provide at least twice the detection power of the m-sequence design with less than a 10% decrease in overall randomness. For four trial types, the clustered m-sequence and mixed design yield 57% and 100% increases in detection power, respectively, with a 20% decrease in overall randomness. The conditional entropy of the mixed designs reflects an average of high conditional entropy due to the m-sequence portion of the design and low conditional entropy due to the block design portion. This is shown in the plots of first- and second-order local conditional entropy where a window width of $W = 40$ was assumed. The mixed designs show a significant decrease in local entropy during the block design portion, while the clustered m-sequence and permuted block designs exhibit a smaller variation of local entropy around their respective average values. The large variance in local entropy may make the mixed designs unsuitable for certain experiments despite the advantages they offer in terms of statistical efficiency.

Discussion

In this paper, we have examined the design of multiple trial type fMRI experiments that attempt to achieve an optimal trade-off between estimation efficiency, detection power, and conditional entropy. Block designs offer maximal detection power but low estimation efficiency and conditional entropy, while m-sequence designs offer nearly optimal estimation efficiency and conditional entropy but low detection power. Intermediate trade-offs between estimation efficiency and detection power can be obtained with permuted block designs, clustered m-sequence designs, and mixed designs. For two- and four-trial-type experiments, these designs were shown to closely approach the theoretically predicted trade-off over the entire range of estimation efficiencies. As an example of an advantageous trade-off, it was shown that for a two-trial-type experiment, a permuted block design could provide over twice the detection power of an m-sequence design with only a 10% decrease in randomness and a 20% decrease in estimation efficiency. Such a design would be useful for an experiment in which detection power is the primary goal, but a high degree of randomness is needed to reduce psychological confounds such as anticipation.

While the design methods presented are generally applicable to experiments with any number of trial types and any length, there are some limitations that could be addressed by future investigations. m-Sequence-based designs currently exist only for experiments where $Q + 1$ is either a prime number or a power of a prime number. Thus, designs based on m-sequences, such as clustered m-sequence and mixed designs, do not exist for experiments with $Q = 5$, 9, or 11 trial types. For such experiments, it is unlikely that a random search will yield a design with optimal estimation efficiency because of the extremely large number of possibilities, for example, 5.7×10^{186} possible sequences for a 240-point design with five trial types (Buracas and Boynton, 2002). Other approaches, such as genetic algorithms (Wager and Nichols, 2003), may be useful for obtaining a design with optimal efficiency. m-Sequence designs are also restricted to experimental lengths $(Q+1)^r - 1$, where r is the number of stages in the shift register. In practice, this is not too severe of a restriction because m-sequence designs appear to maintain most of their estimation efficiency even when they are truncated. For example, the estimation efficiency of mixed designs, which provide nearly optimal trade-offs, is due primarily to the estimation efficiency of truncated m-sequence designs.

As the number of trial types grows, it may become more difficult to find clustered m-sequence and permuted block designs that achieve advantageous trade-offs. The methods described in this paper for generating these designs may not be the most efficient for searching the space of possible designs, and it is possible that other approaches such as genetic algorithms, dynamic stochastic designs (Friston et al., 1999), varying of minimum block durations (Birn et al., 2002), or some combination of all these techniques may provide better performance. For example, the performance of a genetic algorithm that occasionally inserts m-sequences and block designs into the population of designs would be interesting to investigate.

The metrics for estimation efficiency and detection power were defined using the variances of all individual trial types and all unique pairwise contrasts between trial types. With this definition, the optimal frequency of occurrence is $p = 1/(Q+1)$, where the same frequency for each trial type is assumed. For other weightings of individual trial types versus pairwise contrasts, the optimal frequency is different, with a frequency of occurrence greater than $1/(Q+1)$ as pairwise contrasts are weighted more heavily and a frequency less than $1/(Q+1)$ as individual trial types are weighted more heavily. In some cases, only a small subset of estimates may be of interest, and the restriction that all trial types have the same frequency of occurrence may not be appropriate. For example, in a five-trial-type experiment where only estimates for trial types A, B, and C and pairwise contrasts D–E and A–E are of interest, optimal statistical efficiency may be achieved when the individual trial types have different frequencies of occurrence. While an extension of the theoretical framework of Liu and Frank (2004) to calculate the optimal frequency of occurrence per trial type would be useful, the assumption of the same frequency of occurrence across trial types is probably not too restrictive in practice and should serve as a good starting point for most designs.

It was shown that significant gains in detection power can be obtained if a decrease in conditional entropy is acceptable. Further work is required to determine how much entropy is required to sufficiently minimize psychological confounds. In addition, as discussed in the section on m-sequence designs, the theoretical entropy is probably lower than the perceived entropy, especially as the number of trial types increases. A comparison of the theoretical and perceived entropies of sequences would be useful and would potentially allow the experimenter to accept greater decreases in theoretical entropy.

The examples in this paper have assumed that estimation efficiency is calculated without the use of basis function expansions and that the additive noise is uncorrelated. As shown in Liu and Frank (2004), the assumption of either basis function expansions for the HRF or correlated noise modifies the trade-off between estimation efficiency and detection power so that intermediate designs, such as clustered m-sequences and permuted block designs, can provide both higher detection power and estimation efficiency than obtained by m-sequence designs. The form of the trade-off depends, however, on the assumptions the experimenter is willing to make regarding the choice of basis functions and the model for noise correlation. In practice, it may be difficult to make a set of assumptions that are appropriate for all subjects involved in a study. For example, the form of correlated noise can show great variance across subjects and even experimental runs (Buracas and Boynton, 2002). As a result, selecting an optimal design with a minimum of assumptions may prove to be the best overall strategy, especially if the goal is to balance conditional entropy versus detection power while ignoring estimation efficiency. This is because conditional entropy does not depend upon the assumptions, while detection power already assumes a known HRF and is to first order just scaled by a constant that depends on the noise correlation and the assumed HRF (Liu and Frank, 2004).

Together with the theoretical framework established in Liu and Frank (2004), the designs described in this paper provide a starting point for selecting the design that is optimal for a specific experiment. In practice, the choice of an optimal design requires numerical simulations that allow the experimenter to explore the various trade-offs involved and to assess the impact of experimental assumptions regarding nuisance terms, additive noise, and the shape of the HRF. To facilitate this process, a MATLAB toolbox for the design of multiple trial type fMRI experiments is available for download at http://fmriserver.ucsd.edu/ttliu/.

Acknowledgments

This work was supported in part by a Biomedical Engineering Research Grant from the Whitaker Foundation. Thanks to Larry Frank for his comments on the manuscript.

Appendix A

The definition of conditional entropy stated in Eq. (28) of Liu and Frank (2004) is repeated here for the convenience of the reader. For a stimulus pattern of the form $a_1, a_2, \ldots, a_N$ where the trial type a_i for the ith event is drawn from an alphabet A, the conditional entropy is

$$H_r = -\sum_{a_i \in A} \sum_{a_{i-1} \in A} \cdots \sum_{a_{i-r} \in A} p(a_{i-r}, \ldots, a_{i-1}, a_i) \log_2 p(a_i \mid a_{i-r}, \cdots, a_{i-1})$$

where $p(a_{i-r}, \ldots, a_{i-1}, a_i)$ is the probability of the sequence $a_{i-r}, \ldots, a_{i-1}, a_i$ of $(r+1)$ trial types occurring, and $p(a_i | a_{i-r}, \ldots, a_{i-1})$ is the conditional probability of trial type a_i occurring after the sequence of r trial types $a_{i-r}, \ldots, a_{i-1}$.

As derived in Liu and Frank (2004), the theoretical expressions for detection power and estimation efficiency are

$$R_{\mathrm{tot}} = \frac{Nk}{2(Q+1)} \left(\alpha \cos^2\theta + \frac{1-\alpha}{k-1} \sin^2\theta \right)$$

$$\xi_{\mathrm{tot}} = \frac{Nk}{2(Q+1)} \frac{\alpha(1-\alpha)}{1+\alpha(k^2-k)}$$

where $1/k \le \alpha \le 1$ and the angle θ is used to empirically model the relative detection power of different block designs.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8, 360–369.

Birn, R.M., Cox, R.W., Bandettini, P.A., 2002. Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. NeuroImage 15, 252–264.

Bischoff-Grethe, A., Martin, M., Mao, H., Berns, G.S., 2001. The context of uncertainty modulates the subcortical response to predictability. J. Cogn. Neurosci. 13, 986–993.

Buracas, G.T., Boynton, G.M., 2002. Efficient design of event-related fMRI experiments using m-sequences. NeuroImage 16, 801–813.

Burock, M.A., Buckner, R.L., Woldorff, M.G., Rosen, B.R., Dale, A.M., 1998. Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. NeuroReport 9, 3735–3739.

Buxton, R.B., Liu, T.T., Martinez, A., Frank, L.R., Luh, W.-M., Wong, E.C., 2000. Sorting out event-related paradigms in fMRI: the distinction between detecting an activation and estimating the hemodynamic response. NeuroImage 11, S457.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Cox, R.W., 1996. AFNI-software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Friston, K.J., Frith, C.D., Turner, R., Frackowiak, R.S.J., 1995. Characterizing evoked hemodynamics with fMRI. NeuroImage 2, 157–165.

Friston, K.J., Josephs, O., Rees, G., Turner, R., 1998. Nonlinear event-related responses in fMRI. Magn. Reson. Med. 39, 41–52.

Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N.A., Dale, A.M., 1999. Stochastic designs in event-related fMRI. NeuroImage 10, 607–619.

Godfrey, K., 1993. Perturbation Signals for System Identification. Prentice-Hall, Englewood Cliffs, NJ.

Liu, T.T., Frank, L.R., 2004. Efficiency, power, and entropy in event-related fMRI with multiple trial types—Part I: theory. NeuroImage 21, 387–400.

Liu, T.T., Frank, L.R., Wong, E.C., Buxton, R.B., 2001. Detection power, estimation efficiency, and predictability in event-related fMRI. NeuroImage 13, 759–773.

Miller, K.L., Luh, W.-M., Liu, T.T., Martinez, A., Obata, T., Wong, E.C., Frank, L.R., Buxton, R.B., 2001. Nonlinear temporal dynamics of the cerebral blood flow response. Hum. Brain Mapp. 13, 1–12.

Wager, T.D., Nichols, T.E., 2003. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. NeuroImage 18, 293–309.