This article was originally published in Brain Mapping: An Encyclopedic Reference, published by Elsevier, and the attached copy is provided by Elsevier for the author’s benefit and for the benefit of the author’s institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who you know, and providing a copy to your institution’s administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier’s permissions site at:

http://www.elsevier.com/locate/permissionusematerial

**Design Efficiency**

RN Henson, MRC Cognition and Brain Sciences Unit, Cambridge, UK

© 2015 Elsevier Inc. All rights reserved.

---

**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(p)</td>
<td>Autoregressive model of order p</td>
</tr>
<tr>
<td>BOLD</td>
<td>Blood oxygenation level-dependent (signal normally measured with fMRI)</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete cosine transform</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite impulse response (basis set)</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional magnetic resonance imaging</td>
</tr>
<tr>
<td>GLM</td>
<td>General linear model</td>
</tr>
<tr>
<td>HRF</td>
<td>Hemodynamic response function</td>
</tr>
<tr>
<td>ReML</td>
<td>Restricted maximum likelihood</td>
</tr>
<tr>
<td>SOA</td>
<td>Stimulus-onset asynchrony (SOA_{min} = minimal SOA)</td>
</tr>
<tr>
<td>TR</td>
<td>Interscan interval (repetition time)</td>
</tr>
</tbody>
</table>

**Formal Definition of Efficiency**

The general linear model (GLM) normally used for mass-univariate statistical analysis of functional magnetic resonance imaging (fMRI) data can be written for a single voxel as

\[ y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 C_\epsilon) \]  \[\text{[1]}\]

where \( y \) is an \( N \times 1 \) column vector of the data time series sampled every TR for \( N \) scans, \( X \) is an \( N \times P \) design matrix in which the \( P \) columns are regressors for the time series of predicted experimental effects, \( \beta \) is a \( P \times 1 \) column vector of parameters for each regressor in \( X \) (whose values are estimated when fitting the model to the data), and \( \epsilon \) is \( N \times 1 \) vector of residual errors. The second expression in eqn [1] denotes that the residuals come from a zero-mean, multivariate normal (Gaussian) distribution with covariance \( C_\epsilon \). Normally, the residuals are assumed to be drawn independently from the same distribution (white residuals), or if not, then the data and model are filtered, or prewhitened, by an estimate of the error covariance (see later). This means that \( C_\epsilon = \sigma^2 I \), corresponding to an \( N \times N \) identity matrix (I) scaled by a single variance term \( \sigma^2 \).

Assuming white residuals, the parameters can be estimated by minimizing the sum of squares of the residuals, to give the so-called ordinary least squares (OLS) estimates, \( \hat{\beta} \). The planned comparisons we want to test with our experiment are a linear combination of these parameter estimates, specified by a \( 1 \times P \) contrast vector, \( c \). For example, \( c = [1 \quad -1] \) would test whether the parameter estimate for the first of two regressors is greater than the second. Significance can be assessed by a \( T \)-statistic, defined by

\[ T(\text{df}) = \frac{\hat{\beta}^T c}{\sqrt{c^T (X^T X)^{-1} c \sigma^2}} \]  \[\text{[2]}\]

where \( \sigma^2 \) is the error variance estimated by \( \epsilon^T \epsilon / \text{df} \) (where \( \epsilon \) denotes the transpose of a matrix) and the degrees of freedom, \( \text{df} \), are defined by \( N - \text{rank}(X) \). The probability, \( p \), of getting a value of \( T \) or greater under the null hypothesis that \( \hat{\beta} = 0 \), given the \( \text{df} \), can then be calculated from Student’s \( T \)-distribution, and the null hypothesis rejected if, for example, \( p < 0.05 \).

We are now in the position to define the efficiency of a contrast, \( e \), as

\[ e = 1 - \frac{c(X^T X)^{-1} c}{c^T c} \]  \[\text{[3]}\]

which can be seen as inversely related to the denominator of the \( T \)-statistic in eqn [2]. Thus, if we increase \( e \), we also increase \( T \). (For multiple contrasts, where \( e \) is an \( M \times P \) matrix of \( M \) contrasts, such as an \( F \)-contrast, we can define the average efficiency as \( 1/\text{trace}\{c(X^T X)^{-1} c^T\} \}.)

Note that the scaling of \( e \) is arbitrary (depending on the scaling of the contrast, scaling of regressors, and number of scans), so the precise relationship between \( e \) and \( T \) is best assumed only to be monotonic. Note also that this statement assumes that the estimate of the error variance (\( \hat{\sigma}^2 \)) is independent of the design (\( X \)), which may not always be true (see later). Given these assumptions, and that the contrasts are specified a priori, then to maximize the efficiency of our design, we simply need to vary \( X \). We now consider how \( X \) is defined for fMRI.

**HRF Convolution**

We can start by assuming that stimuli elicit brief bursts of neural activity, or events, which are modeled by delta functions every time a stimulus is presented. Then, for the \( j \)th of \( N_j \) event types (conditions), the neural activity over time, or neural time course, \( u_j(t) \), can be expressed as

\[ u_j(t) = \sum_{i=1}^{i=N_j(j)} \delta(t - T_{ji}) \]

where \( T_{ji} \) is a vector of \( i=1 \ldots N_j(j) \) onset times and \( \delta \) is the Dirac delta function. With fMRI, we do not measure neural activity directly, but rather the delayed and dispersed BOLD impulse response, \( b(\tau) \), where \( \tau \) indexes poststimulus time (e.g., from 0 to 30 s). Given that \( b(\tau) \) may vary across voxels (and individuals), it can be modeled by linear combination of \( N_k \) hemodynamic response functions (HRFs), \( h_k(\tau) \):

\[ b_i(\tau) = \sum_{k=1}^{k=N_k} \beta_{ik} h_k(\tau) \]

where \( \beta_{ik} \) are the parameters to be estimated for each HRF and condition (and voxel).
Assuming that BOLD responses summate linearly (though see later), the predicted BOLD time course over the experiment, $x(t)$, can then be expressed as the convolution of the neural time courses by the HRFs:

$$x(t) = \sum_{j=1}^{N_j} u_j(t) \otimes b_j(t) = \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} \sum_{i=1}^{N_i} b_{ij} h_k(t - T_{ij})$$ \[4\]

resulting in a linearly separable equation that can be represented by a design matrix $X$ with $P = N_j N_k$ columns.

At one extreme, we can assume a fixed shape for the BOLD response by using a single canonical HRF (i.e., $N_k = 1$). At the other extreme, we can make no assumptions about the shape of the BOLD response (up to a certain frequency limit) by using a so-called finite impulse response (FIR) set (see Figure 1; for multiple basis functions, the contrasts become $c \otimes I_{N_k}$, where $c$ is a contrast across the $N_j$ event types and $I_{N_k}$ is an $N_k \times N_k$ identity matrix for the $N_k$ basis functions). Normally, one is only interested in the magnitude of a BOLD response, in which case a single canonical HRF is sufficient to estimate efficiency a priori (by assuming that a canonical HRF is a sufficient approximation on average across voxels and individuals). If however one is interested in estimating the shape of the BOLD impulse response, then a more general set such as an FIR is necessary. Using a canonical HRF would correspond to what Liu, Frank, Wong, and Buxton (2001) called detection power, while using an FIR would correspond to what they called estimation efficiency. This is important because the choice of HRF affects the optimal experimental design (see later).

### Filtering

So far, we have considered definition of the signal, $x(t)$, but the other factor that affects the $T$-statistic in eqn [2] is the noise variance, $\sigma^2$. fMRI is known to have a preponderance of low-frequency noise, caused, for example, by scanner drift and by biorhythms (e.g., pulse and respiration) that are aliased by slower sample rates (1/TR). A common strategy therefore is to high-pass filter the data. An example matrix, $F$, for implementing high-pass filtering within the GLM using a discrete cosine transform (DCT) set is shown in Figure 1. The reduction in noise will improve sensitivity, as long as the filtering does not remove excessive signal too. Heuristics suggest that an...
approximate inflection in the noise power spectrum typically occurs at around 1/120 s, which is why it is inadvisable to have designs with changes in signal slower than this (e.g., alternating blocks of more than 60 s; see later).

High-pass filtering also helps render the noise white, that is, constant across frequencies, though there is often still temporal autocorrelation (color) in the residuals. A common strategy to deal with this is to hyperparameterize the error covariance matrix \( C_e \) using an AR(\( p \)) model, which can be estimated using ReML (Friston et al., 2002). Once estimated, \( C_e \) can be inverted in order to prewhiten the data and model, which therefore also affects efficiency. Mathematically, both high-pass filtering and prewhitening can be implemented by multiplying data and model by a single filter matrix, \( K \), such that efficiency becomes

\[
1/(C^{-1} K^T (K C^{-1} K)^{-1} C^T) \tag{5}
\]

### Parameterizing Experimental Designs

For events, the neural time course is determined by \( T_{ji} \) in eqn [4], which itself can be captured by two parameters: (1) the minimal time between events, or \( \text{minimal stimulus-onset asynchrony}, \text{SOA}_{\text{min}} \), and (2) a transition table, which can be defined by two matrices, an \( N_J \times N_T \) matrix, \( T(\rho) \), describing the history of the previous \( N_J \) event types, of which there are \( N_p \) possible sequences (in the extreme case \( N_p = N_J^{N_T} \), though it can be smaller), and an \( N_J \times N_T \) matrix, \( T(\gamma) \), describing the probability of the next event being one of the \( N_J \) event types, given each of those possible previous sequences. So for a fully randomized design, where the probability of each event is equal and independent of previous events (\( N_T = 1 \)), \( T(\rho) \) would be \( N_J \times 1 \) matrix and \( T(\gamma) \) would be \( N_J \times N_J \) matrix with values of \( 1/N_J \) for each event type. So for \( j = 1, 2 \) event types,

\[
T(\rho) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, T(\gamma) = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}
\]

This implies that there is an equal chance of event type 1 being followed by event type 1 as there is for it being followed by event type 2 and likewise for what follows event type 2. Specifying a design in terms of probabilistic transition matrices allows one to treat the design matrix as a random variable and derive the expected efficiency by averaging over all possible design matrices (see Friston, Zarahn, Josephs, Henson, & Dale, 1999, for details). In other words, one can express design efficiency in terms of the probabilistic contingencies entailed by the design matrix.

### Randomized Designs

For a randomized design with two event types, we can plot the efficiency against \( \text{SOA}_{\text{min}} \) for each of 2 contrasts, \( c = [1 \ -1] \), the differential effect of event types 1 and 2, and \( c = [1 \ 1] \), the common effect of event types 1 and 2 versus the interstimulus baseline. For a canonical HRF, the efficiency of these two contrasts is plotted against \( \text{SOA}_{\text{min}} \) in Figure 2(a). As can be seen, the optimal SOA for the common effect is around 18 s, whereas the optimal SOA for the differential effect increases exponentially as \( \text{SOA}_{\text{min}} \) decreases.

The basic reason for these results is that higher efficiency corresponds to greater variability of the signal over time (where the signal is a function of the contrast and regressors). At short SOAs, the sluggish (low-pass) nature of the HRF means that when we do not distinguish the two event types (by using a contrast for their common effect), the BOLD responses for successive events summate to give a small oscillation around a raised baseline (leftmost inset in Figure 2(a)), that is, low signal variance. The random ordering of the two event types means that for the differential effect, however, there is a large variance in signal. At longer SOAs, there is time for the BOLD response to return to baseline between events, so signal variance (efficiency) is increased for the common effect, but the variance for the differential effect decreases relative to short SOAs (rightmost inset in Figure 2(a)). For further explanations of this behavior of efficiency, for example, in terms of signal processing or correlations between regressors, see http://imaging.mrc-cbu.cam.ac.uk/imaging/DesignEfficiency. The important point of Figure 2(a) is that if one is interested in the difference between two randomly ordered event types, then a shorter SOA is generally better, though the price one pays is reduced efficiency to detect the common effect of both versus baseline. In reality though, this increasing efficiency with decreasing \( \text{SOA}_{\text{min}} \) (for a differential effect) cannot occur indefinitely, because at some point, there will be saturation of the neural and/or hemodynamic response to stimuli that are too close together in time, an example of nonlinear behavior that is ignored under the linear superposition assumptions made so far (see later).

One can improve efficiency for the common effect versus baseline at short SOAs by ensuring that the probabilities in \( T(\gamma) \) do not sum to 1 (across columns). This means that for some \( \text{SOA}_{\text{min}} \), no event occurs (sometimes called a null-event or fixation trial, Dale, 1999). This effectively produces a stochastic distribution of SOAs, with an exponentially decreasing probability of each SOA. The efficiency for the common effect now also increases as \( \text{SOA}_{\text{min}} \) decreases, at a cost to the efficiency of the differential effect (Figure 2(b)). These types of design are suitable for an FIR basis set (insets in Figure 2(b)) because a stochastic distribution of SOAs allows the BOLD response shape to be estimated, particularly with short \( \text{SOA}_{\text{min}} \). (Note that efficiency is not directly comparable across canonical and FIR HRFs, since it depends on the scaling of the basis functions, and an FIR will also entail a reduction in the df, which will affect the final \( T \)-statistic in eqn [2].)

### Blocked Designs

Events of the same type can be blocked into short sequences, which can increase the detection power relative to a randomized design. For a blocked design with two event types, there would be \( N_L \) events per block; for example, for blocks of three events,
The magenta dashed line in Figure 2(c) shows efficiency for the differential effect between two event types as a function of block length using a canonical HRF when SOA\textsubscript{min} is 2 s and the high-pass cutoff is 120 s. Short blocks (in the extreme case, alternating event types when block length is 1) have low efficiency, for the same reason that the main effect is inefficient at short SOAs: any variance in neural activity is smoothed out by the HRF. As in Figure 2(a), efficiency is maximal for block lengths around 18 s (since 1/18 Hz is close to the highest passband of the canonical HRF filter; Josephs & Henson, 1999), but for block lengths of 30 s or more, efficiency plummets again because of the high-pass filter: for such long blocks, most of the signal variance (particularly that at the fundamental frequency of the block alternation) is low enough to be removed by the high-pass filter.

\[
TM^{(p)} = \begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 2 \\
1 & 2 & 2 \\
2 & 2 & 2 \\
2 & 2 & 1 \\
2 & 1 & 1
\end{bmatrix}
\]

\[
TM^{(a)} = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
0 & 1 \\
0 & 1 \\
1 & 0 \\
1 & 0
\end{bmatrix}
\]
pass filter, which is appropriate since it is likely to be masked by fMRI noise anyway. In short, one does not want fMRI designs where the signal changes too fast (since it will be attenuated by the sluggish BOLD response) or where it changes too slow (since it will be removed by any high-pass filtering or swamped by noise).

The blue dotted line in Figure 2(c) shows efficiency as a function of block length when using an FIR basis set rather than canonical HRF. In this case, efficiency is low, at least until the block length exceeds the duration of the assumed BOLD response (30 s here). This is because of the high covariance between regressors for each basis function (rightmost inset in Figure 2(c)), as explained later. This illustrates that, while blocked designs are efficient for estimating the amplitude of a known BOLD response, they are not efficient for estimating the unknown shape of a BOLD response (Liu et al., 2001).

Unique Effects

In addition to the differential and common effects for two event types, one might be interested in the unique effect of each event type, having adjusted for effects of the other event type, which corresponds to contrasts of \([1 \ 0]\) or \([0 \ 1]\). A common example is when one type of event must always follow another type, such as a motor act that is contingent on a stimulus (e.g., a working memory trial). The dashed line in Figure 2(d) shows that, for SOAs below 9 s, efficiency is low for a \([0 \ 1]\) contrast in such an alternating design. To improve efficiency, one needs to reduce the correlation between the two event types (since the term X'X in eqn [3] relates to the covariance of the regressors in the design matrix). One way to do this is to randomize the SOA between stimuli and motor acts; another is to only require a motor act on a random fraction, say 50%, of trials, as shown in the solid line in Figure 2(d). Although this entails fewer motor events in total, the efficiency for separating the BOLD response to the motor act from that to the stimulus is increased at short SOAs.

A similar issue arises in so-called mixed designs, in which one wishes to separate a sustained, or state, effect from a transient, or item, effect (Chavla, Rees, & Friston, 1999). This requires blocks of events, in which the SOA is varied within each block so as to reduce the correlation between the (epoch) regressor modeling the state effect and the (event) regressor modeling the item effect. The downside of such designs is that this requirement to decorrelate the two regressors, in order to estimate both unique effects reasonably efficiently, considerably reduces the efficiency for estimating either effect alone, relative to designs with only epochs, or only events.

Optimizing Designs

The examples in Figure 2 represent just a subspace of possible designs, chosen to help illustrate some of the properties of eqn [5]. Other formal explorations of design space can be found in, for example, Dale (1999), Josephs and Henson (1999), Friston et al. (1999), and Hagberg, Zito, Patria, and Sanes (2001). There are automated ways of maximizing efficiency by searching through possible designs, using, for example, genetic algorithms (Wager & Nichols, 2003). Additional constraints are often needed however, such as limits on runs of the same event type; otherwise, an optimization scheme is likely to converge on a blocked design, which is always most efficient (for detection power) from the fMRI perspective, but may not be appropriate from the psychological perspective (e.g., if the presence of structure in the sequence of events affects brain activity). An interesting class of pseudorandomized design that has optimal estimation efficiency is an m-sequence (Buracas & Boynton, 2002). This is a deterministic sequence that presents all combinations of event histories up to \(N_{ij} = m\) (i.e., has a large, deterministic transition matrix) but that is nonetheless effectively unpredictable to participants. Such sequences have been computed for \(N_j = 2, 3, \) and 5 event types but require a fixed number \((N_j^N_{ij} - 1)\) of events in total.

Nonlinearities and Assumptions

The main assumption made in the examples mentioned earlier is that the brain’s response to successive events is a linear superposition of responses that would occur to each event on its own. Of course, this is unlikely in practice, particularly for short SOAs, where there is likely to be saturation of neural and/or hemodynamic responses. Such saturation has been demonstrated empirically and can be modeled with the GLM by using Volterra kernels (Friston, Josephs, Rees, & Turner, 1998). Once fit, these kernels can be used to adjust predictions for efficiency. For kernels fit to auditory cortex responses to auditory stimuli of varying SOAs, the negative impact of saturation on efficiency still only became appreciable for SOAs below 2 s: that is, there was still an advantage of short SOAs down to 2 s in randomized designs.

A second assumption is that changing the design \(X \) in eqn [5] does not affect the estimation of the error \((\hat{\sigma}^2\) in eqn [2] or even \(K \) in eqn [5]). Different SOAs (even across trials within a design) may entail differences in HRF shape, or different nonlinearities, resulting in inaccurate model fits and therefore different residuals. Indeed, when comparing blocked and randomized designs, differences in the error estimate have been shown empirically (Mechelli, Price, Henson, & Friston, 2003). Nonetheless, without such a priori knowledge about the noise, one can only rely on the general heuristics about maximizing the signal outlined earlier.

Summary

Efficiency is a well-defined mathematical property of the GLM, and under the linear assumptions of a convolution model for the BOLD response, efficiency can be optimized for a priori contrasts of the conditions of an fMRI experiment by selecting the optimal SOA and stimulus transition table.

Acknowledgments

This work was supported by the UK Medical Research Council (MC_US_A060_5PR10).
See also: INTRODUCTION TO METHODS AND MODELING: Contrasts and Inferences; Convolution Models for FMRI; The General Linear Model.

References


Relevant Websites


http://imaging.mrc-cbu.cam.ac.uk/imaging/DesignEfficiency – General advice about how to optimise an fMRI experiment.


http://surfer.nmr.mgh.harvard.edu/optseq/ – Tool for optimising randomised designs.