Stochastic Designs in Event-Related fMRI

K. J. Friston,* E. Zarahn,† O. Josephs,* R. N. A. Henson,* and A. M. Dale‡

*The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, United Kingdom;
†Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania 19104; and ‡Massachusetts General Hospital NMR Center,
149 13th Street, Charlestown, Massachusetts 02129

This article considers the efficiency of event-related fMRI designs in terms of the optimum temporal pattern of stimulus or trial presentations. The distinction between "stochastic" and "deterministic" is used to distinguish between designs that are specified in terms of the probability that an event will occur at a series of time points (stochastic) and those in which events always occur at prespecified time (deterministic). Stochastic designs may be "stationary," in which the probability is constant, or nonstationary, in which the probabilities change with time. All these designs can be parameterized in terms of a vector of occurrence probabilities and a prototypic design matrix that embodies constraints (such as the minimum stimulus onset asynchrony) and the model of hemodynamic responses. A simple function of these parameters is presented and used to compare the relative efficiency of different designs. Designs with slow modulation of occurrence probabilities are generally more efficient than stationary designs. Interestingly the most efficient design is a conventional block design. A critical point, made in this article, is that the most efficient design for one effect may not be the most efficient for another. This is particularly important when considering evoked responses and the differences among responses. The most efficient designs for evoked responses, as opposed to differential responses, require trial-free periods during which baseline levels can be attained. In the context of stochastic, rapid-presentation designs this is equivalent to the inclusion of "null events." © 1999 Academic Press

Key Words: functional neuroimaging; fMRI; stochastic; event-related; experimental design.

INTRODUCTION

A current issue in event-related fMRI is the choice of interstimulus interval or more precisely stimulus onset asynchrony (SOA). The SOA, or the distribution of SOAs, is a critical factor in experimental design and is chosen, subject to some constraints, to maximize the efficiency of response estimation. The constraints on the SOA clearly depend upon the nature of the experiment but are generally satisfied when the SOA is small and derives from a random distribution. Rapid presentation rates allow for the maintenance of a particular cognitive or attentional set, decrease the latitude that the subject has for engaging alternative strategies, or incidental processing, and allows the integration of event-related paradigms using fMRI and electrophysiology. Random SOAs ensure that preparatory or anticipatory factors do not confound event-related responses and ensure a uniform context in which events are presented. These constraints speak to the well-documented advantages of event-related fMRI over conventional blocked designs (e.g., Buckner *et al.,* 1996; Dale and Buckner, 1997; Josephs *et al.,* 1997; Zarahn *et al.,* 1997; Rosen *et al.,* 1998).

There has been a growing interest in the choice of SOAs that has been focused by the emergence of a dichotomy in event-related fMRI using multiple trial/event types. The first approach involves the use of very short SOAs (e.g., Dale and Buckner, 1997; Clark *et al.,* 1998; Burock *et al.,* 1998). In these paradigms SOAs of a second or less are commonplace. The other approach used relatively long SOAs of several seconds or more (e.g., Friston *et al.,* 1998b). The aim of this article is to compare the relative efficiency of these approaches. To do this a framework that accommodates stochastic or nonstochastic event or epoch-related designs is introduced. The sections below consider (i) how to assess the sensitivity or efficiency of a design, (ii) how different sorts of designs can be parameterized in a common framework to facilitate comparisons, and (iii) how the efficiency of a particular design depends on whether one is looking for evoked responses per se or differences in evoked responses.

SENSITIVITY AND EFFICIENCY

In Friston *et al.* (1994), we characterized the form of the hemodynamic impulse response function (HRF) using least-squares deconvolution and a linear time invariant (LTI) model, where evoked neuronal re-

sponses are convolved with the HRF to give the measured hemodynamic response. This convolution framework is the cornerstone for making statistical inferences about activations in fMRI using the general linear model. In Friston *et al.* (1995), we introduced the notion of temporal basis functions to model evoked responses in fMRI and applied it to event-related responses in Josephs *et al.* (1997). The advantage of using temporal basis functions (as opposed to an assumed form for the HRF) is that one can model voxel-specific forms for hemodynamic responses and differences among responses to different sorts of events or trials (Friston *et al.,* 1998a). Temporal basis functions allow for a graceful transition between the efficiency of fixed-form response models (like a conventional box-car) and the flexibility of finite impulse response (FIR) characterizations that embody no constraints on the modeled response.

Irrespective of the form of the HRF, its extended time course effectively smooths neuronal responses, elicited in rapid succession with a fixed SOA, to render them flat. This puts a constraint on the minimum SOA that can be usefully employed in fMRI that depends upon the experimental design. By using a suitable distribution of SOAs one can make some events sufficiently sparse so that the hemodynamic response can be estimated. The objective of experimental design is to make this estimation as efficient as possible. The efficiency of an estimator is inversely related to its variance. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \,, \qquad (1)$$

where $\mathbf{Y}$ is a column vector corresponding to the observed hemodynamic response at any voxel, $\mathbf{X}$ is a design matrix modeling the expected hemodynamic responses, β is a vector of parameters or coefficients pertaining to the explanatory variables in $\mathbf{X}$, and $\mathbf{e}$ is a normally distributed error term $\sim N(0, \sigma^2 \mathbf{I})$. Here we assume that the data have been prewhitened if necessary (see Dale *et al.,* in preparation, and Burock *et al.,* in preparation, for a discussion of efficient, unbiased estimation of event-related fMRI responses in the presence of serial correlations). The efficiency of the estimation is inversely related to the (co)variance of the parameter estimates. This (co)variance is a function of, and only of, the design matrix and noise variance σ^2.

$$\text{cov} \,[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}, \qquad (2)$$

where $\mathbf{X}$ embodies all the information about experimental design. From the current point of view it contains information about the SOAs, which event occurs when, and the constraints on the form of the hemodynamic response. In general $\mathbf{X}$ can be created by convolving a set of delta functions, indicating the presence of a particular event, with a small set of basis functions that model the hemodynamic response to those events. A special case of this general approach obtains when the basis functions are delta functions placed at a discrete set of peristimulus times. This special case corresponds to a FIR model and is formally related to "selective averaging" (Dale and Bucker, 1997), which requires stimulus presentation and data acquisition to be synchronized.

Maximizing efficiency means minimizing the (co)variances in Eq. (2), which is equivalent to maximizing the (co)variance of the design matrix itself $\mathbf{X}^T \mathbf{X}$ (cf. the energy ratio used by Zarahn *et al.* submitted for publication). Put simply, maximizing the efficiency of the estimators reduces to maximizing the variance (or power) accounted for by the explanatory variables or compounds thereof that are interesting. If there is only one event type, and a fixed form for the HRF is assumed, then there is only one column in $\mathbf{X}$ and one parameter estimate. In this instance the (co)variance of the parameter estimate reduces to a scalar and this can be used to optimize the distribution of interstimulus intervals. More generally, however, there will be several columns in $\mathbf{X}$ due to the fact that more than one event or trial type has been used in the experiment and/or the hemodynamic response is modeled with two or more basis functions. In this instance we can specify a compound or weighted sum of the parameter estimates for which we want to maximize the efficiency. For example, when using two trial types (each modeled with a single basis function) the differential evoked responses would be estimated using the weights -1 and $+1$ (these are referred to as contrast weights and will be denoted by the column vector $\mathbf{c}$). In general the variance of any estimator of interest, specified as a contrast of parameter estimates is

$$\text{cov} \,[\mathbf{c}^T \hat{\beta}] = \sigma^2 \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c} \qquad (3)$$

and

$$\text{Efficiency} \propto \text{trace}[\mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}]^{-1}.$$

The trace operator allows for cases when multiple contrasts are specified (e.g., when all basis functions are equally interesting such as in selective averaging; Dale *et al.,* in preparation). The critical thing to note, at this stage, is that the best design matrix for one contrast, in terms of efficiency, may not be the best for another contrast. In the example above the best distributions of SOAs for detecting an event-related response itself ($\mathbf{c} = [1 \, 0]^T$) may not be the best for detecting the difference between two events ($\mathbf{c} = [1 \, -1]^T$).

PARAMETERIZING AND COMPARING
fMRI DESIGNS

fMRI designs can vary over a large number of parameters. In order to compare the efficiency of different designs it is useful to have some common framework that accommodates them all. The efficiency can then be examined in relation to the parameters of the design. A general taxonomy of designs might be the following: Any design can be stochastic or deterministic. In stochastic designs (Heid *et al.*, 1997) one needs to specify the probabilities of an event occurring at all times those events could occur. In deterministic designs the occurrence probability is unity and the design is completely specified by the times of stimulus presentation or trials.

The idea of stochastic designs can be made more accessible by relating them to designs that we typically use. For example a stochastic design with one event allows one to ask whether the response to that event differs from zero. Here the event of interest can be presented at a number of regularly spaced time points during the scanning period. Whether it is actually presented at these times or not is determined probabilistically, usually by some random number generator. The nonoccurrence at certain time points can be considered as "null events" and corresponds to the inclusion of a baseline in a more traditional framework. The case of two event types can be treated as a basic subtraction design where one can ask if the differential response is significant. Generally stochastic designs subsume any design that calls upon a random process for its specification.

The distinction between stochastic and deterministic designs pertains to how a particular realization or stimulus sequence is created. The efficiency afforded by a particular event sequence is a function of the event sequence itself and not of the process generating the sequence (i.e., deterministic or stochastic). In fact, a stochastic process is highly unlikely to generate an optimal sequence (in the sense of maximizing efficiency) for any experiment of finite duration. With a stochastic process, the design matrix $\mathbf{X}$ and associated efficiency are random variables, where the efficiency converges only in the limit of infinite length sequences. However, for finite length sequences the expected efficiency, over an infinite number of realizations of $\mathbf{X}$, is easily computed and this is the tenet of the approach adopted by Zarahn and colleagues and extended here to deal with nonstationary stochastic designs (see below). The motivation for this approach is to compare various stochastic designs and compare their expected efficiency in relation to equivalent deterministic designs.

Under deterministic designs the events can occur at fixed intervals (e.g., fixed-interval event-related design;

Friston *et al.*, 1998b) or the SOAs can vary over time (e.g., variable-interval event-related and conventional block or epoch-related designs). Under stochastic designs the temporal support (i.e., times at which a trials could occur) must be specified. This could of course be the entire scanning period but generally comprises some discrete time points. One parameterization that has emerged in event-related fMRI (Dale and Buckner, 1997; Zarahn *et al.*, submitted for publication) is in terms of a minimum SOA_{min} and the probability P^i that the ith event type will occur every SOA_{min}. Clearly the sum of P^i over all event types i is unity. The corresponding SOAs have a geometric distribution. The parameterization of these designs is in terms of the scalars SOA_{min} and P^i.

Here we adopt a more general formulation that accommodates both stochastic and deterministic designs. This model is based on that employed by Zarahn *et al.*, (submitted for publication) to look at stationary stochastic designs. In the present model the probability of any event occurring is specified at each time it could occur. Here $\mathbf{P}^i$ is a vector with an element for every SOA_{min}. This formulation engenders the distinction between stationary stochastic designs, in which the occurrence probabilities are constant and nonstationary, and "modulated" stochastic designs, in which they change over time. For deterministic designs the elements of $\mathbf{P}^i$ are 0 or 1, with the presence of a 1 denoting the occurrence of an event. An example of $\mathbf{P}^i$ might be the box cars used in conventional block designs. Stochastic designs of the sort proposed by Dale and Buckner (1997) correspond to a vector of identical values $\mathbf{P}^i = [P^i \ P^i \ P^i \ \ldots P^i]^T$ and are therefore stationary in nature. Stochastic designs with temporal modulation of occurrence probability (e.g., Elliot *et al.*, submitted for publication) would correspond to probability vectors with time-dependent probabilities varying between 0 and 1 (inclusive).

The advantage of this parameterization is that the (expected) efficiency of any design is easily computed with Eq. (3) using

$$\langle \mathbf{X}^T\mathbf{X} \rangle = \mathbf{P}^T(\mathbf{S}^T\mathbf{S} - \mathbf{D})\mathbf{P} + \sum_m P_m D_{mm} \qquad (4)$$

for one trial type and basis function (see the Appendix for a full description and more general cases). $\langle . \rangle$ denotes expectation or average. Here $\mathbf{S}$ is a prototypic design matrix that embodies information about SOA_{min}, the length of the time-series, and the basis functions employed. $\mathbf{D} = \text{diag}(\mathbf{S}^T\mathbf{S})$. The nice thing about this expression is that by setting the mean of the probabilities $\mathbf{P}$ to a constant, one can compare different deterministic and stochastic designs given the same number of events (or equivalently the same mean SOA). Some

common examples are given in Fig. 1 for an SOA$_{min}$ of 1 s and 32 expected events or trials over a 64-s period (except the fixed deterministic example with 8 events). It can be seen that the least efficient design is a fixed-interval deterministic design (despite the fact that the SOA is roughly optimal for this class), whereas the most efficient is a block design. See Burock *et al.* (1998) for an empirical illustration of the increased efficiency afforded by randomized designs relative to fixed SOA designs.

A slow modulation of occurrence probabilities gives high efficiency while perhaps retaining the advantages of stochastic designs and may represent a very useful compromise between the high efficiency of block designs and the psychological benefits and latitude of stochastic designs. The fact that slow modulations, as opposed to high-frequency modulations, are more efficient is consistent with the fact that the eigenvectors of ($\mathbf{S}^T\mathbf{S} - \mathbf{D}$), with large positive eigenvalues, are low frequency in nature for this SOA$_{min}$ and basis set (a canonical HRF). It should be noted that the relative efficiencies will change with changes in design parameters such as SOA$_{min}$, basis set, and the chosen contrast. In our experience there is always some nonstationary stochastic design that is substantially more sensitive than the equivalent stationary design.

The optimization of $\mathbf{P}$, in terms of efficiency, is a nonlinear problem that is compounded by the fact that the elements of $\mathbf{P}$ must lie between 0 and 1. Nonlinear optimization of $\mathbf{q}$, where $\mathbf{P} = (\sin(\mathbf{q}) + 1)/2$ (as implemented in MatLab (MathWorks Inc., Natick, MA)) suggests that the block design is at least a local maximum and may be the global maximum given that we cannot find a more efficient design.

WHAT IS THE MINIMUM SOA FOR A VALID LINEAR APPROXIMATION?

Generally, for any design, the smaller the SOA$_{min}$, the greater the number of trials or events that can be presented and the more efficient that design (there are important exceptions that will be mentioned below). What then is the smallest SOA one could entertain and still rely on predictions about efficiency based on Eqs. (3) and (4)? The lower limit on SOAs is dictated by nonlinear interactions among events, and the evoked responses, when they are very proximate in time. These nonlinearities can be thought of as saturation phenomena or "refractoriness" at a neuronal or hemodynamic level. We have already presented a fairly comprehensive analysis of these effects in the context of single word presentation in Friston *et al.* (1998b) that will be briefly reprised here: Conventional fMRI models use linear time invariant models (e.g., Friston *et al.,* 1994; Boynton *et al.,* 1997). However, there is evidence to suggest that nonlinear effects may predominate at very

short SOAs (e.g., Vazquez and Noll, 1998). In Friston *et al.* (1998b) we used a nonlinear convolution model (Volterra series) to estimate high-order convolution kernels in a way that is directly analogous to the estimation of event-related responses in terms of a linear or first-order kernel (i.e., the HRF). Having estimated these kernels, we were then able to examine the responses to stimuli over continuous ranges of SOAs. The results of a typical analysis are given in Fig. 2. This represents the average response, integrated over a 32-s train of stimuli, as a function of SOA within that train. The data were based on kernel estimates from a voxel in the left posterior temporal region of a single male subject obtained during the presentation of single words at a variety of rates. The task was a simple passive listening task. The solid line represents the estimated response and shows a clear maximum at just less than 1 s. The dots represent estimates based on empirical data from the same experiment. The broken line shows the expected response in the absence of nonlinear effects (i.e., that predicted by setting the second-order kernels to zero). It is clear that nonlinearities become important at around 2 s, leading to an actual diminution of the integrated response at subsecond SOAs. This effect probably corresponds to a hemodynamic, as opposed to a neuronal, refractoriness because it was not observed when the same experiment, with the same subject, was replicated with PET. In brief, over the range studied, increasing presentation rate caused linear increases in regional cerebral blood flow (roughly equivalent to integrated synaptic activity). The fMRI signal on the other hand evidenced saturation, suggesting that nonlinearities enter at the translation of neuronal activity into the hemodynamic response as measured with fMRI. This, of course, does not preclude nonlinearities at the neuronal level but does suggest that they are not a sufficient explanation for the nonlinearities observed. The implications of these results are that (i) SOAs should not really fall much below 1 s and (ii) at short SOAs the assumptions of linearity, upon which the analyses in this paper rely, are violated. It should be noted that these data pertain to single word processing in auditory association cortex. More "linear" behaviors may be expressed in primary sensory cortex as suggested by the work of Dale and Buckner (1997). Furthermore Burock *et al.* (1998) have demonstrated the feasibility of using minimum SOAs as low as 500 ms, even when sampling at a lower rate (e.g., TR = 1 s).

DIFFERENT DESIGNS FOR DIFFERENT EFFECTS

In this section we consider why, in some instances, a very short average SOA is best, whereas in others a longer SOA is more appropriate. Here we deal explicitly with multiple trial types and define the trial onset
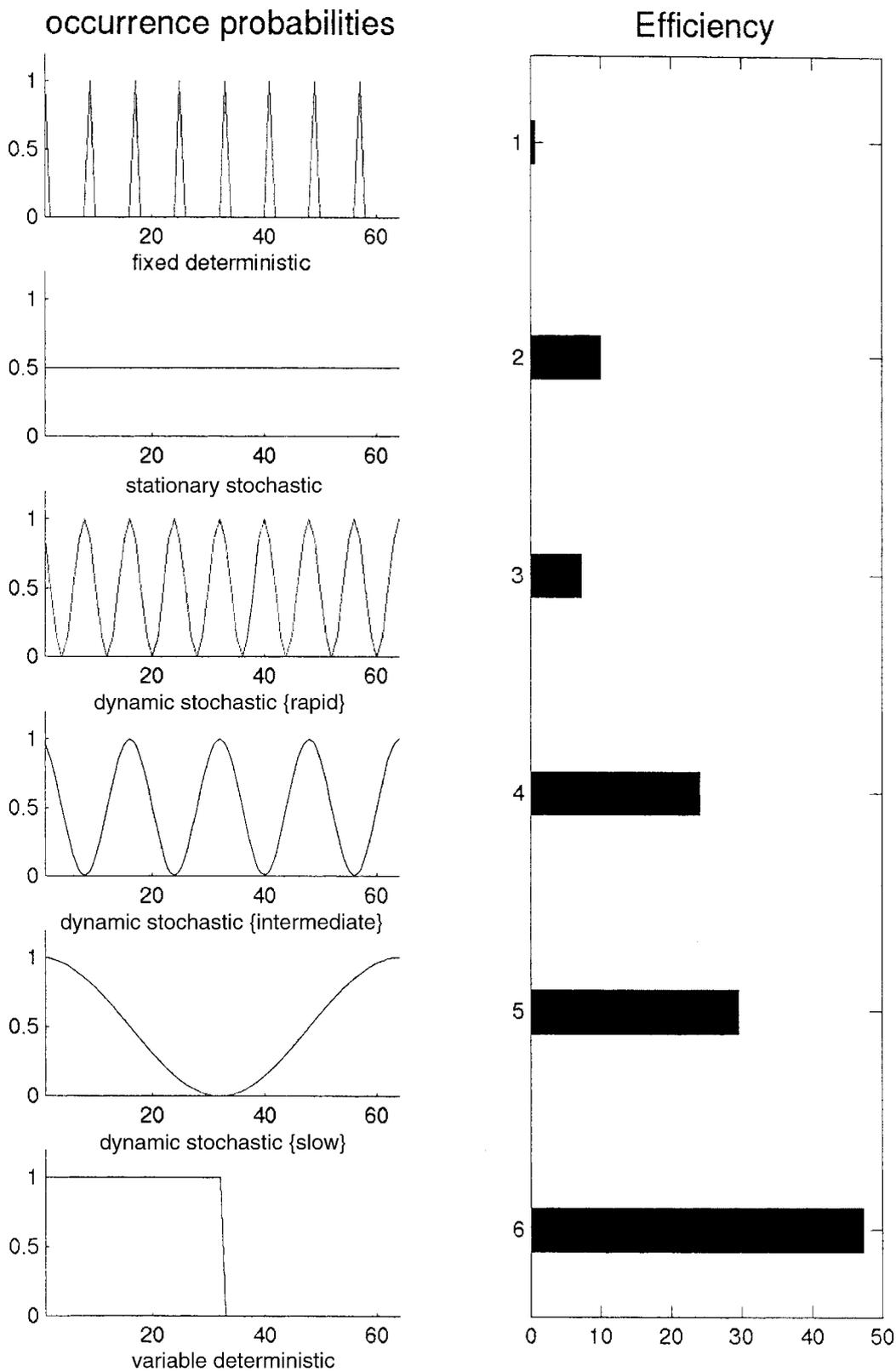
FIG. 1. A comparison of some common (and some untried) designs. The left-hand column is a graphical representation of the occurrence probabilities expressed as a function of time (seconds). The efficiency of each design is shown along the ordinate assuming an $\text{SOA}_{\min}$ of 1 s, a time-series of 64 s, and the first of the basis functions shown in Fig. 4. The expected number of events (i.e., the mean value of **P**) was 0.5 in all cases (apart from the first), corresponding to an expected SOA of 2 s or 32 events.
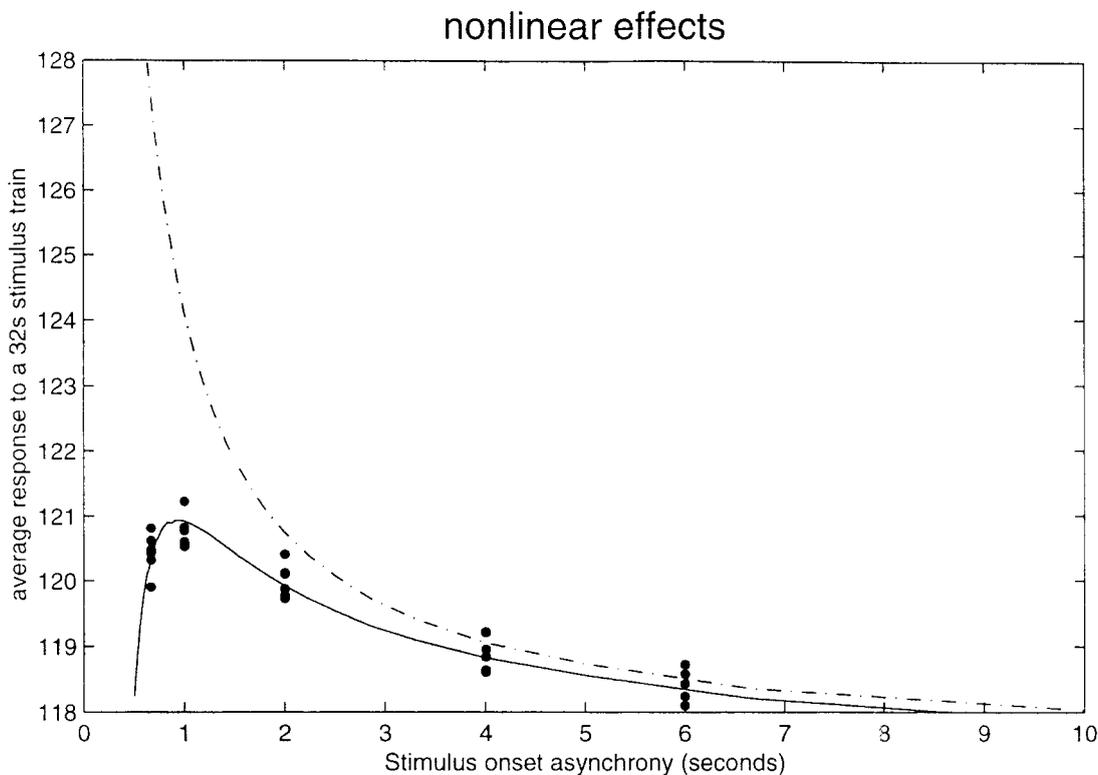
FIG. 2. Plot of integrated response over a 32-s stimulus train as a function of interstimulus interval. Solid line: Estimates based on a nonlinear convolution model and empirical responses to single words presented in a passive listening task to a single subject at different rates. Dots: Empirical averages based on the presentation of actual stimulus trains. Broken line: The responses expected in the absence of second-order effects.

asynchrony (TOA) as the interval between onsets of a particular trial type. The best TOA depends upon the nature of the characterization of evoked responses that is required. For any given event type the associated $\mathbf{P}^i$ determines the average or expected TOA for that event type. This is simply $\mathrm{SOA_{min}}/\overline{\mathbf{P}}^i$, where $\overline{\mathbf{P}}^i$ is the mean probability for the ith trial type. This relationship is illustrated in Fig. 3, where the efficiency of a stationary stochastic design, with one trial type and two basis functions, is plotted as a function of the occurrence probability P and the equivalent TOA (where mean TOA = $\mathrm{SOA_{min}}/P$). In this design the most efficient probability is 0.5, giving a mean TOA of $2 \cdot \mathrm{SOA_{min}}$. Here $\mathrm{SOA_{min}}$ was 1 s and the basis functions corresponded to a canonical HRF (a mixture of two gamma functions that emulate the early peak at around 5 s and a subsequent undershoot) and the derivative of this with respect to latency (Fig. 4).

In designs with multiple trial types the effects of interest may relate to the responses evoked by each event type or they may be differential responses on comparing event types. These two sorts of effects are tested with different compounds or contrasts of the parameter estimates and lead to different optimal TOAs. This distinction can be addressed from the

perspective of a number of designs. In experiments with more than one trial type it will be shown that very short TOAs are appropriate for analyses in which only the difference between responses to one event type and another are of interest. When the responses themselves are important longer TOAs are appropriate and are ensured by the inclusion of null events into the design.

In this section we present results using Eqs. (3) and (4) to show how efficiency varies as a function of occurrence probability and the implications for the optimum mean TOA. In the analyses below the expectation of the parameter estimate (co)variances for two trial types was computed as described in the Appendix using the temporal basis functions in Fig. 4. The ensuing parameter estimates correspond to estimates of the amplitude and latency of the evoked response for both trial types. We present results for inferences about the amplitude of the responses (identical results obtain for latency and differential latency estimates provided the basis functions employed are orthonormal).

Stationary Stochastic Design with $P = 0.5$

In this design an event occurs every $\mathrm{SOA_{min}}$ that may be either trial type 1 or 2. In this example we take the
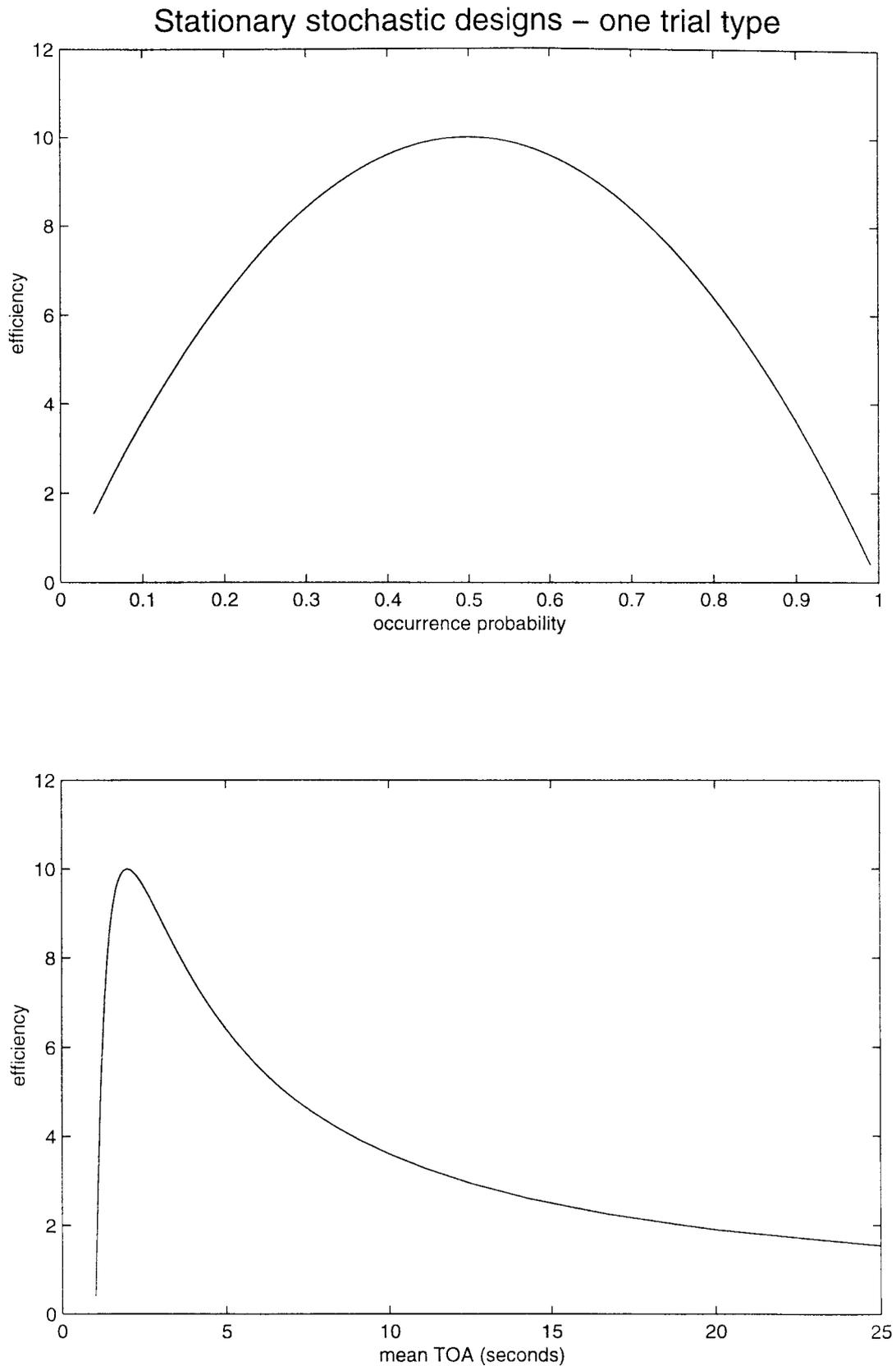
FIG. 3. Efficiency as a function of occurrence probability for stationary stochastic designs. (Top) Expressed directly in terms of occurrence probability. (Bottom) Expressed in terms of the equivalent expected trial onset asynchrony (TOA). These results, and those in the remaining figures, are for 64-s time-series.
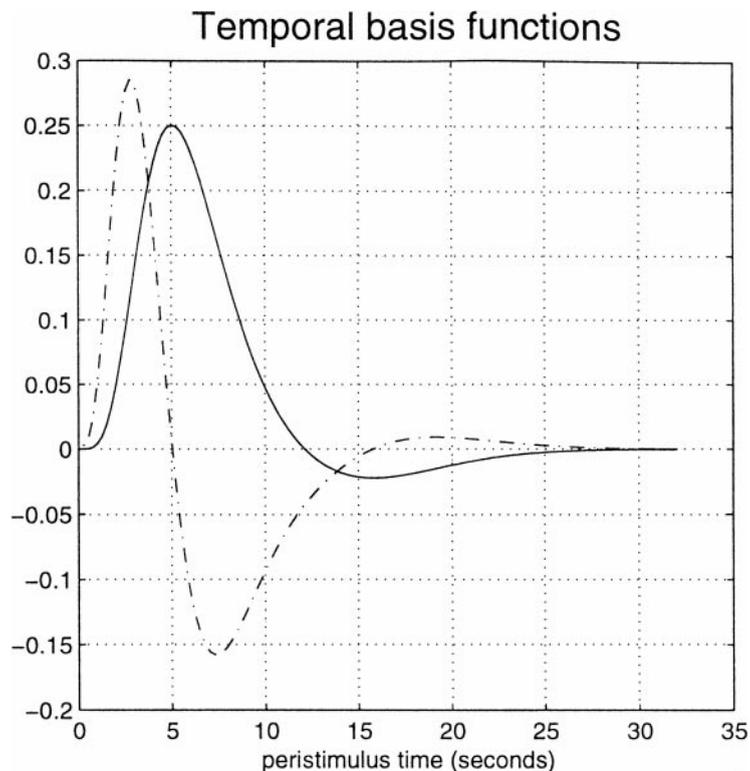
FIG. 4. Basis functions used in constructing the design matrix expectations (see Appendix). These are a mixture of two gamma functions (solid line) and the temporal derivative (broken line).

stationary probabilities of the two types to be the same, $P^1 = P^2 = 0.5$. In this example the SOA_{min} was varied between 2 and 18 s. Figure 5 shows the resulting changes in efficiency for a contrast testing for the amplitude of response evoked by the first event $\mathbf{c} = [1\ 0\ 0\ 0]^T$ (solid line) and differential responses $\mathbf{c} = [1\ 0\ -1\ 0]^T$ (broken line). It is immediately apparent that although the most efficient SOA_{min} for differential responses is very small, longer SOAs of around 16 s are necessary to estimate the responses themselves. This is a crucial observation that speaks to the difference between short and long SOA designs. Although rapid presentation rates are very efficient for detecting differences, they are incapable of estimating the form or degree of evoked response per se or do so with negligible efficiency unless there are at some periods that provide a baseline reference. There is a problem here though in that using long TOAs generally leads to inefficient designs. A more powerful approach (adopted in Buckner *et al.,* 1998; Wagner *et al.,* 1998; Burock *et al.,* 1998) is to explicitly include null events as an extra event type and estimate the evoked response in relation to those null events:

Stationary Stochastic Designs with Null Events

Above we showed that relatively long TOAs are required to estimate responses (as opposed to differen-

tial responses) when an event occurred every SOA_{min}. However by reducing P to less than 0.5 there will be some trials on which no event occurs. These null events now provide a baseline against which the response to either trial type 1 or 2 can be estimated even using a very small SOA_{min}. Figure 6 shows this for a fixed SOA_{min} of 1 s. Here the most efficient occurrence probability for detecting differences is, as one might expect, 0.5. However, the best probability for detecting the responses themselves is about 0.3. This gives a mean TOA for each event type of just over 3 s.

In the case of null events the SOA_{min} was fixed (at 1 s) and P was varied. In the preceding example P was fixed (at 0.5) and SOA_{min} was varied. Clearly these are just two variations on the general stationary stochastic two-trial design, defined in terms of SOA_{min} and P. Figure 7 shows the efficiency as a function of these two parameters and demonstrates that ensuring the inclusion of baseline periods with null events supervenes over increasing the SOA_{min}.

In summary by treating the baseline or control condition as any other condition (i.e., by including null events, with equal probability, as other event types), the efficiency in estimating the evoked response is the same as that for any difference between nonnull trials. Thus, by making the probability of null events and all other events equal to $1/(N + 1)$, where N is the number

Optimum SOA for two trial types
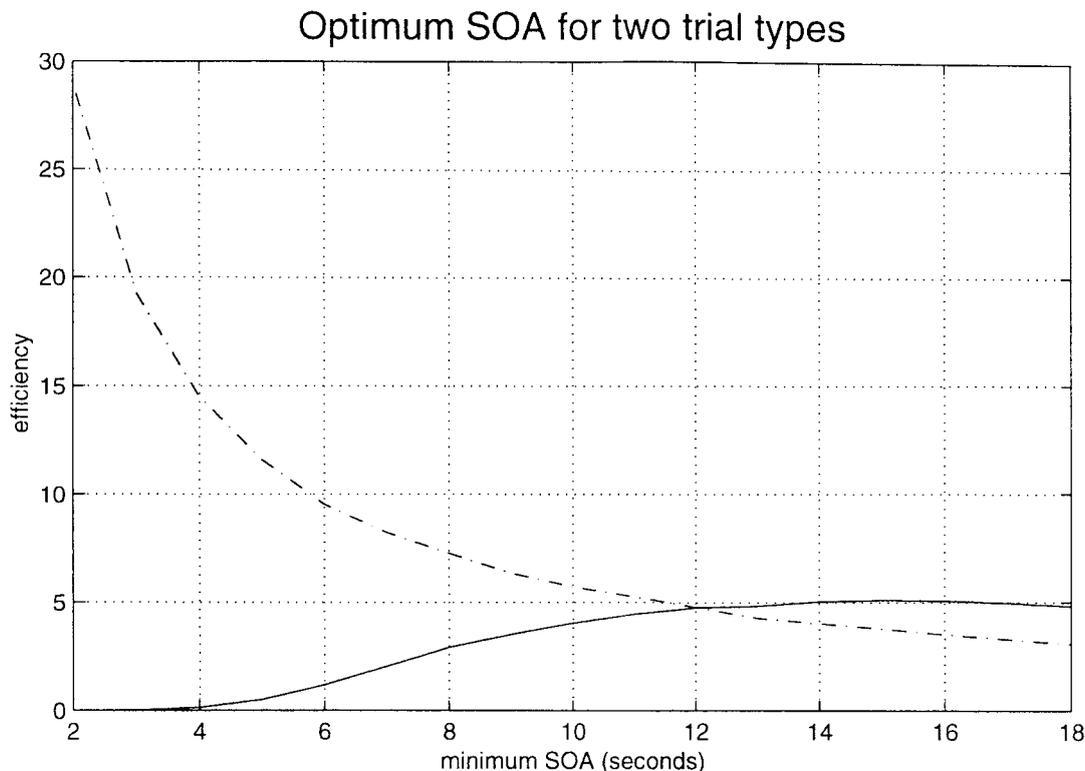
FIG. 5. Efficiency in a stationary stochastic design with two event types each presented with a probability of 0.5 every SOA_{min}, as a function of SOA_{min}. Solid line: Efficiency of estimates of the response to a single trial type. Broken line: Efficiency of estimators of differential responses. Here the time-series is 256 s.

of event types, we obtain a mean TOA of $(N + 1) \cdot$ SOA_{min}. Such designs result in optimal and equivalent efficiency for all comparisons (within stationary stochastic designs). On the other hand, if one were only interested in the differences among the N trial types, a mean TOA of $N \cdot SOA_{min}$ is best. It should be noted that efficiency depends only on the variance of the estimator and sensitivity also depends on the actual scaling of the true underlying effects; therefore one cannot say *a priori* that sensitivity is equivalent for all comparisons.

CONCLUSION

The distinction between stochastic and deterministic has been used above to distinguish between designs that are specified in terms of the probability that an event will occur at specified times and designs where events always occur at prespecified times (clearly deterministic designs are the limiting case of stochastic designs, where all the probabilities are zero or one). Stochastic designs may be stationary, where the probability is constant, or nonstationary, where occurrence probabilities change with time. All these designs can be parameterized in terms of a vector of occurrence probabilities **P** and a prototypic design matrix that embodies constraints (such as the minimum interval between

stimulus onsets) and the basis functions employed to model hemodynamic responses **S**. Universally long SOA (e.g., Friston *et al.,* 1998b) designs are less efficient than rapid presentation designs. Nonstationary designs, with modulation of occurrence probabilities, can be more efficient than stationary designs. The most efficient design of all is the conventional deterministic block design. A slow modulation of occurrence probabilities gives high efficiency while retaining the advantages of stochastic designs and may represent a useful compromise between the high efficiency of block designs and the psychological benefits and latitude of stochastic designs.

In experiments with multiple trial types the most efficient design for one effect may not be the most efficient for another. This is particularly important when considering evoked responses and the differences among responses. The optimum SOAs for the former are longer than for the latter and attain when a null event is included with the same occurrence probability as the index events. In summary the choice of optimal design is critically dependent upon the nature of the inferences that are sought and in turn the effect that one is trying to estimate. Although there may be an optimal SOA for a given effect or compound of effects, the best SOA for one sort of inference may be different
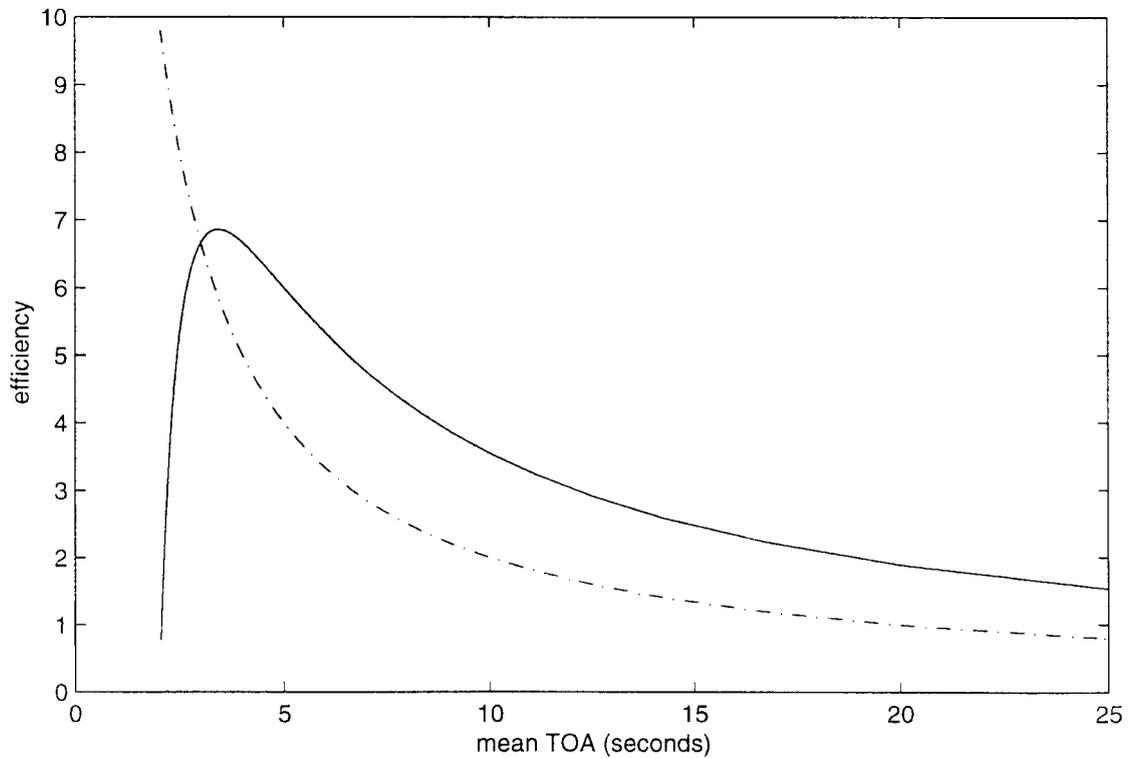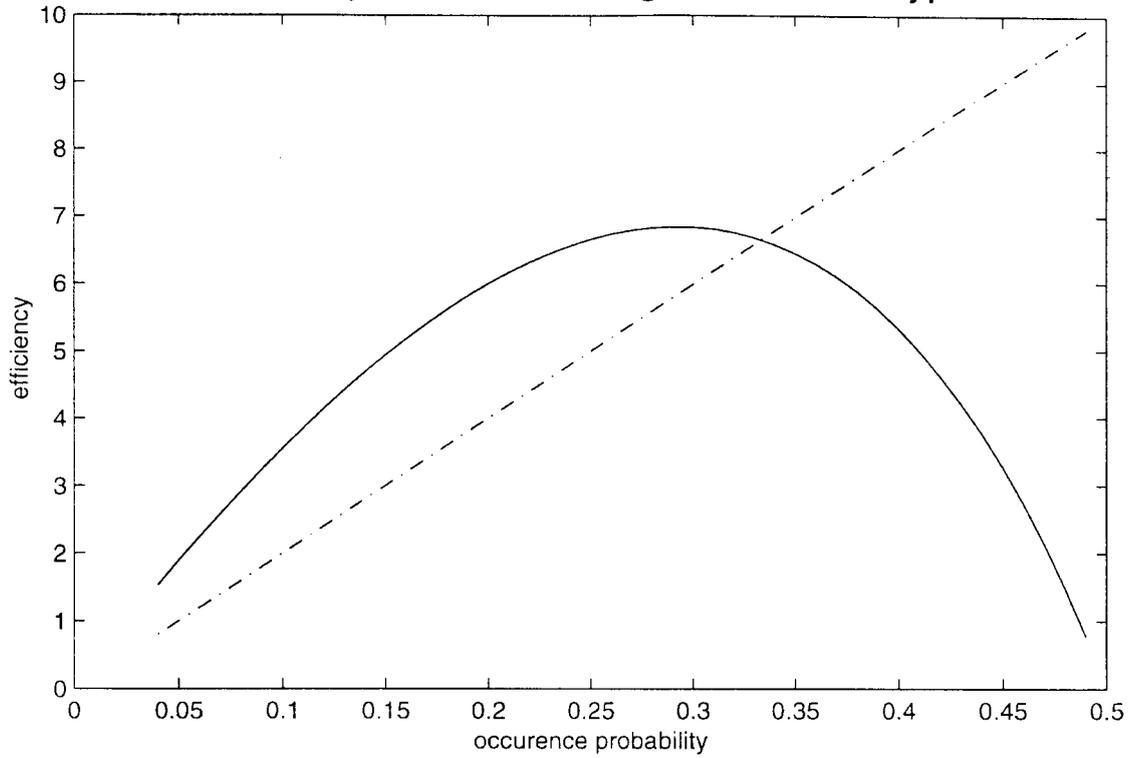
FIG. 6. As for Fig. 3 but now for two trial types. Solid line: Efficiency of evoked responses per se. Broken line: Efficiency of differential response. Here the time-series is 64 s.
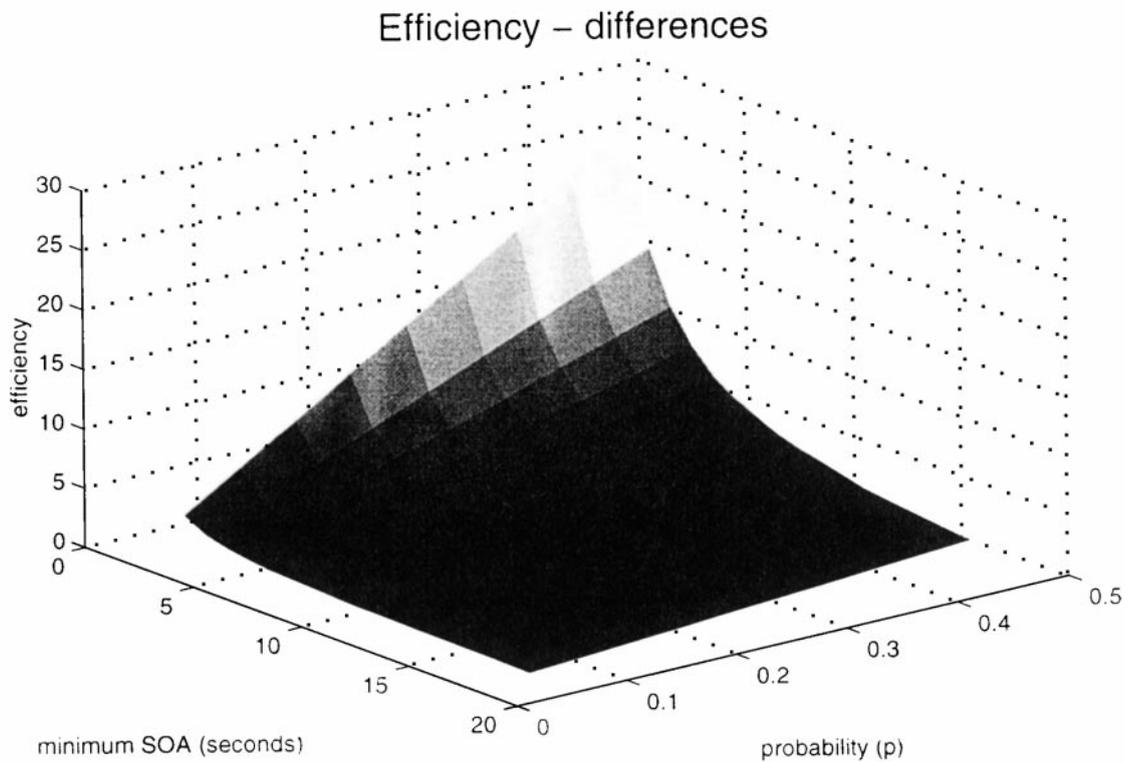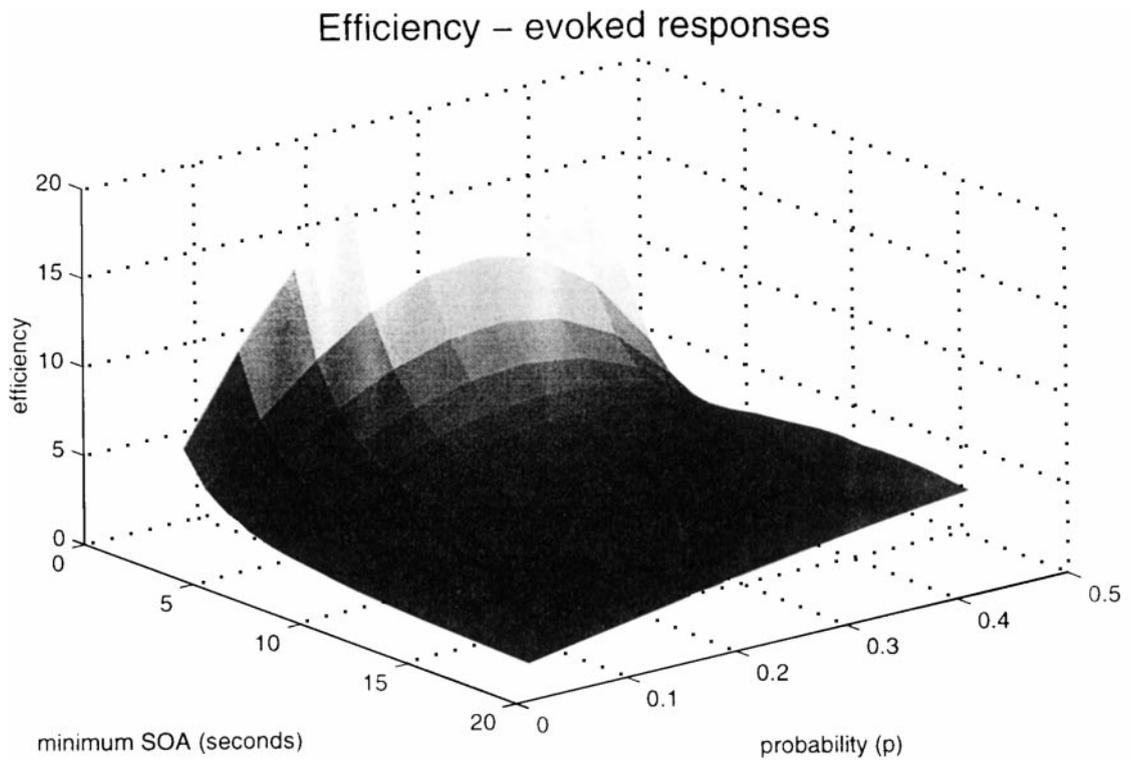
FIG. 7. As for Figs. 5 and 6 but now efficiency is expressed as a functions of both P and SOA_{min}. Here the time-series is 256 s.

from that required by another. A novel idea that emerged from this work is the notion of nonstationary or modulated stochastic designs that combine the facility to specify high-efficiency stimulus sequences and still render them unpredictable from the subject's point of view.

In event-related fMRI experiments with rapid presentation of multiple stimulus types the inclusion of a null event is a critical consideration and is necessary to characterize event-specific responses *and* differences in their form or latencies. One of the main motivations to characterize event-specific responses, in addition to differential responses, is to select areas that are activated by both event types and then see whether or not they show a difference. For example, consider a photic stimulation experiment under two different levels of attention. If one were interested only in the attentional modulation of evoked responses, then the most efficient design would involve the presentation of a stimulus every SOA_{min}. Note, however, that this experimental design would preclude any estimation of photic responses per se because there may be many areas that did not show attentional modulation and visually evoked responses in these areas would never be detected. In this example it may be useful to identify all regions that showed a visual evoked response and then test for attentional modulation in, and only in, these regions.

As a final comment, it should be noted that designs used in electrophysiology can be emulated in fMRI with very small SOAs as long as the occurrence probabilities are sufficiently low to render the expected SOA suitable for the inferences sought. In electrophysiology most of the interesting work has been predicated on differential evoked responses and, if this generalizes to event-related fMRI, very short SOA designs will have a central, if not exclusive, role in functional neuroimaging.

APPENDIX

Let the hemodynamic response to an event or trial be modeled in terms of k orthonormal basis functions $f_i(t_n)$, where t_n are the peristimulus times at which scans are acquired ($t_n - t_{n-1}$ is assumed to be small for simplicity):

$$\sum_n f_i(t_n) \cdot f_j(t_n) \, dt = \begin{cases} 1 & (i = j), \\ 0 & (i \neq j). \end{cases} \quad (A.1)$$

The $N \times Pk$ design matrix $\mathbf{X}$ will contain a column for each basis function, for each of the P trial types. N is the number of scans in the time series. For simplicity we will deal with one basis function and one trial type and generalize later. In this case $\mathbf{X}$ reduces to a $N \times 1$ column vector. Each realization of $\mathbf{X}$ can be expressed as $\mathbf{X} = \mathbf{S} \cdot \mathbf{b}$, where $\mathbf{S}$ is a $N \times M$ matrix whose mth column $\mathbf{S}_m$ contains a basis function at $t_m = m \cdot SOA_{min}$

($m = 1, 2, \ldots, M$), the times at which the mth event could occur. If $S_{n,m} = f(t_n - t_m)$, where t_n is the time of the nth scan, then $\mathbf{S}_m$ is the mean corrected vector $[S_{1,m}, S_{2,m}, \ldots, S_{N,m}]^T$. When constructing $\mathbf{S}$ it is important to consider events that may "cause" experimental variance even if they occurred before the start of scanning. $\mathbf{b}$ is a $M \times 1$ vector of independent variates taking the values 0 and 1 with probabilities $1 - \mathbf{P}$ and $\mathbf{P} = [P_1, P_2, \ldots P_M]^T$, respectively. Note that this formulation allows for occurrence probabilities that are specific to each of the M possible occurrences of a trial. The expectation of $\mathbf{X}^T\mathbf{X}$ is given by:

$$\mathbf{C} = \langle \mathbf{X}^T\mathbf{X} \rangle = \langle \mathbf{b}^T\mathbf{S}^T\mathbf{S}\mathbf{b} \rangle = \mathbf{P}^T(\mathbf{S}^T\mathbf{S} - \mathbf{D})\mathbf{P} + \sum_m P_m D_{mm}. \quad (A.2)$$

$\mathbf{D}$ is a diagonal matrix corresponding to diag($\mathbf{S}^T\mathbf{S}$). Equation (A.2) follows from the fact that the expectation $\langle b_i b_j \rangle = P_i P_j$ unless $i = j$ in which case $\langle b_i b_i \rangle = P_i$. The expected efficiency for a contrast of effects $\mathbf{c}$ is given by

$$\text{Efficiency} \sim \text{trace}[\mathbf{c}^T\mathbf{C}^{-1}\mathbf{c}]^{-1}. \quad (A.3)$$

Together Eqs. (A.2) and (A.3) give the expected efficiency for any contrast of effects in a design specified by the fixed parameters (SOA_{min}, basis functions, and number of scans) embodied in $\mathbf{S}$ and those pertaining to the probabilistic or stochastic components implied by $\mathbf{P}$. In the general case of multiple basis functions and trial types the element of $\mathbf{C}$ corresponding to the ith and jth basis functions for the uth and vth trial types is

$$\mathbf{P}^{uT}(\mathbf{S}^{iT}\mathbf{S}^j - \text{diag}(\mathbf{S}^{iT}\mathbf{S}^j))\mathbf{P}^v + \sum_m \langle b_m^u b_m^v \rangle \mathbf{S}_m^{iT}\mathbf{S}_m^j, \quad (A.4)$$

where $\mathbf{S}^i$ is the mean corrected basis function matrix for the ith basis function, $\mathbf{P}^u$ is the probability vector for the uth trial type, and

$$\langle b_m^u b_m^v \rangle = \begin{cases} p_m^u & (u = v) \\ 0 & (u \neq v). \end{cases}$$

That $\langle b_m^u b_m^v \rangle$ is zero for different trial types follows simply from the fact that two trial types cannot occur simultaneously.

Other columns in the design matrix may include fixed confounds. The expected value of the elements of $\mathbf{C}$ corresponding to the covariance between the regressor for the uth event type/ith basis function and any confound $\mathbf{g}$ is

$$\langle \mathbf{g}^T\mathbf{S}^i\mathbf{b}^u \rangle = \mathbf{g}^T\mathbf{S}^i\mathbf{P}^u. \quad (A.5)$$

Note that the constant term has been implicitly included as a confound above by virtue of the fact that the columns of **S** are mean corrected.

REFERENCES

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16:**4207–4221.

Buckner, R., Bandettini, P., O'Craven, K., Savoy, R., Petersen, S., Raichle, M., and Rosen, B. 1996. Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* **93:**14878–14883.

Buckner, R. L., Goodman, J., Burock, M., Rotte, M., Koutstaal, W., Schacter, D., Rosen, B. R., and Dale, A. M. 1998. Functional–anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. *Neuron* **20:**285–296.

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

Clark, V. P., Maisog, J. M., and Haxby, J. V. 1998. fMRI study of face perception and memory using random stimulus sequences. *J. Neurophysiol.* **76:**3257–3265.

Dale, A., and Buckner, R. 1997. Selective averaging of rapidly presented individual trials using fMRI. *Hum. Brain Mapping* **5:**329–340.

Elliot, R., Friston, K. J., and Dolan, R. Dissociable neural responses associated with reward, punishment and risk-taking behaviour. Submitted for publication.

Friston, K. J., Jezzard, P. J., and Turner, R. 1994. Analysis of functional MRI time-series. *Hum. Brain Mapping* **1:**153–171.

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

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., and Rugg, M. D., and Turner, R. 1998a. Event-related fMRI: Characterizing differential responses. *Neuroimage* **7:**30–40.

Friston, K. J., Josephs, O., Rees, G., and Turner, R. 1998b. Non-linear event-related responses in fMRI. *Magn. Reson. Med.* **39:**41–52.

Heid, O., Gönner, F., and Schroth, G. 1997. Stochastic functional MRI. *NeuroImage* **5:**S476.

Josephs, O., Turner, R., and Friston, K. J. 1997. Event-related fMRI. *Hum. Brain Mapping* **5:**243–248.

Rosen, B. R., Buckner, R. L., and Dale, A. M. 1998. Event related fMRI past present and future. *Proc. Natl. Acad. Sci. USA* **95:**773–780.

Vazquez, A. L., and Noll, C. D. 1998. Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage* **7:**108–118.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., Rosen, B. R., and Buckner, R. L. 1998. Encoding verbal experiences: Brain activity predicts subsequent remembering and forgetting. *Science* **281:**1188–1191.

Zarahn, E., Aguirre, G. K., and D'Esposito, M. 1997. A trial-based experimental design for fMRI. *NeuroImage* **5:**179–197.

Zarahn, E., Friston, K. J., Tang, Y. J., and Aguirre, G. K. A model of the relationship between temporal design parameters and sensitivity to experimental effects in BOLD fMRI. Submitted for publication.