Nonlinear Coupling between Evoked rCBF and BOLD Signals: A Simulation Study of Hemodynamic Responses

Andrea Mechelli, Cathy J. Price, and Karl J. Friston

Wellcome Department of Cognitive Neurology, Institute of Neurology, 12 Queen Square, London, WC1N 3BG, United Kingdom

The aim of this work was to investigate the dependence of BOLD responses on different patterns of stimulus input/neuronal changes. In an earlier report, we described an input-state-output model that combined (i) the Balloon/Windkessel model of nonlinear coupling between rCBF and BOLD signals, and (ii) a linear model of how regional flow changes with synaptic activity. In the present investigation, the input-state-output model was used to explore the dependence of simulated PET (rCBF) and fMRI (BOLD) signals on various parameters pertaining to experimental design. Biophysical simulations were used to estimate rCBF and BOLD responses as functions of (a) a prior stimulus, (b) epoch length (for a fixed SOA), (c) SOA (for a fixed number of events), and (d) stimulus amplitude. We also addressed the notion that a single neuronal response may differ, in terms of the relative contributions of early and late neural components, and investigated the effect of (e) the relative size of the late or "endogenous" neural component. We were interested in the estimated average rCBF and BOLD responses per stimulus or event, not in the statistical efficiency with which these responses are detected. The BOLD response was underestimated relative to rCBF with a preceding stimulus, increasing epoch length, and increasing SOA. Furthermore, the BOLD response showed some highly nonlinear behaviour when varying stimulus amplitude, suggesting some form of hemodynamic "rectification." Finally, the BOLD response was underestimated in the context of large late neuronal components. The difference between rCBF and BOLD is attributed to the nonlinear transduction of rCBF to BOLD signal. Our simulations support the idea that varying parameters that specify the experimental design may have differential effects in PET and fMRI. Moreover, they show that fMRI can be asymmetric in its ability to detect deactivations relative to activations when an absolute baseline is stipulated. Finally, our simulations suggest that relative insensitivity to BOLD signal in specific regions, such as the temporal lobe, may be partly explained by higher cognitive functions eliciting a relatively large late endogenous neuronal component.   © 2001 Academic Press

INTRODUCTION

This paper presents a series of biophysical simulation studies to explore the coupling of the hemodynamic response with the underlying neural activity and how this coupling depends on various parameters pertaining to experimental design. The biophysical simulation studies we present are based on an input-state-output model with input being stimulus/synaptic activity and output being the hemodynamic responses. The model was derived by combining the Balloon/Windkessel model (Buxton *et al.,* 1998), which specifies the coupling of perfusion to BOLD signal, with a dynamical model of the transduction of neural activity into perfusion changes. In an earlier report we showed that the model is sufficient to reproduce empirically-derived Volterra kernels and therefore to account for the more important nonlinearities observed in evoked fMRI responses (Friston *et al.,* 2000). Moreover, the parameters of the model that best reproduced empirically-derived Volterra kernels were all biologically plausible. These parameters included rCBF, signal decay, flow autoregulatory time-constants, transit time, venous compliance/capacitance and resting oxygen extraction fraction. We concluded that the Balloon/Windkessel model was sufficient to account for hemodynamic refractoriness and other nonlinear aspects of evoked responses in fMRI.

In the present work we used the hemodynamic input-state-output model to explore the dependence of PET and fMRI response estimates on various parameters pertaining to experimental design. In these simulation studies, we assumed that substantial nonlinearities enter when rCBF changes are translated into BOLD changes (Buxton *et al.,* 1998). Although nonlinearities between stimulus input and rCBF are apparent in certain contexts (Price *et al.,* 1992), they were not addressed in this paper. By comparing rCBF and

AP

BOLD responses we were able to explore how nonlinearities depend on the stimulus/neuronal input and experimental design. We were concerned with the estimate of the average rCBF and BOLD responses per stimulus or event and not with the statistical efficiency with which these responses are detected. Specifically, this work addressed the sensitivity of BOLD responses to stimulus input/neuronal changes by looking at the difference between rCBF and BOLD responses as functions of (a) the presence of a prior stimulus (hemodynamic refractoriness), (b) epoch length for a fixed SOA, (c) stimulus onset asynchrony (SOA) for a fixed number of events, (d) stimulus amplitude, and (e) a late or "endogenous" neuronal component. Assessing the effects of these parameters on rCBF and BOLD response estimates enables one to identify the optimum parameters for maximizing BOLD signal changes in neuroimaging studies and speaks to differential sensitivities when using different experimental designs, different classes of neural responses and different modalities (i.e., PET vs fMRI).

Hemodynamic Refractoriness

The effect of a preceding stimulus on the BOLD response to a second stimulus has already been characterised using Volterra kernels (Friston *et al.,* 1988). The response to the second stimulus can be isolated by subtracting the response to the first stimulus from the response to both. The resulting BOLD response to the second stimulus is markedly attenuated relative to the response to the same stimulus presented in isolation. Although profound refractory or saturation effects are manifest at SOAs of less than 1 s, they can be detected for SOAs of up to 4–6 s. The fact that the response to the second stimulus is compromised by the first suggests some refractoriness in the BOLD response. In this paper we modeled the rCBF and BOLD responses to a second stimulus in the context of the first using the hemodynamic input-state-output model described above. Since rCBF in this model conforms to a linear convolution of underlying stimulus input/neural activity, the rCBF response to a second stimulus will not be affected by a first stimulus. In contrast, the Balloon/Windkessel component is fully nonlinear and can, in principle, account for the refractoriness of BOLD response. We demonstrate this to be the case below.

Epoch Length

The effect of epoch length on the estimated BOLD response can be thought of in terms of the nonlinear refractoriness of BOLD responses to a second stimulus, in the context of the first, compounded by a whole train of prior stimuli. The epoch length is a variable of great interest in box-car paradigms. In PET, the hemodynamic responses are summed over a 60- to 90-s scan; this time can not be manipulated because of intrinsic limitations of the measurement technique. In fMRI, on the other hand, data are acquired by sampling the hemodynamic response to each stimulus and there are no constraints on epoch length. Thus, it is important to identify the optimum epoch length for maximizing signal sensitivity in fMRI studies. Previous simulations suggest that, for block designs, the BOLD signal improves dramatically above 8 s until it reaches a plateau at around 32 s (e.g., Hutton *et al.,* 1998). These simulations rely on a linear model to address design efficiency. In contrast, we employed our nonlinear model to estimate the average response per stimulus or event. Using the hemodynamic input-state-output model, we modeled the rCBF and BOLD responses as functions of epoch length for a fixed SOA of 1 s. Such an SOA was chosen because, in our unit, this is the shortest SOA that is guaranteed to be relatively immune from profound hemodynamic refractoriness (see next section). While rCBF per stimulus unit will not be affected by epoch length, we anticipated a relative decrease in the estimated average BOLD response with increasing epoch length and were interested in the nature and degree of this nonlinear effect. This characterization may speak to the trade-off between short epochs (that reduce refractoriness by allowing recovery) and longer epochs that harness the natural time constants of the hemodynamic response function.

Stimulus Presentation Rate and SOA

In the last few years there has been a growing interest in the choice of SOAs in event-related fMRI. For any design, shorter SOAs allow a greater number of trials or events for averaging over the same period. Short SOAs also increase the amplitude of the signal thereby increasing the likelihood that it is detected. For these reasons, very short SOAs have been proposed (Dale and Buckner, 1997; Burock *et al.,* 1998; Clark *et al.,* 1998). However, the study by Friston *et al.* (1998) indicated that the nonlinear refractoriness of BOLD responses is expressed at very short interstimulus intervals. In particular, nonlinear effects can lead to a decrease in the integrated response for SOAs of one second or less. Other studies have also shown that nonlinear effects predominate at very short SOAs (Vazquez and Noll, 1998). These findings suggest that a lower limit on SOAs will be dictated by nonlinear interactions among events. Although profound refractory or saturation effects are manifest at SOAs of less than 1 s, they can be detected for longer SOAs. Using the hemodynamic input-state-output model, we estimated the rCBF and BOLD responses for increasing SOAs for a fixed number of events. We expected that the average BOLD response to each event would reduce with decreasing SOAs. Again we were interested in the nature of this nonlinearity.

Stimulus Amplitude

In the above simulations, the simulated input amplitude was kept constant. However, previous studies suggest that this variable affects both rCBF and BOLD responses. For example, Dettmers *et al.* (1996) reported a logarithmic relationship between key press force and rCBF or BOLD signals in the motor cortex. In addition, Vazquez and Noll (1998) detected a nonlinear relationship between BOLD signal and stimulus amplitude in the primary visual cortex. We specifically addressed how the estimated average rCBF and BOLD responses are differentially affected by stimulus amplitude. In particular we were interested in whether BOLD sensitivity to neuronal deactivations relative to an absolute baseline (corresponding to negative stimulus amplitudes) were the same as to neuronal activations (corresponding to positive stimulus amplitudes).

The Form of the Evoked Neural Response

The input parameters discussed above relate to experimental design and assume the same form of neuronal response is evoked by each trial or stimulus. In this section we address the notion that neuronal responses themselves may vary, in terms of the relative contributions of early and late neural components. This may cause differences in the estimated hemodynamic responses when BOLD or rCBF is measured. In the final simulation, we therefore investigated the differential effect of late endogenous components on rCBF and BOLD responses to complex stimuli. By complex stimuli we mean stimuli that elicit multiple components in electrophysiological studies. The term "endogenous component" is taken from the event-related brain potential literature, which distinguishes between two classes of evoked waveform components that follow events (Rugg and Coles, 1995). On the one hand there is a set of components whose characteristics (amplitude, latency, and distribution) depend on the physical properties of sensory stimuli, such as their modality and intensity. These are "exogenous components." On the other, there is a second set of components whose characteristics depend on the nature of the subject's interaction with the stimulus (i.e., processing or top-down effects). These are "endogenous components." The event-related brain potentials that occur within the first 100 ms of stimulus presentation tend to be exogenous, whereas the event-related brain potentials that occur later tend to be endogenous. Irrespective of the subject's state, some stimuli tend to elicit late endogenous components more than others. For instance, in the temporal lobe, visual and auditory words are known to elicit both an exogenous component and a late endogenous component (N400) (Kutas and Hillyard, 1980). These late components are due to the complex nature of verbal stimuli, that involve functional integration among a number of brain regions over time. It has been argued that the N400 is only elicited by linguistic stimuli and reflects high-level lexical processing. A critical question is how the early and late components interact to differentially affect signal based on blood flow and deoxyhemoglobin concentration. In other words, what is the differential sensitivity of rCBF and BOLD to the endogenous component of neural responses in the presence of an exogenous component? We used the hemodynamic input-state-output model to simulate the rCBF and BOLD correlates of different neuronal responses to the same stimuli. While the exogenous component remained the same, the endogenous component was manipulated so that its size, relative to the exogenous component, varied from very small to very large. We expected that the exogenous response would act like a "prior" stimulus and compromise the ability of the endogenous component to elaborate a BOLD response. This might partially account for reported failures in detecting BOLD signal relative to PET in regions involved in language processing, namely in temporal regions where reduced signal can not be accounted for by susceptibility artifacts (Veltman *et al.,* 2000).

METHODS

Estimating the rCBF and BOLD signal per stimulus involved two stages. First, the hemodynamic model was used to simulate rCBF and BOLD responses for single events and trains of stimuli with different epoch lengths, SOAs, amplitudes, and neural responses. Second, the average response per stimulus was estimated.

Simulating rCBF and BOLD Responses

The hemodynamic model previously described by Friston *et al.* (2000) (see Appendix) was used to simulate the rCBF and BOLD responses to changes in stimulus input/neuronal activity. The six parameters of the model and the ranges of values explored are reported in Table 1.

Estimating Average rCBF and BOLD Responses

After simulating the rCBF and BOLD responses using the input-state-output model, we estimated the average rCBF and BOLD response to each event. The average BOLD responses were estimated assuming a linear convolution model with a canonical hemodynamic response function (HRF). The HRF assumed was based on the model's impulse response function (i.e., the response to a unit stimulus of infinitely short duration). This conforms to a conventional event related analysis of fMRI time series where the parameter estimates correspond to the average BOLD response per stimulus or stimulus component. Because only the balloon component of the model is nonlinear, the translation of neural activity into rCBF is effectively a linear

TABLE 1

The Six Parameters of the Hemodynamic Model and the
Ranges of Values Explored

The model's parameters	Values as estimated in Friston *et al.* (2000)	Parameter ranges explored
RCBF component		
Neuronal efficacy	0.54	0.05–1.00
Signal decay	0.86	0.40–1.00
Autoregulation	0.41	0.25–1.00
Balloon component		
Transit time	0.98	0.10–1.00
Stiffness parameter	0.33	0.05–0.70
Resting oxygen extraction	0.34	0.01–0.80

Note. The values of the six parameters are based on an empirical fMRI study of single word presentation at varying frequencies (Rees *et al.,* 1998). These values were identified as those that best reproduced the empirical data in terms of the first and second order Volterra Kernels as described in Friston *et al.* (2000).

convolution. By virtue of the fact we used the model's impulse response function, the parameter estimates for the rCBF response will always be unity. In contrast, the BOLD responses may be over-estimated or under-estimated depending upon how the nonlinearities in the balloon component interact with the experimental and neuronal factors manipulated.

In summary, by using a standard response estimation procedure based upon a linear convolution model, we characterized the effect of various manipulations on the estimated BOLD response in terms of a deviation from unity. The estimated BOLD response is expressed in terms of a single parameter that scales the hemodynamic response function used in the linear model. The estimates are simply the amount of the stimulus function required to explain the observed response under linear assumptions. Estimates of <1 suggest that the BOLD response is being under-estimated, or is less than would have been seen had the stimuli been presented in a more optimal fashion. Estimates of >1 suggest a super-additive effect of stimuli in elaborating a BOLD response, or equivalently, an over-estimation of the response to a stimulus presented in isolation. The estimates were obtained from a standard multiple linear regression analysis, as would be implemented in SPM, using the stimulus or neural activity functions, convolved with the model's HRF, as regressors. The response variable in the simulation studies was the output of the nonlinear hemodynamic model using the same input or stimulus function embodying the nonlinear dynamics.

RESULTS

Hemodynamic Refractoriness

The first simulation addressed the effect of a preceding stimulus on the BOLD responses to a second stim-

ulus, for an SOA of 1 s. In the first instance, we modeled responses to a pair of words presented together and in isolation. We characterized the response to a second word in the context of the first by subtracting the response to the first word from the response to both. The difference between the responses to the second word when preceded by the first and when presented alone, reflects the impact of the first word on the response to the second. The differential effect of a preceding stimulus on BOLD and rCBF responses is shown in Fig. 1. The upper panels illustrate the simulated responses to the stimuli presented together (solid line) and separately (broken lines). The lower panels illustrate the simulated responses to a word presented alone (broken line) and to a second word when preceded by the first (solid line). The rCBF response to a second stimulus is not affected by previous stimulation. However, it can be seen that the BOLD response to the second stimulus is markedly attenuated, with an augmented undershoot, in relation to the response observed when the stimulus is presented in isolation. In other words, the response to a stimulus is modulated by preceding stimuli to give a BOLD refractoriness.

Epoch Length

This simulation addressed the rCBF and BOLD responses as functions of epoch length for a fixed SOA of 1 s. We created 32 stimulus functions composed of one epoch with a length varying from 1 to 32 stimuli. Again, we simulated the rCBF and BOLD responses using the hemodynamic model and estimated the magnitude of the responses under linear assumptions. The differential effect of epoch length on rCBF and BOLD is shown in Fig. 2. The average BOLD response (solid line) falls with increasing epoch length and then levels out when length is around 10 s. In short, the average BOLD signal decreases with epoch length even when the average rCBF response (broken line) is not affected by previous activation or epoch length.

Stimulus Presentation Rate and SOA

The second simulation addressed how the nonlinear refractoriness illustrated in the previous section depends on the interstimulus interval or rate, by looking at the effect of increasing the SOA on BOLD estimates for a fixed number of events.

We created a number of stimulus functions, each containing eight stimuli, with an SOA varying between 0.25 and 16.25 s. We simulated the rCBF and BOLD responses to each stimulus function. The average BOLD response to a single stimulus was estimated as described above. Figure 3 shows the rCBF and BOLD response estimates as functions of the SOA. As stipulated by the model, the rCBF signal (broken line) is not modulated by the SOA. In contrast, the estimated BOLD response (solid line) increases dramatically with
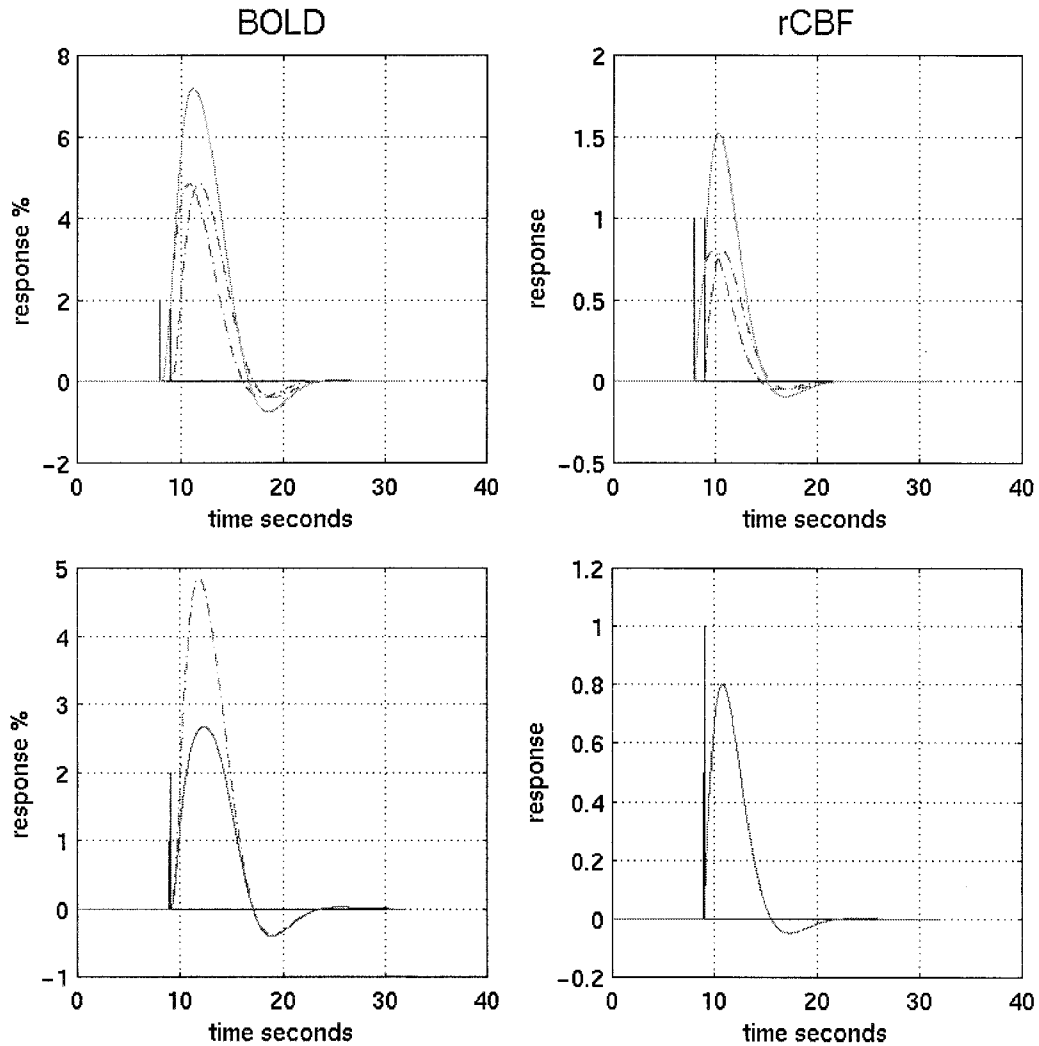
BOLD

rCBF

FIG. 1. Top panels: Simulated BOLD (left) and rCBF (right) responses to a pair of words (bars) one second apart, presented together (solid line) and separately (broken lines). Lower panels: Simulated BOLD (left) and rCBF (right) responses to the second word when presented alone (broken line as above) and when preceded by the first (solid line). The latter obtains by subtracting the response to the first word from the response to both. The difference reflects the effect of the first word on the response to the second. It can be seen that the BOLD response to the second stimulus is markedly attenuated in relation to the response observed when the stimulus is presented in isolation.

SOA, until it reaches a maximum around 8 s. As the SOA increases further, the BOLD signal decreases slightly until it reaches a plateau for an SOA of around 12 s. These data suggest that the estimated BOLD response dramatically decreases with increasing rates of presentation and that an optimal SOA occurs at around 8 s. Thus, the response to a stimulus is modulated by preceding stimuli to give a nonlinear refractoriness that depends on the interstimulus interval or rate.

Stimulus Amplitude

Here we simulated the evoked rCBF and BOLD responses to a stimulus with varying amplitude. We used 32 stimulus functions, each comprising a single stimulus whose amplitude varied from −4.5 to 4.5. Positive

stimulus amplitudes correspond to stimuli that elicit an increase in neuronal activity relative to an absolute baseline. On the other hand, negative stimulus amplitudes correspond to stimuli inducing a decrease in neuronal activity relative to baseline. In the latter case, the presentation of each stimulus transiently interrupts the ongoing activity. Negative neuronal input has a simple and explicit meaning in the context of the model because baseline activity corresponds to the situation where the input is zero. When input falls below zero this corresponds to a decrease from resting or tonic levels of neural activity. Transient neuronal deactivations are easily modeled using negative spikes of varying amplitude. In contrast to the other simulation studies, the behavior was highly dependent on the hemodynamic model parameters, particularly resting
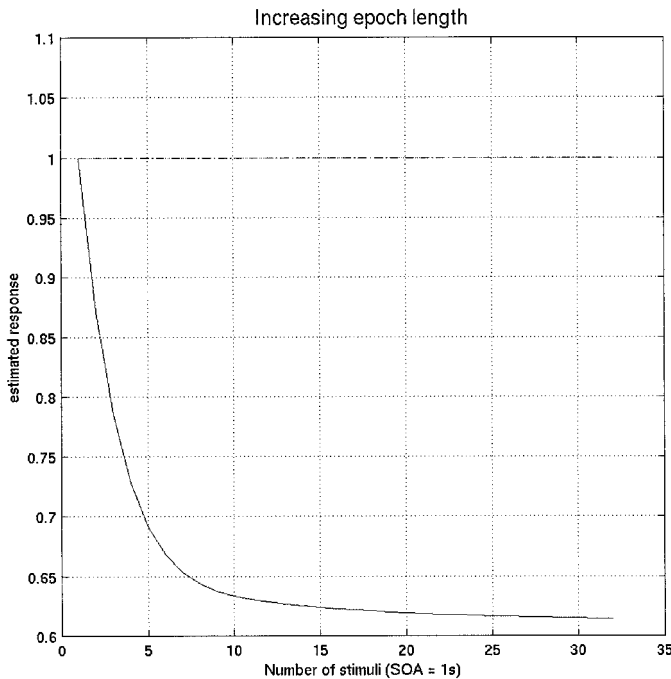
FIG. 2. The differential effect of epoch length on estimated rCBF and BOLD responses per stimulus. The BOLD response (solid line) falls dramatically with increasing epoch length and then levels out at around 10 s. On the other hand, the rCBF response (broken line) is not affected by the epoch length.

oxygen extraction fraction (E_0). We therefore report these simulations for values of E_0 varying between 0.11 and 0.55. The left panel of Fig. 4 shows the BOLD responses as functions of stimulus amplitude and E_0. It can be seen that the estimated BOLD response increases linearly with positive stimulus amplitude, but when stimulus amplitude is negative some highly nonlinear behavior ensues. This nonlinear behavior is more pronounced for relatively high values of E_0 and corresponds to a rectification-like effect. Our use of the term "rectification" comes from the electrophysiology of neurons. Rectifying currents result in nonlinear voltage-current relationships very reminiscent of the stimulus response characteristics shown by our model. We have chosen to use the term "rectification" but qualify its use by making it clear this is a partial rectification. In our context, rectification refers to the selective attenuation of negative signal components. It can also be seen that the increase in BOLD response with stimulus amplitude shows a saturation effect.

The right panel of Fig. 4 shows the simulated rCBF response as functions of stimulus amplitude and E_0, demonstrating how these responses are linearly dependent on stimulus amplitude irrespective of the value of E_0. These results suggest that PET and fMRI may be differentially sensitive to decreases relative to increases in neuronal activity.

The Form of the Evoked Neural Response

This simulation addressed the effect of differential neuronal responses, in terms of the relative contributions of early and late neural components, on BOLD and rCBF responses. We created 16 neuronal response functions, each of them comprising an early exogenous and a late endogenous component. While the exogenous component remained the same, the endogenous component was manipulated in such a way that its size relative to the exogenous component varied from very small to very large (see upper panel of Fig. 5). The timing of the endogenous component was based upon the ERP literature where pronounced cognitive positivities (e.g., P300) and negativities (e.g., N400) represent important and robust electrophysiological correlates of late neuronal components. We used the hemodynamic model to simulate the rCBF and BOLD responses to each neuronal response function and calculated the BOLD parameter estimates as above. The relative size of the late endogenous component differentially affected the signal generated from blood flow and deoxyhemoglobin concentration, as shown in Fig. 5 (lower panel). A decrease in the BOLD parameter es-
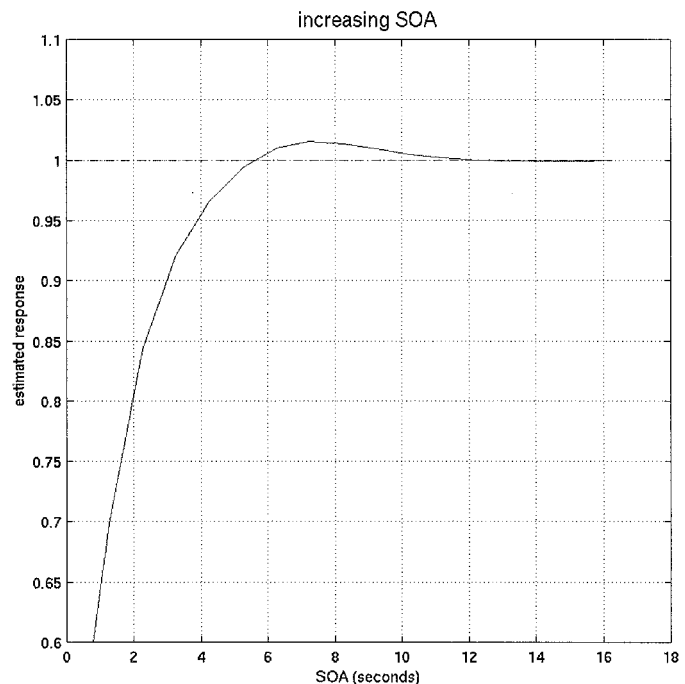

FIG. 3. The estimated rCBF and BOLD responses per stimulus as functions of the SOA. The estimated BOLD response (solid line) is the scalar coefficient that, when multiplied by the hemodynamic response function used to construct the regressor, best predicted the observed simulated response (in a least squares sense). It can be seen that it increases dramatically with SOA until it reaches a maximum for an SOA of around 8 s. As the SOA further increases, the estimated BOLD response decreases slightly until it reaches a plateau for an SOA of around 12 s. On the contrary, the estimated rCBF response (broken line) is not modulated by the SOA.
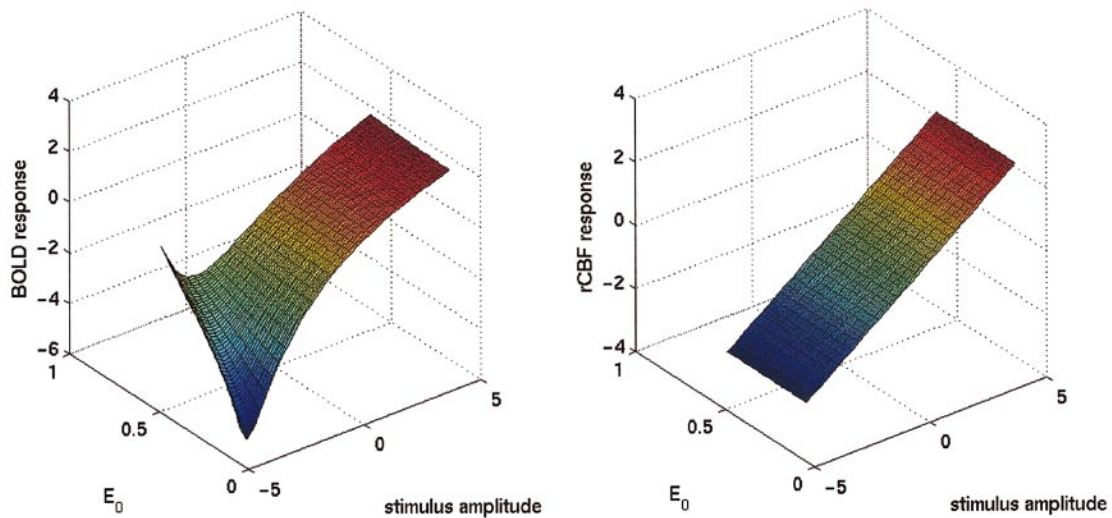
BOLD and rCBF estimated responses as functions of stimulus amplitude and E_0

FIG. 4. BOLD (left panel) and rCBF (right panel) responses as functions of stimulus amplitude and resting oxygen extraction fraction (E_0). It can be seen that the estimated BOLD response increases linearly with positive stimulus amplitude, but when stimulus amplitude is negative some highly nonlinear behaviour ensures. This nonlinear behaviour is more pronounced for relatively high values of E_0. On the other hand, the rCBF responses are linearly dependent on stimulus amplitude irrespective of the value of E_0. It can also be noticed that the increase in BOLD response with stimulus amplitude shows a saturation effect relative to the increase in rCBF response.

timates (solid line) was incurred by increasing the relative size of the endogenous component. Thus, the exogenous response acted like a "prior" stimulus and compromised the ability of the endogenous component to elaborate a BOLD response. On the contrary, as enforced by the model, the rCBF parameter estimates (broken line) were unaffected by the relative size of the endogenous component (i.e. increased in direct proportion to the size of the late component).

DISCUSSION

We have investigated the coupling among stimulus input/neural activity, hemodynamic responses, and changes in fMRI signal. This sort of investigation is important for a better understanding of how the underlying neuronal activity may elicit BOLD signal changes. In the first instance, we explored the dependence of PET and fMRI signals on various parameters pertaining to experimental design. We then addressed the notion that the neuronal responses themselves may differ, in terms of their amplitude and the relative contributions of early and late neural components. Our simulations were based on an input-state-output hemodynamic model whose validity, in relation to empirical nonlinear characterisations of evoked responses in fMRI and other neurophysiological constraints, has been established previously (Friston *et al.,* 2000).

Consistent with previous findings (Friston *et al.,* 1998), we found that the response to a second stimulus

was compromised by the first, evidencing a nonlinear refractoriness of the BOLD response (see Fig. 1). This suggests that the nonlinear behaviour of the Balloon/Windkessel component can, in principle, account for the refractoriness of BOLD response. As a consequence of this refractoriness, the average BOLD signal per event decreases with epoch length (see Fig. 2). However, our data showed that there is only little gain, in terms of BOLD responsiveness, in limiting the epoch length to 10 s instead of 20 or 30 s. In contrast, a considerable gain would be obtained by limiting the epoch length to 1 or 2 s. However, it is important to keep in mind that our results are concerned with the rCBF and BOLD response estimates and not with the statistical efficiency with which these responses are estimated. This means that the choice of a very short epoch length may be ideal in terms of minimizing refractoriness, but is not optimal in terms of design flexibility and statistical power. For example, with epochs containing more than 2 or 3 stimuli, the hemodynamic response reaches a steady state, which increases the likelihood that the signal will be detected irrespective of its size and shape and the precision of sampling. Generally, to maximize efficiency epoch length should approximate the time constants of the HRF. Other variables to consider when choosing the epoch length are attention at set (which may change with time), and habituation (which may modulate neuronal activity associated with task performance).
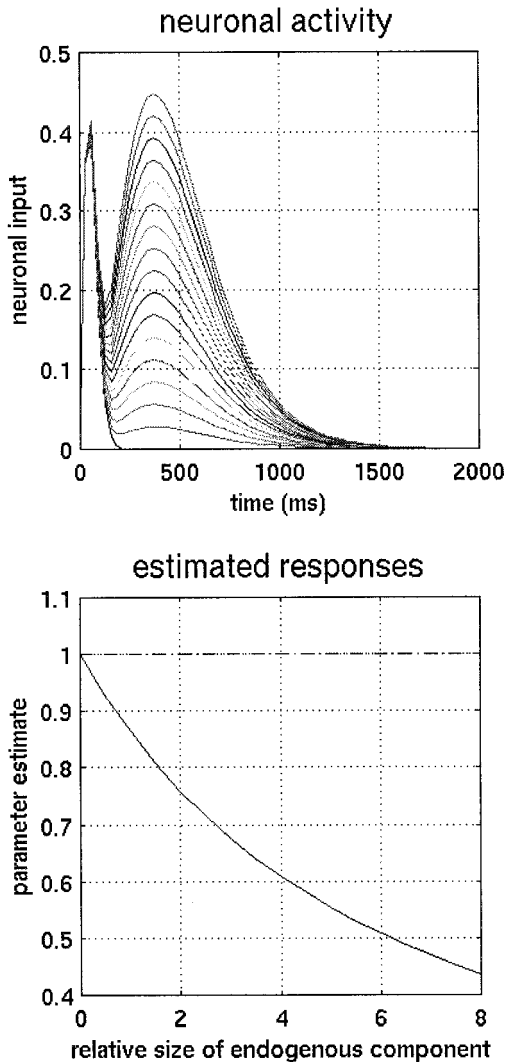
FIG. 5. Upper panel: Neuronal activity functions constructed to calculate the rCBF and BOLD parameter estimates. While the exogenous component remained the same, the endogenous component was manipulated in such a way that its size relative to the exogenous component varied from very small to very large. The early and late components were both modeled with gamma functions peaking at 100 and 400 ms, respectively. Lower panel: rCBF and BOLD parameter estimates as functions of the relative size of the late endogenous component. A decrease in the BOLD parameter estimates (solid line) occurred with increasing relative size of the endogenous component. On the contrary, the rCBF parameter estimate (broken line) was not affected by the relative size of the endogenous component.

Another effect of interest was that the estimated BOLD response to each stimulus decreased dramatically with increasing rate of presentation (see Fig. 3). This illustrates that hemodynamic refractoriness depends on the stimulus onset asynchrony or rate. Previous studies suggested that the optimal fixed SOA for BOLD responsiveness was 12–16 s (Bandettini and Cox, 1998; Hutton *et al.,* 1998). These simulations relied on a linear model and therefore could not account for the nonlinearities observed in evoked fMRI re-

sponses. By contrast our simulation, which relies on a fully nonlinear model and accounts for the nonlinearities observed in real time series, suggests that an average SOA of 8 s may be better because successive stimuli interact in a super-additive fashion. This can be attributed to the fact that, during the flow undershoot following a first stimulus, deoxyhemoglobin concentration is greater than normal, thereby facilitating clearance of deoxyhemoglobin following a second stimulus (Friston *et al.,* 2000). What are the implications of these findings for experimental design? These results might suggest that an SOA of 8 s minimizes BOLD refractoriness and thus the choice of such a long SOA is optimal. However, this is not the case because a long SOA (a) severely restricts the latitude of experimental paradigms for event-related fMRI and (b) greatly reduces the design efficiency by (i) limiting the number of trials available for averaging within a fixed time and (ii) reducing the peak amplitude of the signal (by reducing summation of independent effects). Thus, the choice of the SOA should take into account both the maximization of the BOLD responsiveness given by a long SOA and the statistical power afforded by a short SOA. Miezin *et al.* (2000) reported that, despite BOLD refractoriness effects occurring in motor and visual cortices, the statistical power for detecting a response was significantly greater at fast rates owing to the increased number of trials possible and the increased signal-to-noise ratio. However, our simulations suggest that at some lower limit on the SOA the decreased response to each stimulus may supervene over increases in design efficiency.

We also manipulated the stimulus amplitude to investigate its effect on average rCBF and BOLD responses per event. Our simulations suggest discrepancy between PET and fMRI responses to stimuli of varying amplitude (see Fig. 4). In particular we found that, in relation to the absolute "resting" state afforded by the hemodynamic model formulation, fMRI is asymmetric in its ability to detect deactivations relative to activations, again due to the inherent nonlinearities in BOLD signal transduction. This nonlinearity can be thought of as hemodynamic "rectification." This may account for reported failures in detecting negative effects of stimulus rate when brain activity is measured with fMRI as opposed to PET (Mechelli *et al.,* 2000). The increase in BOLD response with stimulus amplitude also showed a saturation effect, which can be explained in terms of hemodynamic refractoriness. In short, our findings suggest that PET and fMRI may be differentially sensitive to decreases in neuronal activity.

A critical aspect of this work was the finding that varying neuronal responses themselves may have differential effects on rCBF and BOLD signals. By early components we mean evoked electrophysiological changes that occur within 100 milliseconds. By late we

refer to those wave form components that show clear cognitive effects such as the P300 and N400. Interestingly, intermediate components such as face specific N200 are insensitive to a wide variety of top-down processes (Puce *et al.,* 1999). We found that, when the relative size of a late neuronal component increased, there was a decrease in the BOLD parameter estimates even though the rCBF estimates remained constant (see Fig. 5). Such differential responses may account for the reported failures in detecting BOLD signal in temporal regions involved in language processing that are not subject to susceptibility artifacts (see Veltman *et al.,* 2000). Verbal stimuli involve functional integration among a number of brain regions over time and therefore elicit very large late endogenous components (Kutas and Hillyard, 1980). Our findings suggest that BOLD sensitivity decreases when increasing the relative size of the late endogenous neuronal component and therefore the signal reduction observed in some regions of the temporal lobe may be related to the complex nature of verbal stimuli. In this context, the reduced BOLD response to the late endogenous component can be thought of as an effect of prior stimulation given by the exogenous component. In other words, the decrease in BOLD response to increasing the late endogenous component can be explained in terms of BOLD refractoriness. Interestingly, in many other parts of the brain it may not be possible to elicit an early component without inducing a late component a few hundred milliseconds later. This suggests that there may be a regionally specific variation in the sensitivity of BOLD responses to the total amount of neuronal activity that is a function of top-down effects intrinsic to particular brain regions and functions.

In this paper, we have reported and discussed the main effects of preceding stimulation, SOA, epoch length, stimulus amplitude, and the relative size of the late endogenous neuronal component on rCBF and fMRI signals. We also performed simultaneous manipulations of the above parameters in order to address their interactions and found quantitative, as opposed to qualitative, variations in the effect of one parameter when changing the others. These interactions were fully expected since the effects of the above parameters share a common origin, namely BOLD refractoriness. However, for the purposes of the present study, we were only interested in the main effects. Our findings (a) support the idea that varying parameters that specify the experimental design may have differential effects in PET and fMRI, (b) suggest that fMRI is asymmetric in its ability to detect deactivations, relative to activations, and (c) suggest that relative insensitivity to BOLD signal in specific regions, such as temporal areas, may be partly explained by complex stimuli eliciting a large late endogenous neuronal component.

Clearly the relevance of the nonlinear behaviors reported in this paper rests upon the validity of the Balloon/Windkessel component of the hemodynamic model used. Its validity was partially established in Friston *et al.* (2000) in which the model parameters required to reproduce empirically observed nonlinear responses fell exactly into plausible biological ranges. However, the nonlinear behaviors reported in this paper, for example hemodynamic rectification, should be regarded as proof of concept under the assumption that the model employed was sufficient to emulate real hemodynamics at the level modeled. It is quite possible that the phenomena described may be trivial in certain domains of parameter space or under certain experimental conditions. Furthermore, we have not as yet addressed nonlinearities that enter between stimulus input and rCBF, only those entering between rCBF and BOLD.

In relation to the validity of the Balloon/Windkessell component of the hemodynamic model, it is important to distinguish between model assumptions on which our conclusions rest and those model components, which do not affect our results. The central hemodynamic behaviors reported in this paper are nonlinear in nature. The most important nonlinearity in any hemodynamic model is the output nonlinearity that reflects the contribution of both intra- and extravascular signal components. Although, a number of the state equations are nonlinear they are only weakly so. This means that the nonlinearities introduced by oxygen extraction fraction dependency on flow or out-flow dependency on volume are probably not critical for phenomena like hemodynamic refractoriness. This can be inferred from the analysis presented in Friston *et al.* (2000). In this analysis, a bilinear approximation to the state equations was shown to be sufficient to account for nonlinearities observed empirically. Because the bilinear approximation precludes interactions amongst the state variables these can be discounted as important factors when accounting for the overall nonlinear behaviour. One could actually represent the hemodynamic model as two parallel linear convolutions of the input to produce the oxy-hemoglobin content and volume, respectively. The resulting BOLD signal is a nonlinear function of these two variables and it is this nonlinearity that accounts for the behaviors reported above. The implications of this are that any subtle changes to the nonlinear state equations are unlikely to change the phenomenal behavior at the level we have addressed. In short, although there are important controversies concerning, for example, the exact coupling between oxygen extraction fraction and flow, it is unlikely that small changes in the nature of these equations will change the conclusions based upon the simple hemodynamic model presented in this paper.

The qualitative behaviours described in this paper were evident over a large regime of parameter space (see Table 1 for the ranges of the parameters we explored). Clearly, we cannot guarantee that every sub-

ject studied will have a set of parameters that falls within these ranges however based upon the normal physiological ranges we are fairly confident that our simulations encompass most normal subjects and cortical areas. The parameters specified in the hemodynamic model were validated in Friston *et al.* (2000) by demonstrating that they were sufficient to reproduce the empirically derived Volterra kernels from a single subject. The experimental paradigm was specifically designed to address physiological issues about the nature of stimulus to flow and BOLD coupling. Indeed, the experimental paradigm was reproduced in the same subject using Positron Emission Tomography while emulating the fMRI stimulus presentation conditions. This multimodal data set allowed us to ensure that the coupling between input and flow was indeed linear. These and related analyses were reported in Rees *et al.* (1998) and Friston *et al.* (1998).

Although our investigation provides theoretical evidence for refractoriness and rectification effects in fMRI, the generalization of our findings to all situations should be cautioned. For example, neuronal activity and/or its coupling to BOLD signal differ across different brain regions. This is supported by the fact that the amplitude and timing of the BOLD response in one region fails to predict those properties in another region, even for within-subject comparisons (Miezin *et al.,* 2000). However, a number of studies suggest that refractory effects occur not only in the superior temporal lobe (Friston *et al.,* 1998), but also in other parts of the brain such as somatosensory, motor, and visual cortices (Cannestra *et al.,* 1998; Miezin *et al.,* 2000; Huettel and McCarthy, 2000).

We have presented a theoretical contribution on the coupling of the hemodynamic response with the underlying neural activity and how this coupling depends on various parameters pertaining to experimental design. Empirical characterization of the nonlinearities described by the present study is required to further validate our findings and the model itself.

APPENDIX

The model we used essentially combines the Balloon/Windkessel model and a simple dynamical model of changes in rCBF caused by neural activity.

The Balloon Component

The Balloon component links rCBF and the BOLD signal as described in Buxton *et al.* (1998). All variables are expressed in normalized form, relative to resting values. The BOLD signal is taken to be a static nonlinear function of normalized total deoxyhemoglobin voxel content (q), normalized venous volume (v), resting net oxygen extraction fraction by the capillary bed (E_0), and resting blood volume fraction (V_0)

$$y(t) = V_0(k_1(1 - q) + k_2(1 - q/v) + k_3(1 - v))$$
$$k_1 = 7E_0$$
$$k_2 = 2 \quad \quad \text{(a1)}$$
$$k_3 = 2E_0 - 0.2.$$

This signal comprises a volume-weighted sum of extra- and intravascular signals that are functions of volume and deoxyhemoglobin content. The latter are the state variables. The rate of change of volume is simply

$$\tau_0 \dot{v} = f_{in} - f_{out}(v). \quad \quad \text{(a2)}$$

Equation (2) says that volume changes reflect the difference between inflow f_{in} and outflow f_{out} from the venous compartment with a time constant τ_0. This constant represents the mean transit time (i.e., the average time it takes to traverse the venous compartment or for that compartment to be replenished) and is V_0/F_0 where F_0 is resting flow. Note that outflow is a function of volume. This function models the balloon-like capacity of the venous compartment to expel blood at a greater rate when distended. We model it with a single parameter α based on the Windkessel model

$$f_{out}(v) = v^{1/\alpha}, \quad \quad \text{(a3)}$$

where $1/\alpha = \gamma + \beta$. (c.f. Eq. (6) in Mandeville *et al.,* 1999). $\gamma = 2$ represents laminar flow. $\beta > 1$ models diminished volume reserve at high pressures and can be thought of as the ratio of the balloon's capacitance to its compliance.

The change in deoxyhemoglobin $\dot{q}$ reflects the delivery of deoxyhemoglobin into the venous compartment minus that expelled (outflow times concentration)

$$\tau_0 \dot{q} = f_{in} \frac{E(f_{in}, E_0)}{E_0} - f_{out}(v)q/v, \quad \quad \text{(a4)}$$

where $E(f_{in}, E_0)$ is the fraction of oxygen extracted from the inflowing blood. This is assumed to depend on oxygen delivery and is consequently flow-dependent. A reasonable approximation for a wide range of transport conditions is (Buxton *et al.,* 1998)

$$E(f_{in}, E_0) = 1 - (1 - E_0)^{1/f_{in}}. \quad \quad \text{(a5)}$$

The second term in Eq. (4) represents an important nonlinearity: the effect of flow on signal is largely determined by the inflation of the balloon, which results in an increase of $f_{out}(v)$ and clearance of deoxyhemoglobin increasing the BOLD signal. This effect depends upon the concentration of deoxyhemoglobin such that the clearance attained by the outflow will be severely

attenuated when the concentration is low (e.g., during the peak response to a prior stimulus).

In summary, the Balloon component has only three unknown parameters that determine the dynamics: resting oxygen extraction fraction(E_0), mean transit time (τ_0) and a stiffness exponent (α) specifying the flow-volume relationship of the venous balloon. The only thing required, to specify the BOLD response, is inflow.

The rCBF Component

The rCBF component relies on the generally accepted assumption that, over normal ranges, blood flow and synaptic activity are linearly related. Under the constraint that the dynamical system linking synaptic activity and rCBF is linear, the model is

$$\dot{f}_{in} = s, \tag{a6}$$

where s is some flow inducing signal. The signal is assumed to subsume many neurogenic and diffusive signal subcomponents and is generated by neuronal activity $u(t)$

$$\dot{s} = \epsilon u(t) - s/\tau_s - (f_{in} - 1)/\tau_f \tag{a7}$$

ϵ, τ_s, and τ_f are the three unknown parameters that determine the dynamics of this component of the hemodynamic model. They represent the efficacy with which stimulus-related neuronal activity causes an increase in signal, the time-constant for signal decay or elimination and the time-constant for autoregulatory feedback from blood flow. The rCBF component has only zeroth and first order kernels and therefore can not account for the hemodynamic refractoriness and other nonlinearities observed in BOLD responses.

ACKNOWLEDGMENT

REFERENCES

Burock, M. A., Buckner, R. L., Woldorff, M. G., Rosen, B. R., 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.

Buxton, R. B., Wong, E. C., and Frank, L. R. 1998. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon model. *MRM* **39:** 855–864.

Bandettini, P. A., and Cox, W. 1998. Functional contrast in event-related fMRI: Interstimulus interval dependency and blocked design comparison. *Neuroimage* **7:** S552.

Cannestra, A. F., Pouratian, N., Shomer, M. H., and Toga, A. W. 1998. Refractory periods observed by intrinsic signal and fluorescent dye imaging. *J. Neurophysiol.* **80:** 1522–1532.

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. M., and Buckner, R. L. 1997. Selective averaging of repidly presented individuals trials using fMRI. *Hum. Brain Mapp.* **5:** 329–340.

Dettmers, C., Connelly, A., Stephan, K. M., Turner, R., Friston, K. J., Frackowiak, R. S. J., and Gadian, D. G. 1996. Quantitative comparison of functional magnetic resonance imaging with positron emission tomography using a force-related paradigm. *Neuroimage* **4:** 201–209.

Friston, K. J., Josephs, O., Rees, G., and Turner, R. 1998. Nonlinear event-related responses in fMRI. *MRM* **39:** 41–52.

Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. 2000. Nonlinear responses in fMRI: The Balloon model, Volterra kernels and other hemodynamics. *Neuroimage* **12:** 466–477.

Hutton, C., Howseman, A. M., Josephs, O., Zarahn, E., Friston, K., and Turner, R. 1998. The effect of inter-stimulus interval on signal response in fMRI. *Neuroimage* **7:** S591.

Huettel, S. A., and McCarthy, G. 2000. Evidence for a refractory period in the hemodynamic response to visual stimuli as measured by MRI. *Neuroimage* **11:** 547–553.

Kutas, M., and Hillyard, S. A. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* **297:** 203–205.

Mandeville, J. B., Marota, J. J., Ayata, C., Zararchuk, G., Moskowitz, M. A., Rosen, B., and Weisskoff, R.M. 1999. Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J. Cereb. Blood Flow Metab.* **19:** 679–689.

Mechelli, A., Friston, K. J., Price, C. J. 2000. The effects of presentation rate during word and pseudoword reading: A comparison of PET and fMRI. *J. Cogn. Neurosci.* **12:** Supplement 2: 145–156.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., and Buckner, R. L. 2000. Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and possibility of ordering brain activity based on relative timing. *Neuroimage* **11:** 735–759.

Price, C. J., Wise, R. J. R., Ramsay, S., Friston, K. J., Howard, D., Patterson, K., and Frackowiak, R. S. J. 1992. Regional response differences within the human auditory cortex when listening to words. *Neurosci. Lett.* **146:** 179–182.

Puce, A., Allison, T., and McCarthy, G. 1999. Electrophysiological studies of human face perception. III: Effects of top-down processing on face-specific potentials. *Cereb. Cortex* **9:** 445–458.

Rugg, M. D., and Coles, M. G. H. 1995. In *Electrophysiology of mind,* Chapt. 1, pp. 1–17; Chapt. 6, pp. 171–188. Oxford Univ. Press, London.

Talairach, J., and Tournoux, P. 1988. *A Co-planar Stereotactic Atlas of the Human Brain.* Thieme, Stuttgart.

Veltman, D. J., Friston, K. J., Sanders, G., and Price, C. J. 2000. Regionally specific sensitivity differences in fMRI and PET: Where do they come from? *Neuroimage* **11:** 575–588.

Vazquez, A. L., and Noll, D. C. 1998. Non-linear temporal aspects of the BOLD response in fMRI. *Neuroimage* **7:** 108–118.