Published in final edited form as:
Neuroimage. 2008 November 15; 43(3): 509–520. doi:10.1016/j.neuroimage.2008.07.065.

Detection of time-varying signals in event-related fMRI designs

Jack Grinband[1,2,*], **Tor D. Wager**[3], **Martin Lindquist**[4], **Vincent P. Ferrera**[2], and **Joy Hirsch**[1]

1 *fMRI Research Center, Columbia University, New York, New York 10032*

2 *David Mahoney Center for Brain and Behavior Research, Columbia University, New York, New York 10032*

3 *Department of Psychology, Columbia University, New York, New York 10032*

4 *Department of Statistics, Columbia University, New York, New York 10032*

Abstract

In neuroimaging research on attention, cognitive control, decision-making, and other areas where response time (RT) is a critical variable, the temporal variability associated with the decision is often assumed to be inconsequential to the hemodynamic response (HDR) in rapid event-related designs. On this basis, the majority of published studies model brain activity lasting less than four seconds with brief impulses representing the onset of neural or cognitive events, which are then convolved with the hemodynamic impulse response function (HRF). However, electrophysiological studies have shown that decision-related neuronal activity is not instantaneous, but in fact, often lasts until the motor response. It is therefore possible that small differences in neural processing durations, similar to human RTs, will produce noticeable changes in the HDR, and therefore in the results of regression analyses. In this study we compare the effectiveness of traditional models that assume no temporal variance with a model that explicitly accounts for the duration of very brief epochs of neural activity. Using both simulations and fMRI data, we show that brief differences in duration are detectable, making it possible to dissociate of the effects of stimulus intensity from stimulus duration, and that optimizing the model for the type of activity being detected improves the statistical power, consistency, and interpretability of results.

Introduction

Over the past two decades, the development of functional magnetic resonance imaging (fMRI) technology has generated near exponential growth in neuroimaging research and its clinical applications (Bandettini, 2007). The most commonly used method for analyzing the blood oxygenation level-dependent (BOLD) changes in fMRI is based on the general linear model (GLM). A typical experiment consists of generating a hypothetical cognitive or neural model of brain activity and using multiple linear regression to search for voxels correlated with the predicted response. In classic block designs, the duration of each regressor in the regression model matches the duration of the stimulus block. As blocks shorten to four seconds or less, the current convention is to switch to using ‘impulse functions’ of arbitrarily short duration, rather than simply continue shortening blocks to match the length of the stimulus. While this may produce accurate results when the cognitive/neural events are of constant duration (20% of event-related studies; Fig 1B), the majority (80%) of event-related studies involves choice-

*Correspondence should be addressed to 710 W. 168th Street, Neurological Institute-B41, New York, NY 10032, (212) 342-0121, (212) 342-0855 (fax) jg2269@columbia.edu.

related neural processes that can vary in duration with the subject's RT. In 95% of event-related studies containing a decision process (Fig 1E), the duration of the decision period is assumed to be constant and is typically modeled by the convolution of a constant height, finite impulse function (i.e. a Kronecker delta function) positioned at event onset with a canonical hemodynamic response function (Friston, 2003, 1994; Henson, 2003; Josephs, 1997).

Although this method can often detect task-related fMRI activity, it makes the implicit assumption that the underlying neural or cognitive process is a brief, essentially zero-duration, event (i.e. an impulse). This simplification is generally thought to have little or no impact on the results. In fact, due to the low-pass filtering properties of the BOLD response (Zarahn et al., 1997), it has been argued that the shape of the physiological hemodynamic response (HDR) to brief stimuli (< 4 s) is equal to the theoretical hemodynamic impulse response function (HRF), making the constant impulse model a good approximation to the actual BOLD response (Henson, 2003). However, it has been shown that stimulus durations as small as 34 ms (Glover, 1999; Rosen et al., 1998; Savoy, 1995) and onset asynchronies as low as 50 ms (Bellgowan et al., 2003; Henson, 2002; Kim et al., 1997; Menon et al., 1998; Miezin et al., 2000; Richter et al., 2000) can elicit detectable BOLD responses, suggesting that small differences in the onset or duration of modeled events may be important.

If the assumption of equivalence between impulse functions and short (100 ms – 4 s) blocks (or boxcars) does not hold, then it should be possible to dissociate the effects of stimulus intensity from stimulus duration in event-related designs. Moreover, any discriminable differences between these models would suggest that the shape of the BOLD response should be optimized for the type of activity being detected. For example, impulse functions might best model neural activity at stimulus onset, whereas brief epochs might best capture activity that is sustained throughout stimulus processing. Potential differences between impulse- and short epoch-based models may be magnified by the fact that response times (RTs) in many studies vary across trials. Such variations have been shown in animal research to be related to the variations in the *duration* of decision-related neuronal firing (Janssen and Shadlen, 2005; Maimon and Assad, 2006; Ratcliff et al., 2007; Schall, 2003; Shadlen and Newsome, 2001; Snyder et al., 2006), and such variability in response time is the basis for a variety of decision-making models (Ratcliff, 2005). In addition, several fMRI studies have demonstrated that the HDR to a decision process is likely to vary with the time it takes to elicit the subject's response (Connolly et al., 2005; Formisano et al., 2002; Kruggel et al., 2000; Menon et al., 1998). Importantly, temporal variability is rarely an explicit experimental manipulation, but nevertheless exists implicitly as a distribution of response times. RTs for simple, suprathreshold detection tasks (simple RTs) typically range from 200 to 500 ms, whereas choice responses between multiple options (choice RTs) start at around 400 ms and can range up to tens of seconds depending on speed-accuracy tradeoffs, task complexity, arousal, age, clinical status, etc (Verhaeghen et al., 2006; Verhaeghen et al., 2003). Thus, decision-related behavioral responses to many types of brief stimuli—such as those elicited by attention, memory, cognitive control, language, and decision-making processes—are likely to elicit neural activity that (a) persists over much of the time between stimulus presentation and response, and (b) varies in duration from trial to trial (Fig 2A).

It has been proposed that modeling temporal variability in the data increases statistical power and captures an important source of information about the relationship between brain activity and psychophysical performance (Buchel et al., 1998; Friston, 2003; Henson, 2003; Josephs and Henson, 1999). Models that incorporate information about trial-to-trial variation in RT (or other psychophysiological parameters) into the GLM are often called 'parametric modulation' models. In the parametric (or *variable impulse)* approach, a participant's mean-centered RTs are used to modulate the amplitude of an impulse function. The modulated impulse function is then convolved with the HRF and added as an additional regressor in the GLM. Brain regions

for which the amplitude of the primary, unmodulated regressor is significantly non-zero are interpreted as task-related. Conversely, brain regions for which the amplitude of the modulated regressor is significant are interpreted as being sensitive to trial-to-trial variations in RT.

An alternative method is the *variable epoch* approach, which involves modeling each trial with a boxcar epoch function whose duration is equal to the RT of the trial. A single regressor is then constructed from these boxcars to use in the GLM. This approach makes the critical assumption that the cognitive and neural basis of decision-related activity is accurately represented by the diffusion (or race) model of decision-making (Ratcliff, 2005; Ratcliff et al., 2007). The diffusion model is supported by electrophysiological studies in humans (Philiastides et al., 2006) and non-human primates (Janssen and Shadlen, 2005; Maimon and Assad, 2006; Ratcliff et al., 2007; Schall, 2003; Shadlen and Newsome, 2001; Snyder et al., 2006) in which neuronal activity (or firing rate) is sustained or even increases up to the time of the behavioral response. Thus, compared to the *constant impulse* approach, the *variable epoch model* attempts to more faithfully represent the physiological processes related to decision-making in many brain regions. In our previous work, we used this model to locate RT-sensitive brain regions in a decision-making task and confirmed model accuracy using model-free (GLM-free) analysis methods (Grinband et al., 2006).

In the current study, the *variable epoch model* was compared against three other models: a *constant impulse model* (the most common model, used in 70% of event-related studies in our survey; Fig 1E), the *variable impulse model* that includes a mean-centered parametric modulator (11% of event-related studies), and a *constant epoch model* (a variation of the constant impulse approach in which the impulses are binned within each 2 s TR; 14% of event-related studies). This paper uses simulations and fMRI data to explore the differences in the predictions made by these models and demonstrates that, even for brief events, they are not equivalent. Our data suggest that when detecting time-varying signals, such as those generated by a behavioral response, the variable epoch model is physiologically plausible, and has higher power and reliability for detecting brain activation.

Materials and Methods

Analysis of published methods

To determine how often GLM analyses incorporated RT into decision-related regressors, we surveyed all fMRI studies from Jan 1, 2007 to May 30, 2007 published in the following journals: Human Brain Mapping, Nature, Nature Neuroscience, Neuroimage, Neuron, and Science. A total of 170 articles were assessed. Only articles reporting results of original fMRI research were included. A summary of these studies is presented in Fig 1. We characterized the image analysis methods used in each study along six dimensions: use of block vs. event-related designs, inferences about temporal processing, measurement and modeling of RT, impulse- vs. epoch-based modeling, the use of the canonical HRF, and the software package used.

For block vs. event-related designs (Fig 1A), studies that contained both a block and an event-related component were labeled as “event-related.” For constant vs. variable duration designs (Fig 1B), we identified studies that did not modulate durations. These included studies where the stimulus duration was constant, studies where the subject was required to perform a task for a constant duration, or studies where decisions were made but no inferences about decision-related activity was made in the conclusions of the paper. The variable duration designs included studies that made inferences about decision-related activity or studies that explicitly manipulated stimulus duration.

For measurement and modeling of RT (Fig 1C), in cases where the use of RT was not mentioned, it was assumed that RT was neither collected nor included in the regression model.

The percent of studies in which RT was measured/modeled, was calculated by taking the number of studies in which the RT was measured/modeled and dividing this value by the number of studies in which a response was required *and* conclusions were made about the response-related brain activity. To determine whether the length of RT affected the likelihood of incorporating RT information into the model, we recorded the mean RT of the condition with the longest RT. For the use of canonical HRF functions (Fig 1D), it was assumed that the canonical HRF was used unless otherwise specified. For the impulse- vs. epoch-based modeling (Fig 1E), the variable impulse models were defined as those models in which the modulator was a regressor of interest and was used to detect brain activity. Thus, impulse models that only included modulated "confound" regressors (such as those for head motion, respiration, or cardiac-related artifacts) were labeled as constant impulse models. In instances where the nature of the model was unclear (n = 8), the corresponding authors of the papers were contacted. For the software package used (Fig 1F), in cases where multiple software packages were used to process data, only the package used to perform the statistical analysis was included in the frequency calculations.

Simulations

To test the efficacy of the four regression models to detect time-varying brain activity, we simulated a cognitive process that varied in duration across trials and tested model performance by measuring power and false positive rate (FPR). Simulations were performed assuming a linear, time-invariant (LTI) system (Boynton et al., 1996). While nonlinearities are known to exist (Birn et al., 2001; Huettel and McCarthy, 2000; Miller et al., 2001; Vazquez and Noll, 1998; Wager et al., 2005), the LTI system was adequate for exploring the differences between linear models commonly applied to fMRI data. The incorporation of nonlinear effects is a further potential refinement of regression models that is outside the scope of the current paper.

The power simulation consisted of four steps: *(1)* a simulated neuronal time series was created consisting of a series of boxcars with randomly generated durations; *(2)* this neural model was convolved with a canonical HRF to create the simulated BOLD time series; *(3)* AR(1) noise was added to the time series; *(4)* linear regression was performed between the simulated data and each of the four regression models. This process was repeated 10000 times to compute the percent of true positives detected by each model. The false positive simulation consisted only of steps *(3)* and *(4)* to compute the percent of false positives detected by each model.

Creation of simulated cognitive/neural events—The duration and inter-trial variability of the cognitive process to be detected is an important variable that may impact the choice of model. Therefore, we created simulated RTs based on RT distributions from our previously published empirical study of a two-alternative, forced-choice categorization task (Grinband et al., 2006). In the task, 10 subjects categorized line segments as "long" or "short." We fit a gamma function to the RT distributions of each of 10 subjects (Fig S1A) and averaged the gamma parameters to generate a mean gamma distribution ($\alpha = 1.7$, $\beta = 0.4$, minimum value = 0.5; Fig S1B) with RT mean = 0.84 s and s.d. = 0.64 s. Simulated neural process durations were randomly drawn from the resulting gamma distribution. Thus, the simulated neural events consisted of a series of boxcars whose durations are distributed similarly to observed RTs in a simple perceptual decision-making task.

Variable inter-event intervals were selected from a uniform distribution with a minimum of 4 s and a maximum of 7 s, values commonly used in fMRI experiments. The total duration of each simulated run was 5.5 min. To simulate the autocorrelation present in real fMRI time series, we added AR(1) noise to each time series with phi = 0.3.

Linear models tested—The four models used to detect this neural process are illustrated in Fig 2B. The variable epoch model was created using the same variable-length boxcars and, consequently, must necessarily fit better than the other models. However, the question of interest is not which model fits better, but whether there are any appreciable differences between the models in their ability to detect brain activity. Our null hypothesis states that: *For brief (< 4 s) stimulus durations, there are no significant differences between models that assume a constant shape of the HDR (i.e. ignore differences in duration of neural processes) and those that explicitly account for the duration differences between trials.* For the variable impulse and constant impulse models, the model's representation of the neural process was generated by placing a 50 ms epoch at the onset of the process. The constant epoch model assumes that the temporal resolution of the neural process and the BOLD response is equal to the TR and, thus, consists of short epochs equal to the inter-scan interval (TR = 2 s). It was constructed by positioning the 2 s epoch on the TR interval closest to the onset of the neural process and convolving with the HRF. This is in contrast to the constant impulse model, which assumes that the temporal resolution of the BOLD response is equal to the TR but the resolution of the neural process is 50 ms. The constant regressors (impulse and epoch) had amplitudes equal to 1. The modulator regressor was created by mean-centering the neural durations, normalizing the range of durations to ±0.5, and setting the amplitude of each impulse equal to the corresponding normalized duration.

Performance metrics—The performance metrics of interest were true positive rate (power) and false positive rate (FPR). We evaluated power and FPR at different effect sizes using the Pearson correlation coefficient between the model and the data. To compute power, 10,000 simulated data time series of signal + AR(1) noise were generated for each effect size and the fraction of true positives was calculated for each model type. The signal component of each time series consisted of a random sample of gamma distributed RTs and uniformly distributed inter-trial intervals. To compute FPR, we generated 10,000 AR(1) noise time series and the fraction of positive results was calculated for each model type.

To compare the relative contribution of mismodeling shape versus mismodeling amplitude, we computed the effect of shape differences in the variable amplitude case by measuring the amount of variance explained by the impulse model when each trial was fit independently of all other trials. In this case, the best amplitude fit is found using an incorrect shape for each trial (i.e. the canonical impulse response). The degree of error for each trial is determined by the duration of the epoch – longer durations have greater deviation from the canonical response and therefore a larger error. The mean error is determined as a weighted mean of the RT distribution, $\Gamma(\alpha = 1.7, \beta = 0.4$, min value = 0.5, mean = 0.84 s, s.d. = 0.64 s). Although, shape and amplitude are normally coupled, this procedure allows us to estimate the effect of mismodeling amplitude independently from the effect of mismodeling the shape of the HDR.

Visual stimulation experiments

The goal of this study was to determine how model selection affects the ability to detect time-varying BOLD responses. We needed a task that could modulate the duration of a neural process in a precise and reproducible way. However, although an actual decision-making task would introduce temporal variability into the BOLD signal, it would also generate unknown sources of variability related to the cognitive aspects of the task. These unknown cognitive effects would be impossible to model or dissociate from the temporal variability. Thus, to isolate the effect of the temporal variance from other sources, we conducted two flashing checkerboard experiments. Because the flashing checkerboard stimulus primarily activates visual cortex, an area with known response characteristics, it gave us precise and reproducible control over the duration of a neural process. Moreover, the resulting visual activation is known to be sustained throughout visual stimulation (though it is often strongest at the onset of a

stimulus (Logothetis et al., 2001) in the same way as decision-related neuronal activity (Janssen and Shadlen, 2005; Maimon and Assad, 2006; Ratcliff et al., 2007; Schall, 2003; Shadlen and Newsome, 2001; Snyder et al., 2006). We were then able to use the time-varying, visually-evoked responses as a proxy for time-varying, decision-related responses. Thus, using passive visual stimuli instead of a decision-making task allowed us both to (1) isolate the effects of temporal variability in the BOLD response from the confounding effects of decision-related brain activity, and to (2) determine whether the simulations and the actual fMRI results lead to similar conclusions.

Experiment 1 – Dissociating signal intensity from duration—To determine whether differences in HDR shape due to stimulus intensity and stimulus duration are detectable and dissociable at the level of an individual, one subject viewed flashing checkerboards (7.5 Hz) of variable contrast intensity and duration. The stimuli either varied in contrast (5%, 10%, 20%, and 40%) while maintaining a constant duration (0.25 s) or in duration (0.25 s, 0.75 s, 1.3 s, 3.5 s) with a constant contrast intensity (5%). All eight experimental conditions were randomly intermixed. Each trial type was presented twice per run. During the inter-trial interval, a fixation point was presented against a gray background. The subject was scanned for 6 runs of 6 min 48 s each.

Experiment 2 – Sensitivity and consistency of GLM analysis—Eight subjects viewed flashing checkerboards of constant contrast intensity (20%) but variable duration to test the effects of model selection on sensitivity, consistency, and false positive rate across individuals. The duration of each checkerboard stimulus was randomly drawn from the same mean gamma distribution (Fig S1B) used in the simulations, which is typical of human choice RT variability. The inter-trial interval was randomly jittered using a uniform distribution with a minimum of 4 s and a maximum of 7 s. During rest, subjects viewed a fixation point against a gray background. Each subject was scanned for 5 runs of 5 min 30 s each.

Image acquisition—Imaging experiments were conducted using a 1.5T GE TwinSpeed Scanner using a standard GE birdcage head coil. Structural scans were performed using the 3D SPGR sequence (124 slices; 256 × 256; FOV = 200 mm). Functional scans for Experiment 1 were performed using EPI-BOLD (TE = 39; TR = 1.0 sec; 13 slices; 64 × 64; FOV = 200 mm; voxel size = 3 mm × 3 mm × 4.5 mm). Functional scans for Experiment 2 were performed using EPI-BOLD (TE = 60; TR = 2.0 sec; 29 slices; 64 × 64; FOV = 200 mm; voxel size = 3 mm × 3 mm × 4.5 mm). All image analysis was done using the FMRIB Software Library (FSL; http://www.fmrib.ox.ac.uk/fsl/) and Matlab (Mathworks, Natick, MA; http://www.mathworks.com). The data were motion corrected (FSL-MCFLIRT), high-pass filtered (at 0.02 Hz), and spatially smoothed (full width at half maximum = 5 mm).

Image analysis—Standard statistical parametric mapping techniques (FSL-FEAT) were performed prior to registration to MNI152 space (linear template). Multiple linear regression was used to identify voxels that correlated with specific sensory events (i.e. flashing checkerboards). A primary statistical threshold for activation was set at $p = 0.01$. Since our goal was to evaluate the absolute number of voxels detected by the different models and because all comparisons were made between different models for the same data set, no correction for multiple comparisons was made. Inter-subject group analyses were performed in standard MNI152 space by applying the FSL-FLIRT registration transformation matrices to the parameter estimates. For each run, the transformation matrices were created by registering via mutual information (1) the midpoint volume to the first volume using 6 degrees of freedom, (2) the first volume to the SPGR structural image using 6 degrees of freedom, and (3) the SPGR to the MNI152 template using 12 degrees of freedom. These three matrices were concatenated and applied to each statistical image. Ventricular masks were created using FSL-FAST by first

segmenting each high-resolution brain into three tissue types: gray matter, white matter, and CSF. Then the CSF partial volume maps were transformed into the subject's functional space of each individual run and thresholded at 0.95 to ensure that no more than 5% of the volume of each voxel in the mask contained gray or white matter.

Consistency measurements—The quality of each model can be evaluated by its ability to consistently detect the same pattern of activated voxels for a given stimulus. We used Cronbach's alpha (Cronbach, 1951), a measure of inter-subject consistency, to calculate the mean correlation between the spatial activation patterns of the occipital cortex across the five runs. Alpha was calculated for each subject as

$$\alpha = \frac{N\bar{r}}{(N-1)\bar{r}+1} \tag{1}$$

where N is the number of runs (N = 5) and $\bar{r}$; is the average of all pairwise Pearson correlation coefficients between the Z-statistic maps (across voxels). To test for significant differences between models the alphas were normalized using Fisher's Z transform:

$$Z_\alpha = 0.5\log\left(\frac{1+\alpha}{1-\alpha}\right) \tag{2}$$

A paired Student's t-test to the Z_α scores was used to compare the variable epoch model against each of the other models.

Variability of HRF estimates—The constant impulse model is often used to estimate the theoretical hemodynamic impulse response. We compared the quality of the impulse estimate with that of the variable epoch model by measuring the variance in the HRF estimate across runs. The HRF estimate was computed using FLOBS (Woolrich et al., 2004), a three-function basis set that restricts the parameter estimates of each basis function to generate physiologically plausible results. The two models were convolved with each of the FLOBS basis functions and an F-test on the three convolved functions was performed; for each significant voxel within the occipital cortex, the parameter estimates for the basis functions were used to reconstruct the fitted HRF shape, i.e. $HRF = PE_1 * f_1 + PE_2 * f_2 + PE_3 * f_3$, where *PE* is a parameter estimate and *f* is a FLOBS basis function. The HRF estimates were then averaged across all significant voxels in the occipital cortex. Region of interest masks of the occipital cortex were manually generated for each subject using the calcarine and parieto-occipital sulci as landmarks. To determine whether the variances of the two HRF estimates were different from each other, we used Levene's Test for Equality of Variance (Levene, 1960):

$$W = \sum \frac{(N-k)\sum_i n_i(\bar{z}_{i\bullet} - \bar{z}_{\bullet\bullet})^2}{(k-1)\sum_i\sum_j(\bar{z}_{ij} - \bar{z}_{j\bullet})^2} \tag{3}$$

where N is the total number of observations, n_i is the set of observations within group i, and k is the number of groups, $z_{ij} = |x_{ij} - \bar{x}_i|$ is the absolute deviation from within group means, $\bar{z}_i = \sum_j z_{ij}/n_i$ is the average absolute deviation from the group mean, and $\bar{z}_{\bullet\bullet} = \sum_i\sum_j z_{ij}/N$. Here, N = 16, and k = 2.

Results

Results

Of 170 fMRI studies, 48% were blocked and 44% were event-related; the remaining 8% were not easily classifiable (Fig 1A). Stimulus or response duration was important in 80% of event-related designs (Fig 1B). The remaining 20% involved tasks that maintained constant stimulus/response durations (for example, primary sensory or primary motor-related studies) or did not make inferences about decision-related brain activity. In event-related studies in which decision-making was important, RT was measured 82% of the time but modeled only 9% of the time (Fig 1C). Furthermore, only 16% of the studies (Fig 1E) actually included RTs, in some form, in their regression model; that is, 84% of event-related studies with a decision component made the assumption that the time necessary to process a stimulus or to generate a response was constant for all trials and trial types. Similarly, only 4% of event-related studies (Fig 1D) estimated individual HRFs for each subject; 96% assumed no differences in HRF shape existed between subjects.

The majority of event-related studies with a decision component (69%) used constant impulses convolved with the canonical HRF to represent the decision events (Fig 1E); 11% used the variable impulse model to account for parametric modulations in their design, 14% used the constant epoch model, and 5% used the variable epoch approach. Moreover, the large majority (95%) of event-related designs assumed that there were no significant differences in the shape of the HDR between trials or trial types; 84% assumed there were no significant differences in either shape or intensity of the HDR. The mean RT of studies that incorporated temporal information was 1270 ms (s.d. = 727 ms, min = 680 ms, max = 2560 ms). The mean RT for studies that did not incorporate temporal information was 1036 ms (s.d. = 529 ms, min = 256 ms, max = 2100 ms). There was no significant difference between the two groups (T-test, $p = 0.35$). There was a broad distribution of software packages used with no apparent systematic differences in how regressors were modeled between fMRI analysis packages (Fig 1F).

Simulations

We tested the null hypothesis that the constant epoch and the two impulse models were as effective as the variable epoch model at detecting brain activity that varies in duration from trial to trial (i.e. the type of activity generated in decision-making experiments). The effectiveness of the four regression models to detect neural activity was computed by comparing the fraction of true positives (i.e. power) and false positives for each model. Fig 3A shows that the variable epoch model has greater statistical power for detecting time-varying neural activity. This was necessarily the case, as the simulated true response was identical to the regressor used in the variable epoch model. However, the critical issue was whether the variable epoch model produced *significantly* better results than the other models in the presence of physiological noise. At a correlation coefficient of 0.1 (a plausible value for a robust fMRI response; Fig 3A), the power of the variable epoch model was 0.55, while the constant impulse model was only 0.23 — a 58% reduction in power. Furthermore, the modulator regressor from the variable impulse model had the lowest power of any of the regressors (power = 0.18), a 67% reduction compared to the variable epoch model, suggesting that impulse modulation did not provide an adequate model to capture activity related to variable-duration events. The constant epoch model performed only slightly better than the constant impulse model (power = 0.28), a 49% reduction compared to the variable epoch model. The results were reversed when the variable epoch model was used to detect constant duration activity; that is, the constant impulse model outperformed the variable epoch model when detecting a constant neural process (Fig S2). The false positive rate was controlled appropriately at $p = 0.05$ in all analyses and was not affected by model type (Fig 3B). These data suggest that for event-related designs, the type of model matters and that commonly used approximations, which ignore

durations of events, are less able to detect time-varying BOLD activity than has been previously appreciated.

Since the variable impulse model has one more regressor than the other models, it is more flexible at fitting the data than the single regressor models. We used an F-test to determine if the two-regressor variable impulse model had more detection power than the one-regressor variable epoch model. The variable epoch model had higher power than the variable impulse model despite having one less degree of freedom (Fig 3C). The false positive rate was controlled appropriately at p= 0.05 (Fig 3D).

These simulations show that even for events that have a mean duration of less than 1.0 s, the procedure used to model trial-to-trial variability has a substantial impact on statistical power. This is also true for durations typical of simple RTs, which have mean durations significantly smaller than choice RTs (though the effect diminishes as RTs become smaller; Fig S3). Two main factors explain these results. First, the constant impulse and constant epoch models assume constant HDR amplitudes for all trials. While a seemingly reasonable simplification, our results show that ignoring trial-to-trial variations even for brief events results in substantial mismodeling. Secondly, the impulse models (constant and variable) assume that the duration of the neural activity is not different from zero for brief events. Thus, although the variable impulse model allows the amplitude of the HDR to be modulated in height, the shape of the HDR for all trials in the two models is required to remain constant. Our results suggest that trial-to-trial variations in duration are not fully captured by modulation of amplitude.

Fig S4 and Fig 4 illustrate in more detail why the models are not equivalent. Fig S4 shows the size of the difference between the predicted and actual response when modeling a variable-duration neural process (Fig S4A) as if it was a constant impulse (Fig S4B); the two models make quite different predictions about the shape and amplitude of the fMRI time series (Fig S4C). Fig 4 describes the differences in the shape of the HDR during changes in stimulus amplitude (red) and stimulus duration (blue). Fig 4A shows predicted responses after convolution with a canonical HRF for events ranging in amplitude and duration from 0 to 4000 ms (in steps of 500 ms). The graph shows that modulations in stimulus amplitude and duration produce divergent responses—even after convolving very brief events with a canonical HRF. For a stimulus duration of 1.0 s, the correlation between the impulse response and the epoch response is $R^2 = 0.92$, whereas for a stimulus duration of 3.0 s, the correlation is much lower, $R^2 = 0.54$ (Fig 4B).

To compute the relative contribution of mismodeling shape versus amplitude of the HDR, we computed the mean percent variance explained by the impulse model across trials when the durations are gamma distributed in the same way as our RT distribution. In our simulations, mismodeling of shape accounted for 12% of the mismodeling effect; mismodeling of amplitude accounted for the other 88%.

Imaging

Based on our simulations, we developed several predictions that we tested using fMRI. First, if the convolution of neural activity with a HRF is an accurate model of HDRs, then duration-modulation and amplitude-modulation should generate different HDR shapes (Fig 4A). Specifically, amplitude modulation should vary the rise time of the HDR, with more intense stimuli resulting in more rapid signal increases and greater evoked amplitudes. Duration-modulation, by contrast, should result in a quickly saturating (constant) rise-time, but a linearly increasing time-to-peak for the HDR. This hypothesis was tested in Experiment 1. Second, the variable epoch model should be more powerful than other models at detecting variable-duration, visually-evoked activity in the presence of physiological noise, resulting in a significantly greater number of suprathreshold voxels in occipital cortex. Our third and fourth

predictions were that the variable epoch model should result in activation patterns with higher reliability across runs and lower inter-subject variability than other models. These predictions were tested in Experiment 2.

Experiment 1 – Dissociating signal intensity from duration—As the stimulus contrast increased, the amplitude of the response increased (Fig 4C, red). More importantly, as the model in Fig 4A predicts, the initial slope increased linearly (Fig 4D, red; linear regression, slope = 5.8×10^{-5}, p = 0.0050, intercept = 0.0011, p = 0.0093, df = 52) but the time to reach maximum BOLD response did not change (Fig 4E, red; linear regression, slope = −0.0059, p = 0.70, intercept = 3.2, p = 2×10^{-13}, df = 47). The reverse pattern was evident when increasing stimulus duration. Although the amplitude increased with stimulus duration (Fig 4C, blue; polynomial regression, quadratic term = −0.0010, p = 0.0037, linear term = 0.0044, p = 0.0006, intercept = 0.0027, p = 0.84, df = 47), the initial slope reached a constant value at a stimulus duration of 1.3 s, but showed a linearly increasing time of peak response (Fig 4E, blue; linear regression, slope = 0.70, p = 0.00032, intercept = 2.9, p = 1×10^{-14}, df = 50).

Experiment 2 – Sensitivity and consistency of GLM analysis—We compared the performance of each of the regression models to detect visually evoked responses in occipital cortex by comparing the number of significant voxels generated by each model. We also compared the consistency of the results generated by each model using three measures: (1) peak Z-score across runs, (2) reliability of the spatial activation pattern across runs (assessed with Cronbach's alpha), and (3) inter-subject variability of the estimated HRF (assessed with Levene's Test). We predicted that the variable epoch model would have significantly better fits to the data, a more consistent distribution of active voxels, and lower variance in the HRF estimate.

Figure 4A shows that the variable epoch model was able to detect more active voxels than the other models across a range of thresholds. We counted the number of significant voxels detected within the visual cortex for each functional run and tested whether the epoch model detected significantly more active voxels (paired t-test, significant at $p < 0.05$, df = 7). At a voxel detection threshold of p = 0.01, the variable epoch model detected an average of 525 (s.e. = 84) active voxels per run, compared to only 383 (s.e. = 83, p = 0.0005) voxels for the constant impulse model (Fig 5A). This comprises a 27% decrease in the number of detected voxels. The constant epoch model detected 31% fewer voxels than the variable epoch model (mean detected = 361, s.e. = 78, p = 0.0008). The constant impulse regressor and modulator generated 406 (s.e. = 86, p = 0.0006) and 250 (s.e. = 68, p = 0.0005) activated voxels, respectively. The variable epoch model also detected more voxels than the F-test of the variable impulse model (mean detected = 380, s.e. = 56, p = 0.0014, 28% reduction; Fig S5). Thus, for individual functional runs, the variable epoch model detected significantly more active voxels than all other models (including the F-test of the variable impulse regressors) at all detection threshold levels (p-values).

The improvement in performance of the variable epoch model also generalized to the group level, across subjects. At a threshold of p = 0.01, the variable epoch model detected 8793 significant voxels compared to only 6520 voxels (26% decrease) for the constant impulse model and 5694 voxels (38% decrease) for the constant epoch model (Fig 5B). The largest decrease in power was demonstrated by the two-regressor, variable impulse model; the constant impulse and modulator regressors generated only 147 and 473 significantly activated voxels (Fig 5B). The corresponding group activation maps (thresholded at p = 0.01 and unthresholded) are illustrated in Fig 5C. This large reduction in sensitivity of the variable impulse model at the group level is due to the decreased consistency of the spatial activation pattern at the individual level for the variable impulse model when compared with the variable epoch model

(see Cronbach's alpha below). The primary reason for the decreased consistency at the individual level is that the variable impulse model attempts to account for a single source of temporal variance as if it were two independent sources: a constant intensity, zero duration component and a zero duration, variable intensity component with intensity proportional to duration. These two arbitrary transformations produce regressors that do not closely match the data, resulting in low power for each regressor.

To test for differences in the false positive rate between the models, we counted the number of significant voxels detected within the ventricles. Ventricular masks were created by thresholding the partial volume maps such that gray and/or white matter accounted for no more than 5% of the volume of each CSF voxel. The number of active voxels within each mask was counted and normalized by the total CSF volume. There were no significant differences between the variable epoch regressor and any of the other regressors (paired t-test, $p = 0.58$ for constant epoch model, $p > 0.93$ for all others). However, the F-test for the variable impulse model detected 40.8% more significant voxels in the ventricles and surrounding CSF than the variable epoch model (paired t-test, $p = 0.039$).

Since the "true" activation pattern is unknown, it is possible that the greater number of significant voxels detected by the variable epoch model may be due to factors other than greater model detection power. An additional test of model quality is the degree of consistency in the detected brain activity. We evaluated the consistency of the brain activation generated by each model in three ways: (1) significance of activation, (2) Cronbach's alpha, and (3) variability of the HRF estimate.

The variable epoch model generated higher statistical significance values (Z-scores) compared to the other models. The mean peak Z-score across runs was determined for the visual cortex of each subject. The mean peak Z-score was significantly higher for the epoch model using a paired t-test ($p < 0.05$, $df = 7$; variable epoch model: $\mu = 6.4$, $\sigma = 1.7$; constant epoch model: $\mu = 5.4$, $\sigma = 1.5$, $p = 0.0007$; constant impulse model: $\mu = 5.6$, $\sigma = 1.5$, $p = 0.019$; variable impulse, constant regressor: $\mu = 5.8$, $\sigma = 1.7$, $p = 0.0030$; variable impulse, modulator: $\mu = 4.8$, $\sigma = 0.5$, $p = 0.0051$), indicating that the variable epoch model explains a greater proportion of the variance than the other models. The difference in mean peak Z-scores also extended to the group activation map (variable epoch model: $\mu = 5.7$; constant epoch model: $\mu = 4.9$; constant impulse model: $\mu = 4.8$; variable impulse, constant regressor: $\mu = 2.5$; variable impulse, modulator: $\mu = 2.6$).

To evaluate the consistency of the spatial activation pattern across runs for each subject, we compared Cronbach's alpha (Cronbach, 1951) of the unthresholded Z-scores. Within the occipital cortex, the variable epoch model generated a more consistent spatial pattern of activation ($\alpha = 0.74$) than the other models (constant impulse $\alpha = 0.62$; constant epoch $\alpha = 0.56$; variable impulse, constant regressor $\alpha = 0.64$, modulator $\alpha = 0.59$). Higher alpha values for the variable epoch model, as compared with the other models, were significant at $p < 0.05$ using a paired t-test (Table S1).

As a final measure of consistency, we compared the variance of the estimated HRF. The constant impulse model is often used to estimate a custom HRF for individual subjects. It is commonly assumed that the theoretical HRF is equal to the measured HDR. However, as was demonstrated in Fig 4, the HRF and HDR are equal only when the stimulus duration is close to zero. To determine the effect of estimating the HRF by treating time-varying signals as impulses, we used FLOBS, a basis set constructed to generate plausible HRF shapes (Woolrich et al., 2004). Fig 6A shows that the variance of the estimated HRF using the constant impulse model (red) is dramatically larger than the variable epoch model (blue). Fig 6B shows that this

difference in variance was significant for most of the time points using Levene's Test for Equality of Variance (Levene, 1960).

Discussion

In fMRI studies that use block designs, the regressors are almost universally constructed as boxcar functions with block durations equal to the duration of the stimulus. Regressors that are designed in this way are meant to detect neural activity with onset and offset times that match the stimulus. As blocks shorten to 4 s or less, the convention in the field is to switch to using impulse functions, rather than to continue shortening blocks to match the length of the stimulus. The resulting regressor assumes that the hemodynamic impulse response function (HRF) and the hemodynamic response (HDR) are equal, and that every trial produces an identical BOLD response. Although the variable epoch method is rarely used (Fig 1E), it has several advantages. First, the variable epoch model provides higher detection power for neural activity whose durations vary with a known psychophysical parameter (such as RT). Second, in such regions, the variable epoch method generates higher Z statistics, more reliable patterns of activation, and HRF estimates with lower inter-subject variability. Finally, the variable epoch model is a more physiologically plausible representation of decision-related activity — neuronal activity bursts have an appreciable (non-zero) duration that in many cases covaries with response time.

Convolving variable duration epochs with a canonical HRF results in HDRs that are different from those generated by convolving impulses (Fig 4 & S3). A critical issue is whether this theoretical difference is detectable under conditions of physiological noise. It has been argued that the HDR to a short duration neural process is not appreciably different from the HRF (Henson, 2003) due to high hysteresis (temporal smoothness) in the BOLD response (Zarahn et al., 1997). However, accounting for inter-subject variability in the shape of the HRF by using a subject's own HRF in the regression model produces more robust statistical parametric maps than using the canonical HRF (Aguirre et al., 1998; Handwerker et al., 2004). Furthermore, mismodeling of the HRF, as well as temporal mismatches between the neural activity and the regression model, can result in significant decrements in power (Hernandez et al., 2002). These results suggest that differences in the shape of the HDR are detectable and important even under conditions of high auto-correlated noise. Our imaging data demonstrate that even for durations as brief as a few hundred milliseconds, modulation of intensity or duration results in different shapes of the BOLD response (Fig 4). Explicitly modeling these differences increases statistical power (Fig 3 & S2) and can result in dramatic increases in the number of voxels detected (Fig 5).

Another weakness of the constant impulse model is its prediction that the size of the HDR remains constant across events. While this may be a good model for passively viewed stimuli of equal durations, it is unlikely to generate optimal results when a response is required from the subject and when psychophysical measures can be incorporated into the regression model. Even for simple reaction-time tasks in which a subject presses a button in response to a signal onset, response times can vary by hundreds of milliseconds (Menon et al., 1998; Verhaeghen et al., 2006; Verhaeghen et al., 2003). Recordings from single neurons have shown that the time taken to make a response depends on neural processing times in decision-related brain regions (Janssen and Shadlen, 2005; Maimon and Assad, 2006; Ratcliff et al., 2007; Schall, 2003; Shadlen and Newsome, 2001; Snyder et al., 2006). Furthermore, positive correlations between decision time and the onset of the BOLD response have been demonstrated using fMRI (Connolly et al., 2005; Formisano et al., 2002; Kruggel et al., 2000; Menon et al., 1998). Thus, since the time necessary to perform a mental operation has variable, finite duration and is related to neural processing time, it is not surprising that time-varying decision-related

activity may be better detected by a RT-related, variable epoch regressor than a non-varying, zero-duration impulse regressor.

Studies that account for decision time usually do so by modulating the height of the impulse response (Buchel et al., 1998; Friston, 2003; Henson, 2003; Josephs and Henson, 1999). If the true neural activity varies in time, the variable impulse model significantly underperforms the variable epoch model when used for signal detection despite the extra degree of freedom (Fig 5 & Fig S5). Thus, the additional flexibility provided by the modulator is not sufficient to accurately represent time-varying neural activity.

It is important to point out that response time is merely an estimate of the time necessary to make a decision. Specifically, it provides an upper limit on the duration of the neural activity that is involved in forming the response. It may be possible to generate more precise bounds on the decision period by using psychophysical models or electrophysiological recording techniques. For example, RT tasks can provide estimates of the durations of the sensory, motor, and choice components of the decision (Posner, 1978; Sternberg, 2001) whereas EEG can provide estimates of different neural components that can then be included in the regression model (Gerson et al., 2005; Goldman et al., 2002; Goncalves et al., 2006; Osman et al., 1992).

Alternative Methods of Modeling fMRI Data

Although the variable epoch model outperforms the variable impulse model when detecting known time-varying signals, the flexibility of the variable impulse model could potentially allow a better fit to the data when the nature of the time-varying signal is unknown. In fact, flexibility can be maximized by adding a series of orthogonal basis functions to the regression design matrix. Deciding whether to use a single, variable epoch regressor or multiple, orthogonal regressors for modeling time-varying signals is determined by the specific aims of the study.

There are two common aims when performing regression analysis: prediction and detection. Some studies are interested in developing models that can make accurate predictions of the brain's response. Because a predictive model is not known *a priori*, it has to be determined from the data itself using an optimized set of basis functions. Optimized basis functions are typically orthogonalized, but, as a result, do not necessarily represent physiologically plausible neural responses (e.g. a series of FIR filters, or sinusoids, or a basis set created from the principal components of HDRs such as FLOBS).

In other studies, an *a priori* cognitive or neural model is assumed to be true and a set of basis functions is created that represents the expected physiological response in the brain. In this case, the regression model is used to detect voxels that have a similar temporal pattern as predicted by the cognitive/neural model. Such voxels are said to be involved in the computation of the cognitive or neural process that generated the model. Importantly, such models are not orthogonalized because it is important to maintain equivalence between the cognitive/neural model and the regression model in order to provide explanatory power. However, it is possible to add 'nuisance' regressors that reduce the residual, but that are not used for inference and that do not affect the explanatory power of the cognitive/neural model. For example, the main regressor of interest is often orthogonalized with respect to some other variable, for example, head motion parameters to account for non-linear effects of head motion, temporal derivatives to account for constant temporal offsets of the model, or even a constant epoch regressor to improve the specificity of a variable epoch regressor.

Moreover, when regression is used for signal detection, the shape of each response-related regressor should match the predicted response-related neural activity. Thus, for certain cases, the impulse model may be optimal; in fact, the detection of constant duration activity

requires a regressor consisting of boxcars or impulses of constant duration. For example, stimulus onsets and offsets, as well as simple motor responses, are best modeled by impulse functions. A passively observed two-second tone or a three-second sinusoidal grating is best modeled by a two- and three-second constant epoch, respectively. However, the majority of event-related studies make inferences about decision-making activity (Fig 1B). Since decision processes have variable duration, incorporating estimates of RT (or other temporal measures of the decision process) into fMRI analyses using the variable epoch approach can produce improvements in power, reliability, and interpretability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Ted Yanagihara, Peter Freed, Arno Klein, Jason Steffener, Tobias Teichert, Franco Pestilli, Kristen Klemenhagen, Eric Zarahn, and Luiz Pessoa for reading the manuscript and providing many useful comments, and Steve Dashnaw for spending many, many hours helping collect data. This research was supported by T32MH015174 (JG), T32EY013933 (JG), NSF 0631637 (TDW), NSF 0631637 (ML), MH073821 (VPF), and fMRI Research Grant (JH).

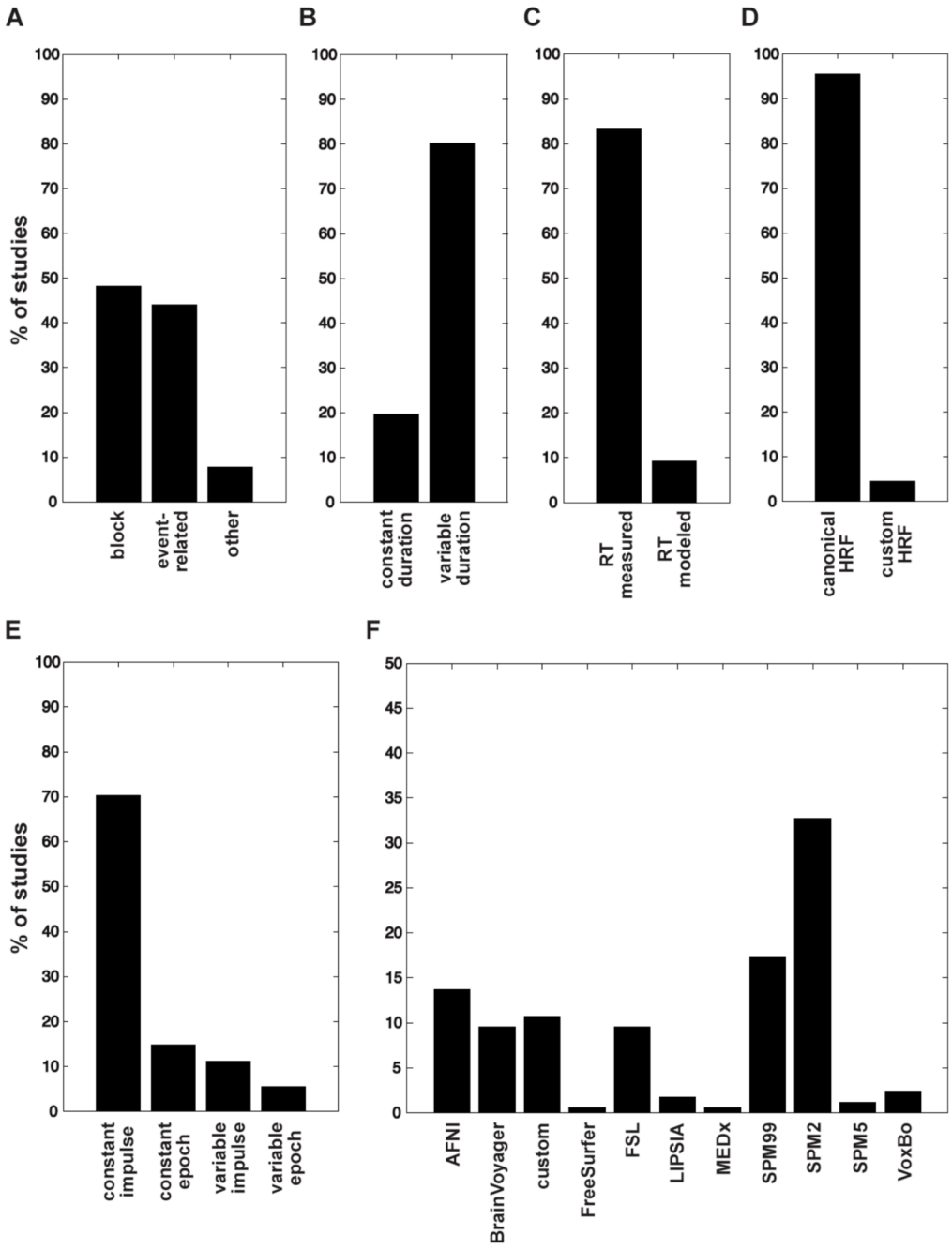

Figure 1. Survey Statistics

We surveyed 170 published fMRI studies to characterize how the GLM is used to analyze imaging data. **(A)** Block and event-related designs were equally common. **(B)** Most event-related studies made inferences about a time-varying decision process. **(C)** Although response times were recorded in 82% of event-related studies with a decision component, only 9% actually used this information to construct a regression model for detecting brain activity. **(D)** Most studies assumed that there were no significant differences in HRF shape across subjects. **(E)** In event-related fMRI studies that made inferences about variable duration decision processes, a majority (95%) assumed that the shape of HDRs did not vary across trials or trial types, and 84% assumed that both shape and intensity did not vary across trials. **(F)** All

of the major analysis platforms were represented and no obvious relationship between platform type and model type was found.

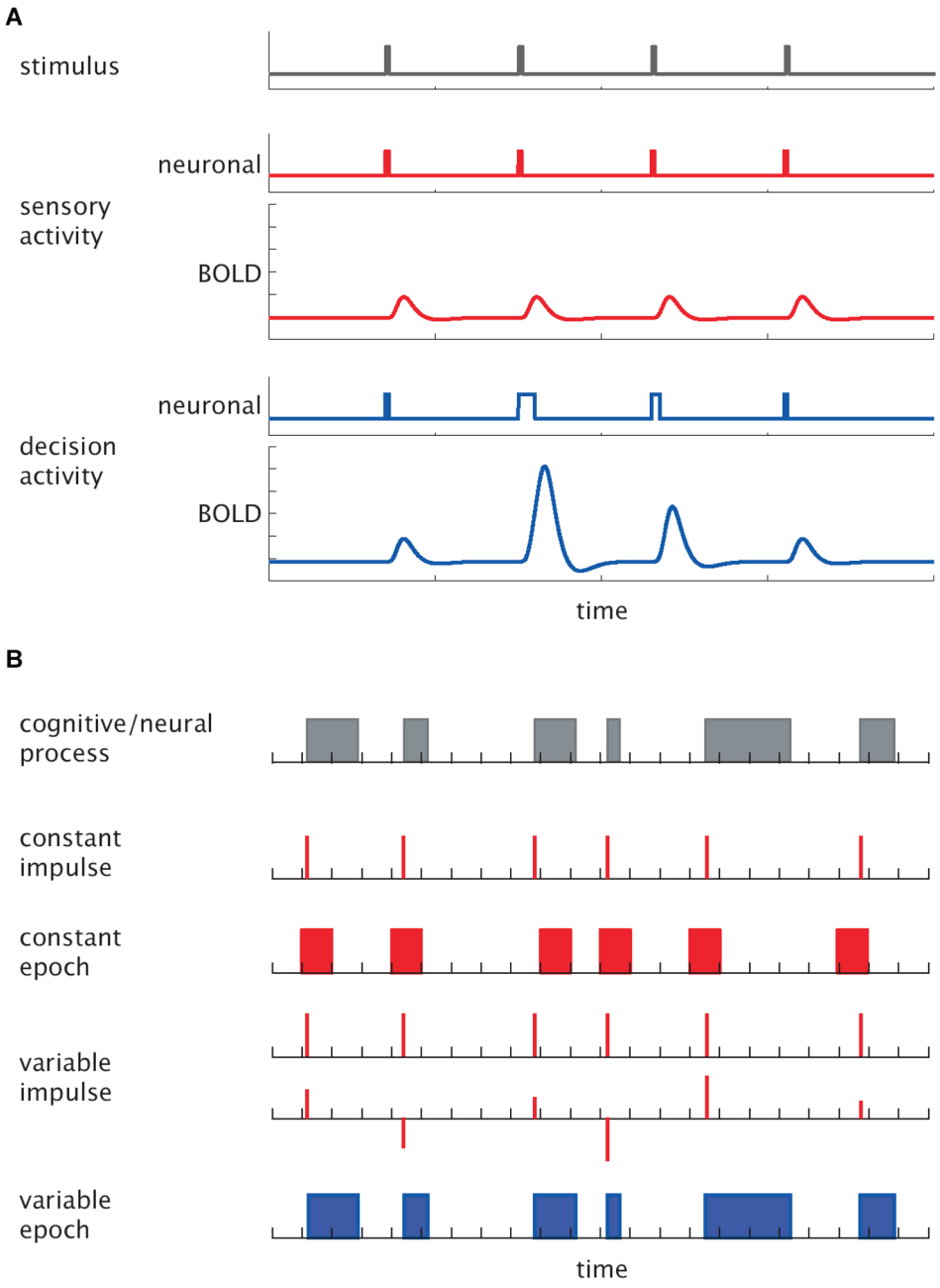

Figure 2. GLM models used in fMRI analysis

(A) A hypothetical event-related fMRI experiment in which the subject responded to the presentation of a stimulus. Assuming a linear time-invariant (LTI) system, the constant duration stimulus produces sensory neuronal responses and sensory BOLD responses that are also constant in duration. However, electrophysiological evidence (Janssen and Shadlen, 2005; Maimon and Assad, 2006; Ratcliff et al., 2007; Schall, 2003; Shadlen and Newsome, 2001; Snyder et al., 2006) shows that the same constant duration stimulus will typically produce variable duration decision-related neural responses, characterized by a subject's RT distribution. An LTI system predicts that the decision-related BOLD response will vary in both shape and intensity from trial to trial. **(B)** We tested the efficacy of four regression models against a hypothetical decision-related cognitive/neural process that varied in duration on each

trial. The *constant impulse model* consists of an impulse function positioned at the onset of each event. The *constant epoch model* consists of a 2 s epoch positioned on the TR nearest the onset of the neural event. The *variable impulse model* is a two-regressor model that uses a constant impulse regressor and a modulator regressor whose height is proportional to the demeaned durations of the process. All the models were convolved with a canonical double gamma hemodynamic response function.

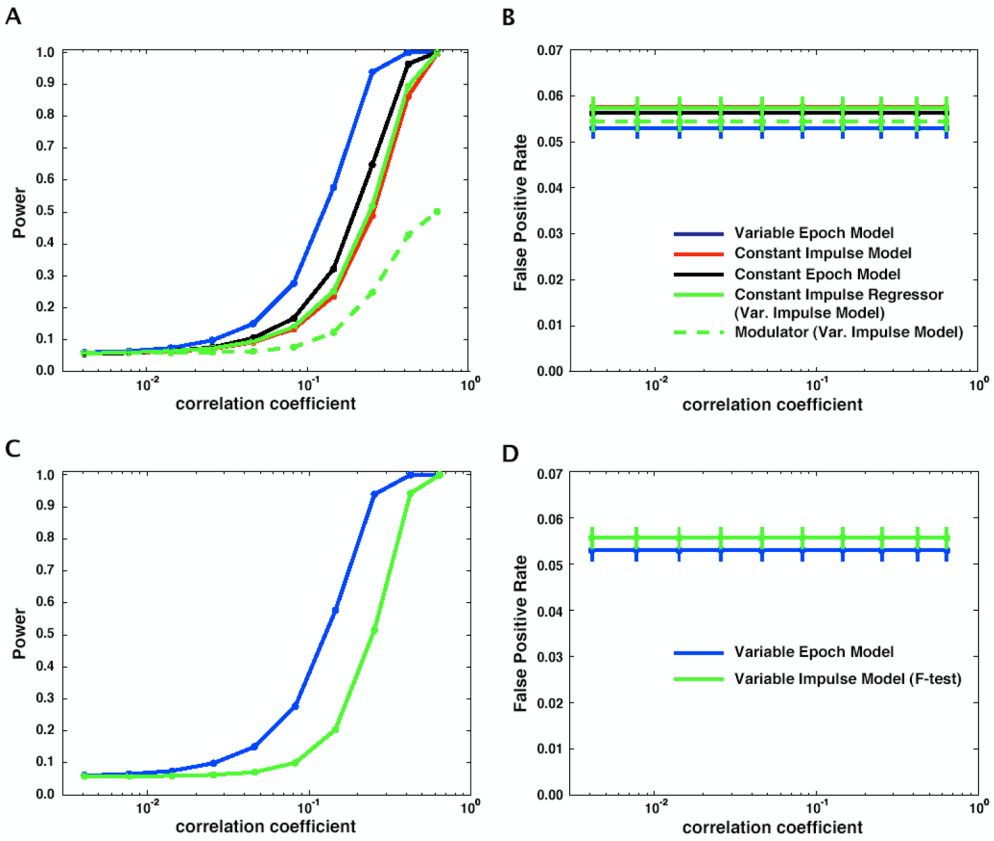

Figure 3. Detection power and false positive rate for each model

Each model was used to detect a simulated time-varying signal. Each data point consisted of 10,000 simulated runs. **(A)** The *variable epoch model* (blue) has significantly higher detection power as a function of effect size (Pearson's correlation coefficient) than all the other models. Although the improvement in power is a function of run length, the results provide an estimate of the relative cost of using the other models in detecting time-varying signals. **(B)** All the models have similar false positive rates. **(C)** The *variable impulse model* (two regressors; green) has higher power than the *variable epoch model* (single regressor; blue) for small effect sizes. However, this comes at a substantial cost due to an increase in false positive rate **(D)**. Error bars represent standard deviation (note: error bars are too small to be visible in A and C).

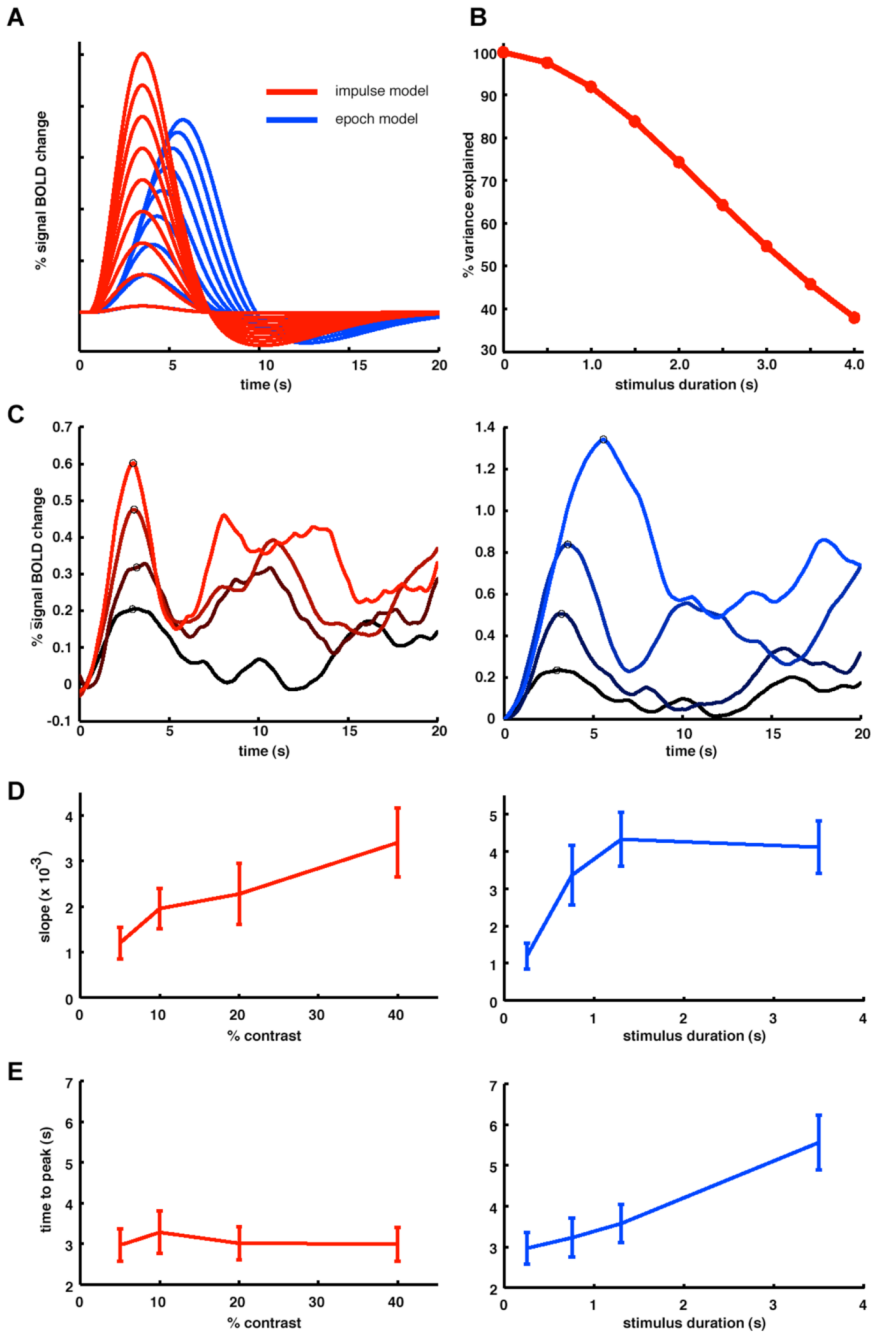

Figure 4. Dissociating changes in intensity from changes in duration

(A) The canonical HRF was convolved with either an impulse of variable height (red) or an epoch of variable duration (0–4000 ms in 500 ms steps; blue). When only intensity is modulated, the shape of the HDR is constant, varies only in height, and is identical to the theoretical HRF (red). However, when duration is the critical variable, both the shape and height of the response vary (blue). **(B)** We calculated the Pearson's correlation coefficient, R^2, between the impulse model (or HRF) and the variable duration HDRs. The percent of temporal variance explained by the impulse model decreases as a function of duration. When the duration of the neural process is 3 s, the impulse model can only explain half of the variance in the data. **(C)** Data from the visual cortex of a single subject viewing flashing checkerboards

of variable contrast (5%, 10%, 20%, 40%) with a constant duration (0.25 s, left panel) and variable duration (0.25 s, 0.75 s, 1.3 s, 3.5 s) but constant intensity (5%, right panel). The circles indicate the peak intensity for each trial type. **(D)** As predicted by the LTI model in (A), the slope of the HDRs in (C, red) increases linearly with stimulus intensity (red). However, for stimulus durations greater than ~1.3 s, the slope of the HDRs in (C, blue) remains constant (blue). Error bars represent standard error. **(E)** As stimulus intensity increases, the time at which the HDRs reach their peak intensity remains constant (red), as predicted by the LTI model in (A). In contrast, the time to peak is linearly related to stimulus duration (blue). Error bars represent standard error.

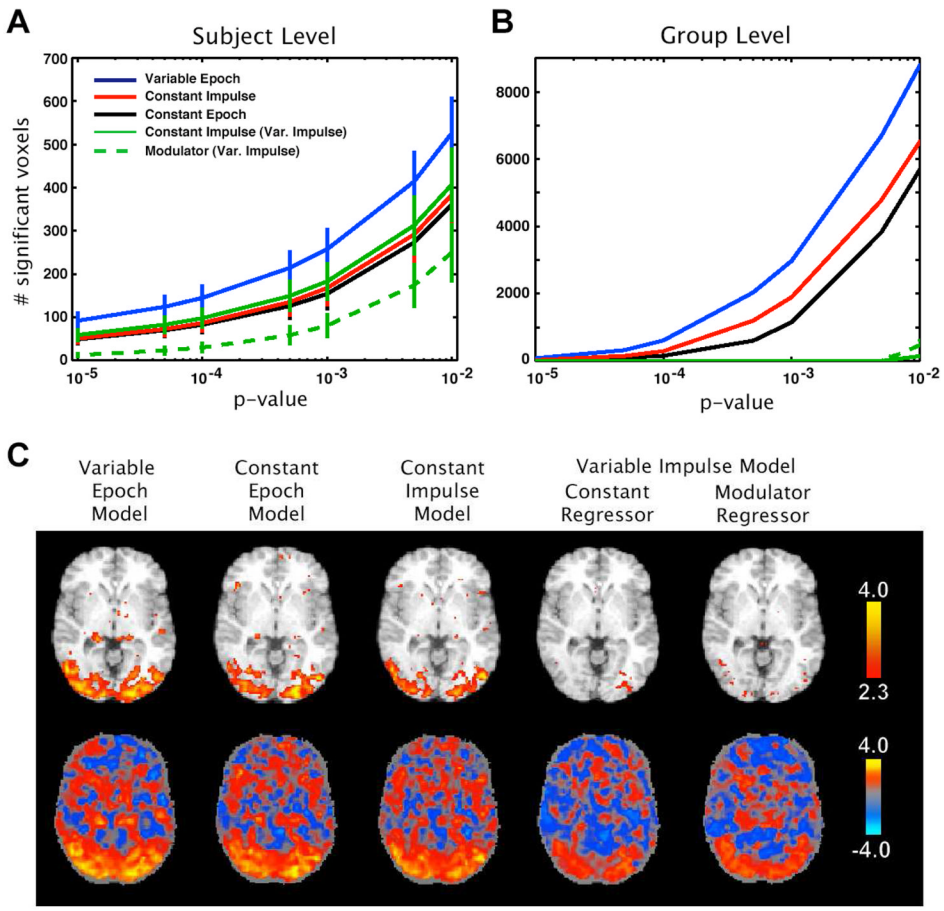

Figure 5. Detection power and consistency of detected response

(A) When a neural process (in this case a flashing checkerboard) has variable duration across trials, the variable epoch model detects a greater number of significant voxels on each 5.5 min run than the other models. **(B)** The same pattern is true at the group level (mixed-effects analysis; n = 8 subjects). The variable impulse model has a much less consistent response across runs and across subjects, resulting in a large drop in sensitivity at the group level. **(C)** An example of the mixed-effects group level activation maps demonstrates that the variable epoch model showed higher Z-statistics and detects many more significant voxels than the other models. The unthresholded activation maps show that the variable impulse model has a similar, but non-significant, spatial distribution of activity in the visual cortex.

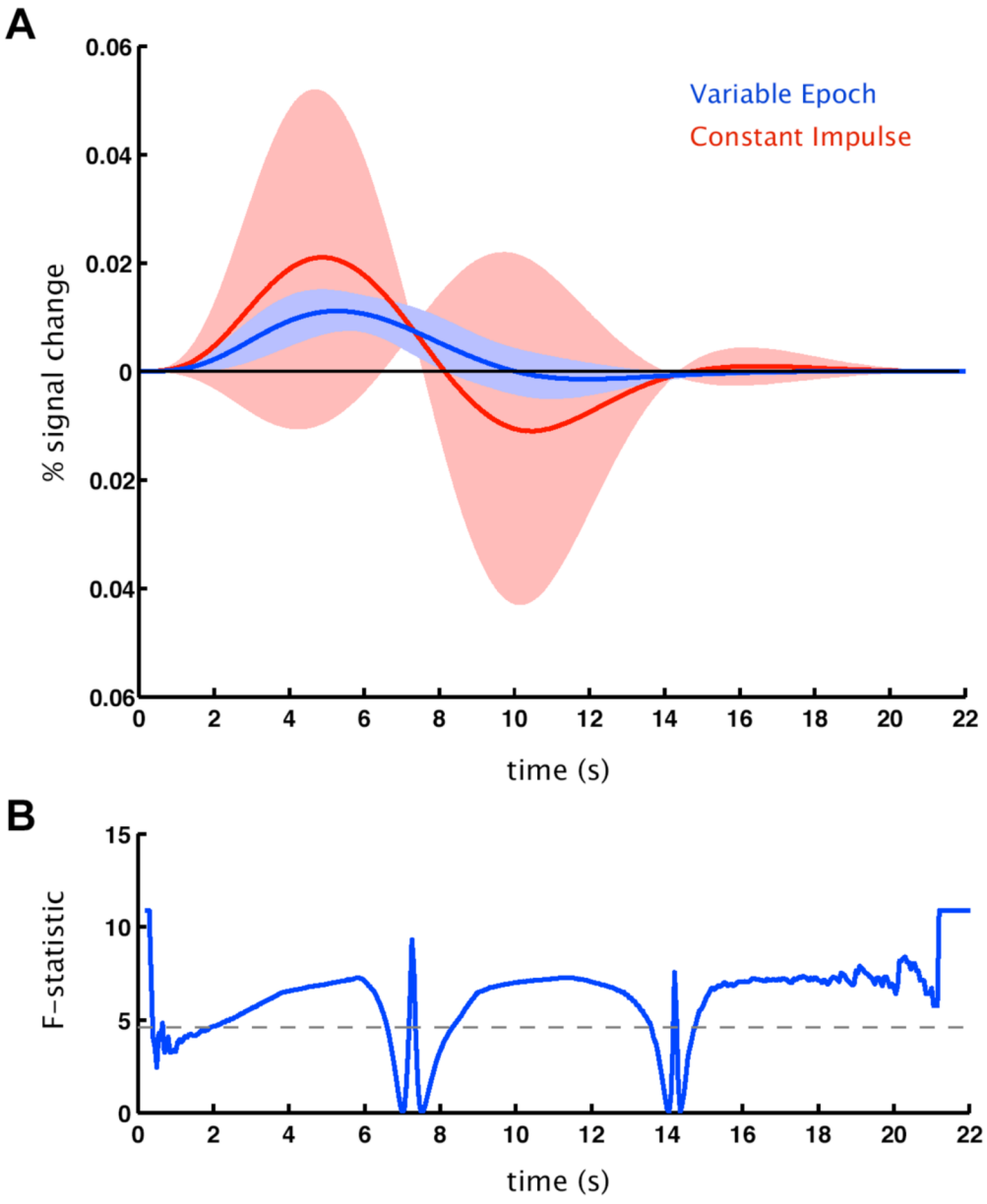

Figure 6. Quality of the HRF estimate

(A) The variable epoch model generates a mean HRF estimate across subjects with lower variance than the constant impulse model. Shaded regions represent one standard deviation. **(B)** Levene's Test for Equality of Variance was used to determine whether the variances were significantly different. The dotted line represents the significance threshold, set at $p < 0.05$, $F(1,14) = 4.6$. The majority of the time points have significantly higher variance for the impulse HRF estimate than for the variable epoch HRF estimate. Note that the crossover points at 7.5 s and 14.2 s are the only regions where the variance for the epoch HRF exceeds the variance for the impulse HRF.