

30 Experimental Design for Brain fMRI

G.K. AGUIRRE and M. D'ESPOSITO

CONTENTS

30.1	Experimental Design for Brain fMRI	369
30.2	BOLD fMRI System and Its Output	370
30.3	Categorical, Subtractive, Blocked Designs	371
30.4	Factorial, Conjunction, and Parametric Designs	374
30.5	Event-Related Designs	375
30.6	Issues in Intergroup Comparisons	377
30.7	Involvement and Implementation Hypotheses	378
	References	379

30.1 Experimental Design for Brain fMRI

This section concerns the design of fMRI experiments. If there is any value to our writing (and your reading) a discourse on the subject, then it must be the case that there are such things as "better" and "worse" experimental designs for fMRI. Normative judgments, in turn, require criteria on which they can be based, and here we consider two. An fMRI experiment should: (1) be inferentially capable of rejecting a hypothesis of interest, (2) maximize sensitivity for a predicted effect. This section will consider a variety of experimental designs for fMRI with regard to these standards.

An fMRI experiment should be capable of rejecting a hypothesis of interest. In one sense this is a trivial point - if the design cannot be used to reject a hypothesis, it is not by definition an experiment (PLATT 1964)! Beyond this glib statement, however, lies a fair bit of complexity. Different experimental designs for fMRI rely upon different sets of assumptions and are sensitive to different types of confounds. The possibility that assumptions are not met or confounds are present can render the interpretation of an experiment equivocal. While all of science

grapples with these challenges, certain fMRI designs are prone to rather plausible inferential failures. In this section, we will explore how some experimental designs reduce reliance upon untenable assumptions and discount the possibility of certain types of confounds.

Given inferential soundness, an experimental design should maximize sensitivity for the effect to be detected. Greater sensitivity allows one to infer with greater confidence that the absence of a significant statistical test is the result of the absence of an effect of a certain magnitude or larger. Because fMRI experiments that demonstrate localized responses to some task are often tacitly interpreted as showing the absence of a response at other locations, sensitivity is a particularly important issue to keep in mind when designing and interpreting neuroimaging studies. The search for optimally sensitive fMRI designs will require us to become acquainted with two properties of the BOLD fMRI system: the transfer function that maps neural activity onto BOLD hemodynamic signal and the temporal autocorrelation structure of the noise. With knowledge of these two characteristics of the system, we will be in a position to rate the relative power of different experimental designs. In considering these issues, reference will be made to the Fourier analysis of periodic time-series. While extensive knowledge of these concepts is not essential to understand the conclusions presented here, the explanations will require some familiarity.

We begin by considering salient aspects of BOLD fMRI data that will guide subsequent discussion. Next, we discuss several experimental designs with regard to their relative statistical power and inferential strength. Finally, we consider the limitations of neuroimaging experiments (including PET, fMRI, evoked-potentials, and electrophysiology) with regard to strong inference about the neural substrates of cognitive processes.

30.2 The BOLD fMRI System and Its Output

As has been well described in previous chapters in this volume, BOLD fMRI data are time-series measurements produced by a nearly linear system. We will unpack this jargon-laden sentence over the next few paragraphs and consider how these features of BOLD fMRI impact experimental design.

First, the absolute level of BOLD fMRI signal values are not readily interpretable. Unlike PET, for example, where the signal measured can be expressed as a physical quantity (e.g., ml blood/100 g of tissue per minute), the BOLD fMRI signal has no absolute interpretation. (Recent developments in perfusion imaging, however, offer the possibility of an fMRI signal that can be interpreted in concrete physical units.) This is because the particular signal value obtained is not *exactly* a measure of deoxyhemoglobin concentration, but is instead a measure that is *weighted* by this concentration (i.e., is T_2^* weighted) and is also influenced by a number of other factors that can vary from voxel to voxel, scan to scan, and subject to subject. As a result, experiments conducted with BOLD fMRI generally test for differences in the magnitude of the signal between different conditions within a scan. One could not, for example, directly contrast the *mean* level of the BOLD fMRI signal obtained within the temporal lobe of Alzheimer's disease patients with that from controls with much hope of obtaining a reliable or unbiased statistical test regarding functional activity. Because the mechanism that mediates a *change* in signal level may be constant across subjects and groups, task x group interactions might be tested (but see Sect. 30.5 below).

Second, BOLD fMRI data can be treated as the output of a system that transforms neural activity

and is time-invariant and approximately linear (BOYNTON et al. 1996). A linear system can be completely characterized by its impulse response function: the output of the system to an infinitely brief and powerful input. In the context of BOLD fMRI, the change in hemodynamic fMRI signal that results from a brief (approximately <1 s) period of neural activity is an adequate stand-in for the impulse response function. The hemodynamic response function can also be represented as a frequency response (like those that characterize a resistance-capacitance circuit). The frequency representation of the hemodynamic response function can be termed the transfer function of the system. The transfer function shows how information at frequencies in the input (in this case neural activity) are scaled by the linear system. The (square of) an estimate of the BOLD transfer function (provided in Fig. 30.1B) shows that low frequencies are preferentially allowed to "pass" through the system. Hence, we call the BOLD system a "low-pass" system.

Low-pass filtering reduces temporal resolution. For instance, changes in fMRI signal corresponding to changes in neural activity every 2 s will be very difficult to detect. Indeed, the higher the paradigm frequency (and presumably the frequency of neural activity changes), the less efficiently the variance of the task will be passed into the fMRI signal. It should be noted that this limitation cannot be readily overcome by more rapid imaging. Even if BOLD fMRI data are collected every 100 ms, the filtering properties of the hemodynamic response (which are dictated by physiology and are not influenced by the scan-to-scan repetition time) will still make rapid alternations in neural activity nearly undetectable. As will be developed below, these limits of temporal resolution do not prevent fMRI experiments from detecting: (1) brief changes in neural activity, (2) dif-

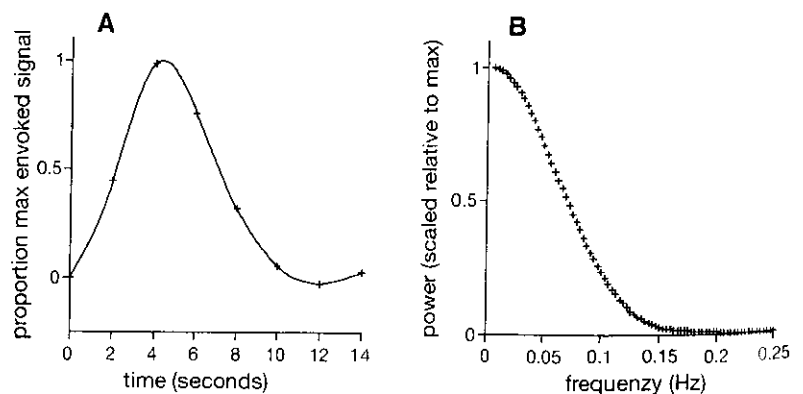


Fig. 30.1A,B. The hemodynamic response. A The change in fMRI signal that results from a brief period of neural activity (AGUIRRE et al. 1998b). B The power spectrum of the response shown in A

ferences in evoked neural activity of randomly ordered, closely spaced events, or (3) neural onset asynchronies on the order of hundreds of milliseconds between different trial types.

Finally, BOLD fMRI data are temporally autocorrelated, or "colored," under the null-hypothesis. That is to say, in fMRI data that are collected from human subjects in the absence of any experimental task or time varying stimuli, greater power is seen at some frequencies than at others. The shape of this distribution of power is well characterized by a 1/frequency (1/f) function (ZARAHN et al. 1997a), and is shown in Fig. 30.2. As can be seen, there is increasing power at ever lower frequencies. In addition to rendering traditional parametric (AGUIRRE et al. 1997; ZARAHN et al. 1997b) and nonparametric (AGUIRRE et al. 1998a) statistical tests invalid, this temporal autocorrelation causes relative reductions in sensitivity for some experimental designs. Specifically, experiments with fundamental frequencies in the lower range (e.g., a boxcar design with 60 s epochs) will have reduced sensitivity, due to the presence of greater noise at these lower frequencies. We noted just above that, because of the low-pass filtering properties of the hemodynamic transfer function, paradigms in which the variance is present at low frequencies will tend to have greater statistical sensitivity. Now we observe that, because of the presence of ever greater noise at lower frequencies, higher frequency paradigms will tend to have greater statistical sensitivity. This suggests a trade-off and the existence of an optimum.

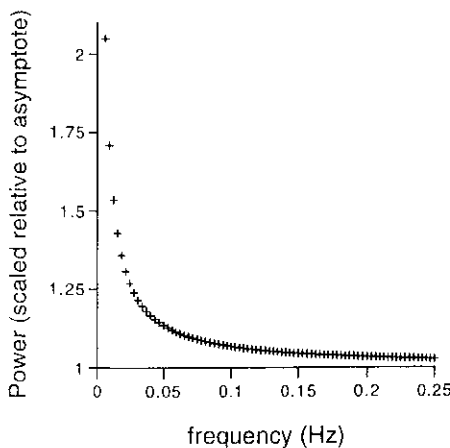


Fig. 30.2. Temporal autocorrelation under the null-hypothesis. Average power spectrum across subjects of fMRI data collected while the subjects rested quietly

The hemodynamic response function and the 1/f power structure of the noise can be considered the Scylla and Charybdis of fMRI experimental design: experimental variance must be present at sufficiently low frequencies to pass through the hemodynamic transfer function but at sufficiently high frequencies to avoid the elevated noise range. As a result, the sensitivity of fMRI experiments will be optimized when the design concentrates its power at frequencies in the mid-range. Throughout this chapter we will comment upon the relative statistical power of different designs by considering the frequency structure of the experimental paradigm in relation to these factors.

30.3 Categorical, Subtractive, Blocked Designs

The prototypical fMRI experimental design is shown in Fig. 30.3. This, the original boxcar approach in which two conditions alternate over the course of a scan, is a categorical, subtractive, blocked design. Categorical because the experiment examines two levels of a category, and blocked because, for most hypotheses of interest, these periods of time will not be utterly homogeneous but will consist of a block of several "trials" of some kind presented together. For example, a given block might present a series of faces to be passively perceived, or a sequence of words to be remembered, or a series of pictures to which the subject must make a living/non-living judgment and press a button to indicate his response. The block of trials is designed to evoke a particular cognitive process, such as face percep-

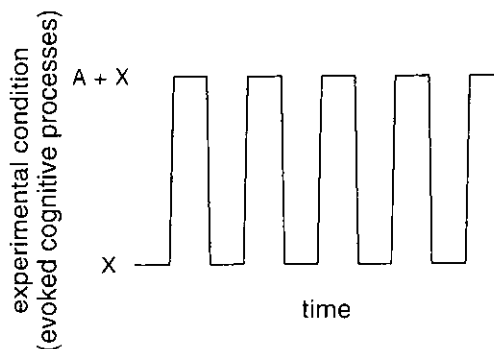


Fig. 30.3. Categorical, blocked, subtractive experimental design

tion, episodic encoding, or semantic recall. These "experimental" blocks alternate with "control" blocks that are designed to evoke all of the cognitive processes present in the experimental block except for the cognitive process of interest. Under the assumptions of "cognitive subtraction" (POSNER et al. 1988), differences in neural activity between the two conditions can be attributed to the cognitive process of interest.

One issue of some importance in the design of blocked experiments concerns the pacing of presentation of experimental trials. In many cases, the particular hypothesis to be tested will dictate the duration of presentation of stimuli and the duration of the inter-trial-interval. For example, the evocation of certain cognitive processes (e.g., semantic priming) requires the presentation of stimuli with a certain duration and minimal spacing. It should be noted, however, that sensitivity will be maximized by making each blocked period evoke the cognitive process of interest in as homogeneous a manner as possible. This may be facilitated by maximizing the period of time in which stimuli are presented and minimizing any inter-trial-interval. For example, a putative face-responsive region would generate a maximal change in fMRI signal given a series of face stimuli with no blank inter-face-interval. In those experiments in which the subject makes a response to each behavioral trial, "self-pacing" offers the ability to minimize the spacing between each trial. In a self-paced design, the rate of presentation of stimuli is dependent upon the reaction time of the subject. Perhaps surprisingly, the selection of self-paced or "fixed-paced" ordering of trials can have a nontrivial impact upon the inferences that may be drawn from a blocked imaging experiment. For details, the reader is referred to (D'ESPOSITO et al. 1997).

What types of assumptions are being made, and what sorts of confounds are present, in categorical, blocked fMRI designs? The discussion of these issues offered here is drawn from (ZARAHN et al. 1997b). Cognitive subtraction (in neuroimaging) generally relies upon two assumptions: "pure insertion" and linearity. Pure insertion is the idea that a cognitive process can be *added* to a pre-existing set of cognitive processes without affecting them. This assumption is difficult to prove because one would need an independent measure of the preexisting processes in the absence and presence of the new process. This problem exists in both chronometric psychological studies (STERNBERG 1969) and neuroimaging studies (FRISTON et al. 1996). If pure insertion fails as an assumption, a difference in neuroimaging signal be-

tween the two conditions might be observed not because of the evocation of the cognitive process of interest, but because of an interaction between the added component and preexisting components. For example, the act of pressing a button to signal a semantic judgment may be different from pressing a button in response to a visual cue. Effects upon the imaging signal that result from this difference would be erroneously attributed to semantic judgment *per se*.

A second assumption of cognitive subtraction is that the transformation of neural activity into fMRI signal is linear. While the BOLD fMRI system has been shown to exhibit behavior close to that of a linear system, there is some evidence for systematic departures (BOYNTON et al. 1996; DALE and BUCKNER 1997; VASQUEZ and NOLL 1996). What effects might the failure of linearity have? Consider a paradigmatic working memory experiment that presents the subject with a stimulus to be remembered, a brief delay, and then a choice stimulus. Multiple trials of this type are contrasted with a control condition in which the delay follows the choice condition, allowing the subject to make responses without relying upon working memory (JONIDES et al. 1993). Cognitive subtraction requires that the total fMRI signal evoked by neural events associated with the temporally separated presentation and choice stimuli and motor response of working memory trials be equivalent to the fMRI signal evoked by the neural event associated with the contemporaneous presentation and choice stimuli and motor response of the control condition. If there is a failure of linearity such that the response to the temporally adjacent stimuli is less than the total response to the separated stimuli, then the experimental design might lead to the erroneous inference that a region displayed delay-correlated increases in neural activity when it actually did not. In fact, failures of cognitive subtraction in these kinds of working memory studies have been empirically demonstrated (ZARAHN et al. 1997b).

Blocked designs also, by their very nature, do not allow the randomization of the order of stimuli and are constrained to grouping trials of the same type with each other in time. The consequent predictability of trial type may act as a confound in a blocked experiment. For example, many imaging studies of the neural substrates of recognition/novelty processing have involved presenting subjects with blocks of either all old (i.e., previously presented during an encoding condition) or all new (i.e., not presented during an encoding condition) stimuli together for judgments of recognition. Such a situation

highlights the a priori undesirability of being constrained to blocked trial structures. The influence of trial order (i.e., blocked or random) on functional neuroimaging data can occur on at least two levels. First, the order of trial presentation may have an effect upon the cognitive processes engaged within the trials themselves (JOHNSON et al. 1997). Second, blocked or random presentation may affect the cognitive processes during the inter-trial interval. For example, changes in imaging signal may be the result of anticipatory behaviors in which the subject engages before the presentation of each stimulus. Another way of stating these observations is that every blocked experiment is confounded by behaviors that may be the result of groups of similar trials being presented together, as opposed to the effect of the individual trials themselves.

Because of these relatively untenable assumptions and plausible confounds, categorical, blocked fMRI experiments are not capable of strong inference. Although a particular instantiation of this design might have better or worse claims to satisfying the assumptions noted above, these failures exist as logical possibilities regardless of their credibility. In some cases, however, a categorical, blocked experimental design is acceptable. When an experiment is to address a purported cognitive process that: (1) is an all-or-none phenomena (i.e., cannot be subjected to parametric manipulation), (2) is homogeneous in its evocation (e.g., there are not correct or incorrect trials) and (3) cannot be separated by several seconds in time from other cognitive processes (i.e., is unlike the delay period of working memory trials) then many of the alternative designs described below offer no advantage.

One reason to actually favor blocked designs is their superior statistical power. The fundamental frequency of the boxcar can be positioned so that variance is maximally passed by the hemodynamic response function but avoids the elevated noise range at low frequencies. Simulations that have used the hemodynamic response function shown in Fig.

30.1A and the $1/f$ noise structure shown in Fig. 30.2 indicate that the optimal block length for a boxcar design is around 14–20 s (0.036–0.025 Hz) (ZARAHN et al. 1997b; Aguirre, unpublished observation). This value assumes a good estimate of the hemodynamic response function (perhaps obtained empirically from the subject; see AGUIRRE et al. 1998b). If this is not available, slightly longer block lengths (e.g., 30 s) become preferable.

When more than two conditions are included in a blocked experiment, the experimenter has several options for how those blocks are to be ordered during the scan (e.g., fixed: A-B-C-A-B-C-A-B-C; or random: A-C-C-B-A-B-C-B-A; see Fig. 30.4A,B). While randomizing the order of the blocks may have appeal from a psychological design standpoint, this randomization decreases sensitivity for differences between the conditions by distributing the variance of the task paradigm over multiple frequencies, including low frequencies that are in the high noise range. If the experiment is to be conducted in several subjects, then a desirable compromise is to fix the order of the blocks within a subject, but to vary (perhaps counterbalance) this order across subjects (see Sect. 30.5 for a discussion of the appropriate statistical model to use in group analyses).

A final application of blocked designs is as an initial experimental probe. Below we discuss several alternative experimental designs that, while providing for stronger inference, have reduced statistical power compared to blocked designs. A possible approach is to combine both the blocked design just described and other designs within a series of experiments. For example, one might use a blocked experiment initially to define regions of interest and then interrogate those regions in subsequent experiments with the designs described below (e.g., AGUIRRE et al. 1998c). The primary drawback of a combined design is that there may exist regions that, because of their response properties, are only detectable using one of the more sophisticated designs. In this case, the combined approach would not be able to detect these regions.

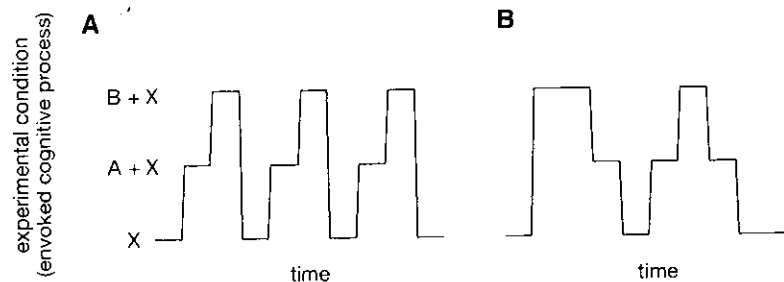


Fig. 30.4A,B. Three condition categorical, blocked experiment, showing A fixed and B randomized block orders

30.4 Factorial, Conjunction, and Parametric Designs

Several experimental designs have as their goal a reduction in the reliance upon the assumption of pure insertion. We will describe these approaches as they apply to blocked experiments, but it should be appreciated that these concepts can also be applied to the event-related designs described below.

Factorial experiments (FRISTON et al. 1996) are explicitly designed to examine the interactions of two different, candidate cognitive processes. The scheme of the design, illustrated in Fig. 30.5, involves (in the simplest case) four conditions, during which two different processes are evoked individually and then jointly. In essence, this amounts to a categorical, blocked design with four conditions. The proposed advantage of the design is that interactions between the two cognitive processes ("A" and "B" in this example) can be examined. The presence of an interaction is indicated if the difference in imaging signal between the presence and absence of cognitive process "A" is itself different when cognitive process "B" is present or absent (i.e., if $[A+B+X]-[B+X]$ is different from $[A+X]-[X]$). While factorial designs do provide a compelling method for gaining greater insight into the neural implementation of cognitive processes, it is a mistake to claim that such designs obviate the need for the pure insertion assumption. Interpretation of the design requires the assumption that the two cognitive processes have, indeed, been isolated. The logic by which this isolation is to occur is the same as that outlined for cognitive subtraction above. That is, process "A" and process "B" must be purely inserted into the other cognitive components ("X") that allow the experiment to evoke these processes.

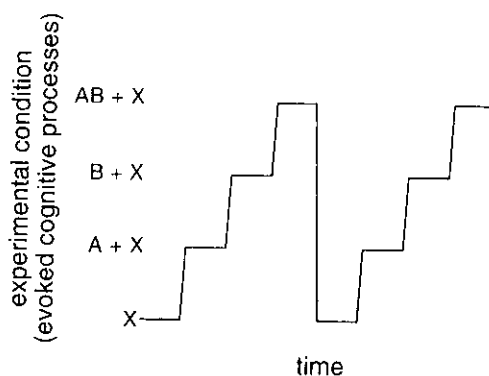


Fig. 30.5. Blocked factorial design

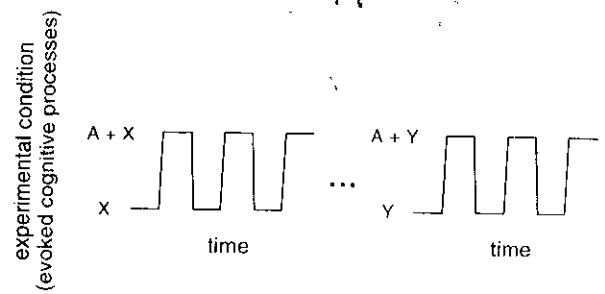


Fig. 30.6. Cognitive conjunction design

The "cognitive conjunction" design (PRICE and FRISTON 1997) has also been proposed to reduce reliance upon the assumption of pure insertion. The logic of the approach is that, if one wishes to discount the possibility of an interaction (i.e., a failure of pure insertion) between the cognitive component to be added and the set of preexisting processes, one should repeat the experiment with a different set of preexisting processes and replicating the result (Fig. 30.6). A rigorous implementation of this notion (PRICE and FRISTON 1997) involves conducting a series of categorical subtraction experiments that all aim to isolate the same cognitive process. The novel twist is that the subtractions need not be complete; that is, the experimental and control conditions can differ in several cognitive processes in addition to the one of interest. The imaging data are then analyzed to identify areas that have a significant, consistent response to the putatively isolated process (i.e., a significant main effect across subtractions in the absence of any significant interactions). Again, while this design reduces the plausibility of some failures of cognitive subtraction, it does not eliminate the possibility. In particular, some cognitive processes, by their very nature, require the evocation of an antecedent process. For example, can working memory be meaningfully present if not preceded by the presentation of a stimulus to be remembered? If not, then any cognitive conjunction design that attempts to demonstrate the presence of neural activity during a delay period will be susceptible to erroneous results due to interactions between the task manipulation and preexisting task components.

Finally, parametric designs, when applicable, offer the opportunity to truly obviate the assumption of pure insertion. In a parametric design, the experimenter presents a range of different levels of some parameter, and seeks to identify relationships (most simply, a linear one) between imaging signal and the values that the parameter assumes (Fig. 30.7A). This can be done to identify the neural correlates of

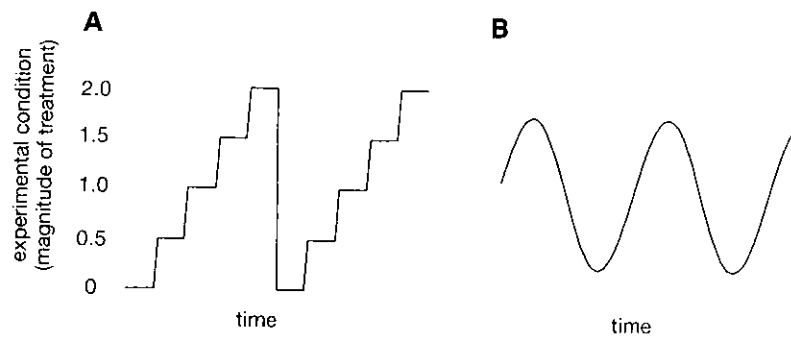


Fig. 30.7A,B. Parametric designs that are A blocked and B continuously varying

straightforward changes in stimulus properties (e.g., the chrominance of a stimulus, ENGEL et al. 1997) or manipulations of a cognitive process (e.g., working memory, BRAVER et al. 1997). The reason that a parametric design may eschew the pure insertion assumption is that only the magnitude of the process of interest is altered. While it is still possible to conceive of inferential failures using this design, this approach is able to avoid several of the questionable experimental assumptions considered above. It should be noted, however, that a parametric design has reduced sensitivity compared to a two-condition design of the same length (i.e., with the same number of observations) that only uses the extreme levels of the parameter. This is because the two-condition design maximizes the ratio of task-induced variance to noise.

The parametric design illustrated in Fig. 30.7A is blocked, but it should be noted that continuously varying parametric designs can also be conducted (Fig. 30.7B). In this case, the parameter of interest is continuously varied over the course of the scan, and correlations between the imaging signal and this independent variable are sought (after accounting for the effects of the hemodynamic transfer function). A twist on the continuously varying parametric design is the so-called traveling wave stimulus, in which the subject is exposed to an expanding annulus of flickering light, or a continuously rotating wedge of flickering light. Interestingly, these designs, used to map early visual areas (ENGEL et al. 1994; SERENO et al. 1995) are parametric in *space*, but blocked in *time*.

30.5 Event-Related Designs

Continuously varying parametric designs serve as an appropriate segue to the discussion of event-related experimental design. An appealing aspect of

the continuously varying parametric design is that it avoids the possibility of behavioral confounds that are the result of blocking of stimuli together. An event-related design enjoys this advantage and others as well, including the ability to: (1) randomize trial presentations, and (2) test for functional changes between different measurable aspects of behavior (e.g., accuracy) or different characteristics of a trial (e.g., stimulus type). These advantages apply to all event-related designs and the reader is advised to consult the chapter by BUCKNER (this volume) and ZARAHN et al. (1997b) for a complete description. As we will discuss, a more refined advantage of a particular event-related fMRI design (ZARAHN et al. 1997b) is the ability to examine separately the neural substrates of components of behavior temporally dissociable on the order of a few seconds *within* a trial.

Event-related fMRI designs attempt to model signal changes associated with individual trials as opposed to a larger unit of time comprised of a block of trials (Fig. 30.8). Each individual trial may be composed of one behavioral event (such as the presentation of a single word) or several behavioral events (such as the presentation of a cue, a delay period, and a motor response in a delayed response task). In the simplest type of event-related experiment, the behavioral trials are distant enough in time from one another to allow the hemodynamic response that results from the hypothesized brief period of evoked neural activity to fully run its course (e.g., 16 s). A variety of analysis approaches are available that allow the statistical evaluation of these responses both with respect to the inter-trial-interval and with respect to one another (DALE and BUCKNER 1997; JOSEPHS et al. 1997; ZARAHN et al. 1997b). While this is properly the topic for a different review, we note here that when trials are spaced far enough apart in time to avoid any overlap in the hemodynamic response of one trial to the next, neither analysis nor inference (in one sense) requires the assumption of

linearity. Additionally, analysis methods exist that require no a priori assumptions regarding the specific shape of the evoked response (JOSEPHS et al. 1997).

Because the analysis focuses upon individual trials, it is possible to ascribe changes in the neuroimaging signal to the effect of one particular trial type, regardless of when it is presented within the experiment. This feature of event-related designs allows for the randomization of stimuli, thus avoiding behavioral confounds that are the result of blocking trials, and allows the separate analysis of functional responses that are only identified in retrospect (e.g., trials on which the subject made a correct or incorrect response).

In the simple case in which an event-related design has only a single trial-type, presented at a regular interval, it is possible to determine what particular trial spacing will result in optimal sensitivity for evoked responses. Assuming the hemodynamic response and noise structure shown in Figs. 30.1 and 30.2, a trial spacing of about 16 s is optimal (Aguirre, unpublished observation). (Optimal meaning: "maximizes sensitivity for an effect over a fixed duration of scanning." If the duration of the scan is free to vary and the number of trials is held constant, ever greater separation between the trials leads to ever greater theoretical sensitivity in all cases, including those considered below.) If the design calls for different trial types to be intermixed, optimal spacing becomes a more complicated issue. If the trials are spaced more closely together, the hemodynamic response from one trial overlaps with the response of the adjacent trial. This closer spacing of trials decreases the sensitivity of the design for evoked changes relative to the inter-trial-interval, but increases sensitivity for differences between the trial types, *as long as they are in a random order*. Additionally, closer spacing of the trials increases the

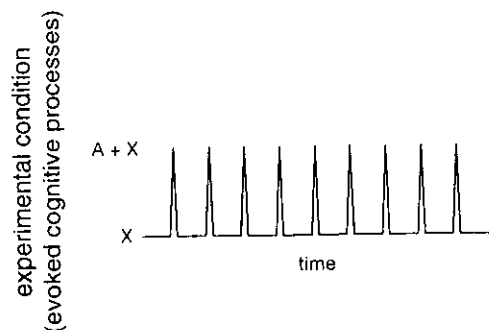


Fig. 30.8. Event-related design

stringency of the assumptions that must be satisfied for a successful analysis. Specifically, the analysis must now assume linearity for the system. Notably, failures of these assumptions can be expected to lead to *false negative* results (due to reductions in sensitivity), as opposed to the false positives that might result from the failures of assumptions in other experimental designs. In the limit, with the assumptions of linearity perfectly satisfied, the optimal spacing to detect a difference between two randomly ordered trial types approaches zero. Again, as this spacing decreases, the ability to detect the evoked response of either trial type relative to the inter-trial interval also approaches zero. Empirical tests conducted within the primary visual cortex are in broad agreement with these assertions (DALE and BUCKNER 1997).

Because the frequency structure of a train of impulses (the neural activity proposed to result from an event-related design) is itself a series of impulses, much of the variance in an event-related design is present at higher frequencies and is therefore lost after passing through the hemodynamic response function. As a result, event-related designs as a class are reduced in sensitivity relative to blocked designs, regardless of the spacing of trials and manipulations of the inter-trial-interval. However, one way to improve sensitivity in an event-related design is to introduce "jitter" into the inter-trial-interval (Fig. 30.9). Variability in the inter-trial-interval acts to distribute some of the variance of the design from higher frequencies to lower ones. Such jitter designs also have the advantage of reducing the ability of the subject to engage in anticipatory behaviors prior to the onset of each trial.

The discussion thus far regarding event-related designs has assumed an ability to randomize perfectly the order of presentation of different event types. There are certain types of behavioral paradigms, however, that do not permit a random ordering of the trial events. For example, the delay period of a working memory experiment always follows the presentation of a stimulus to be remembered. In this case, the different events of the trial cannot be placed arbitrarily close together without risking the possibility of false *positive* results that accrue from the hemodynamic response to one trial event (e.g., the stimulus presentation) being interpreted as resulting from neural activity in response to another event (e.g., the delay period). It turns out that, given the shape typically observed for hemodynamic responses, events within a trial as close together as 4 s can be reliably discriminated within the context of a

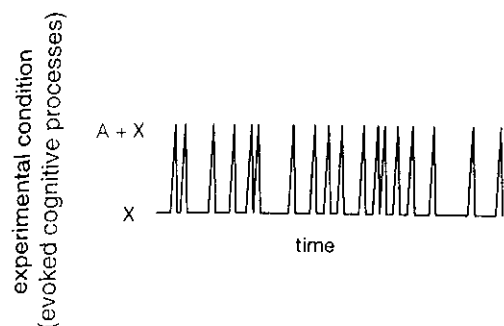


Fig. 30.9. Event-related design with variable inter-trial-intervals

least-squares analysis (ZARAHN et al. 1997b). Thus, event-related designs can be used to examine directly, for example, the hypothesis that certain cortical areas increase their activity during the delay period of a working memory paradigm without requiring the problematic assumptions traditionally employed in blocked, subtractive designs (ZARAHN et al. 1997b; see also the chapter by COHEN, this volume).

The transition to event-related design offers the opportunity to test many hypotheses in a principled, rigorous fashion that was simply not previously possible. As a final example of event-related design, consider an experiment that aims to identify a neural onset asynchrony. We have recently demonstrated that the hemodynamic response observed for a subject during a scanning session is highly reliable in its shape (AGUIRRE et al. 1998b). As a result, it might be possible to use fMRI to detect small neural onset asynchronies. Such a design might present one of two different behavioral trials (in a random order) every 16 s. Because of the reliability of the hemodynamic transformation, differences in the mean time-to-peak of the responses to the two different types of stimuli could be identified and ascribed to an asynchrony in onset of neural activity. Such a design might allow one, for example, to test the hypothesis that a cortical area that responds to pictures of faces responds with a slightly longer latency (on the order of 100 ms) to pictures of inverted faces.

30.6

Issues in Intergroup Comparisons

The discussion so far has focused upon experimental designs that attempt to address hypotheses regarding neural activity in individual subjects. Often,

it is of interest to test the hypothesis that either: (1) the *population* from which the subjects are drawn possesses the hypothesized effect, or (2) that two different populations differ in the evocation of some effect. The movement from individual subject tests to studies of groups of subjects is accompanied by several complications.

The most immediate of these is that different subjects have brains that are shaped differently. If one wishes to test a hypothesis regarding a certain area of the brain, then it is first necessary to identify that same area of the brain across subjects. While there are a variety of sophisticated methods available for registering and aligning the brains of different subjects into a standard space, there are theoretical limits to what such an alignment can achieve. First, there may be inter-subject variability in anatomy that cannot be overcome by warping brains to a standard space. For example, the arrangement of the sulci in ventral occipitotemporal cortex is known to vary between subjects (ONO et al. 1990). Thus, while two subjects may have neural responses at the same "true" cyto-architectonic location, the position of this site with respect to other landmarks in the brain may differ between subjects, leading to spread of these locations when data are converted to a standard space (WOODS 1996). Second, even given rigid alignment of anatomy across subjects, there may be variability in the structure: function relationships between subjects. For example, two subjects may truly have distinct face selective neural regions, but these may be located in different sections of a cortical area as a consequence of differences in nature or nurture. Again, when normalized to a standard space, this variability in location will obscure functional dissociations.

An alternative to anatomical registration is functional identification. The approach here is to first identify a region across subjects by its functional responses. For example, one might identify a region that responds more to pictures of faces than to general objects. Then, hypotheses regarding the response of this functionally defined region to other types of stimuli (for example, faces compared to hands) can be independently tested across subjects within this area (KANWISHER et al. 1996). This powerful approach allows one to make inferences across subjects regarding the responses of some functional area (e.g., the fusiform face area) at the expense of making statements regarding some particular position in a standardized space.

The next issue is a statistical one. In virtually all group PET and fMRI studies reported to date in

which repeated observations have been obtained from each subject, a fixed-effect statistical model has been employed. This type of model makes the assumption that observations from different subjects are the same as observations from within a single subject, i.e., that there are no subject \times task interactions (HOLMES et al, 1998). If one wishes to make inferences regarding an effect that extends to the population from which the subjects were drawn (i.e., the population of neurologically intact young college and medical students for most studies!) then it is necessary to employ a statistical model that explicitly accounts for subject \times task interactions in the data. Such a random effects model (as subject \times task interactions imply that the effect of task in each subject is a random effect) is actually very simple to implement for fMRI. All that is needed is a single value from each subject (from each anatomical location that is to be compared) that reflects the effect of the task for that subject. Within the context of the general linear model, this value should be the parameter estimate (or its corresponding t value) obtained for the predictor that models the hypothesized effect. These values across subjects are then tested using a simple t test against the null-hypothesis that their mean is zero. There is no need to correct for any kind of autocorrelation in the data, as these values are independent. The alert reader may have noted one downside of this statistical approach: the degrees of freedom of the test are only equal to the number of subjects involved in the study (minus 1)! As a result, hypotheses that attempt to demonstrate consistent effects across subjects using fMRI are far less powerful (in the statistical sense) than identical hypotheses that seek this effect only in individual subjects. While this is unfortunate, it is also unavoidable if appropriate inference is to be drawn. Furthermore, if population inference is desired, experimental resources should be devoted towards gathering data from more subjects as opposed to more data per subject.

This type of design can be extended to test hypotheses regarding differential responses of different populations to a task. For example, one might propose that within an anatomically specified area of the fusiform gyrus, copy editors have a greater fMRI signal response to the presentation of letters of the alphabet as compared to normal controls. Again, a mixed-effect model that compared the effect sizes between the two populations would appropriately test this idea. These tests seem straightforward when both populations are healthy, neurologically intact subjects who differ only in their job experience. They

become inferentially more challenging, however, when one population is composed of patients with neurological or psychiatric impairments, the elderly, or the recipients of a pharmacological agent. In these cases, the experimenter must be concerned about a confounding change in physiology or metabolism that might create artifactual results. For example, if elderly subjects have altered *vascular* responses as compared to young controls, then the observation of a decreased fMRI response to certain cognitive tasks cannot necessarily be interpreted as a difference in *neural* response. Ideally, one would also perform a control task that is not hypothesized to have any task \times group interactions, thus disputing the possibility of this confound.

30.7 Involvement and Implementation Hypotheses

So far, our discussion of experimental design has not considered the specific question regarding the relationship between the brain and cognition that is under study. Regardless of the particular type of fMRI design selected, there are two broad categories of question that might be addressed using neuroimaging techniques. We would like to suggest here that neuroimaging can provide for stronger inference regarding one compared to the other. The logical framework from which this discussion is drawn can be found in (ZARAHN 1998).

One class of hypotheses that are frequently tested using neuroimaging methods might be termed "involvement" hypotheses. These are proposals of the kind that a region is causally involved in the production of a particular cognitive process. It turns out that neuroimaging experiments do not allow for strong inference regarding these types of hypotheses, such as interpreting a positive result as a demonstration that the region is involved in the putatively isolated cognitive process. The primary cause of this state of affairs is the observational, correlative nature of neuroimaging (SARTER et al. 1996). Although we make inferences regarding cognitive processes, these processes are not themselves directly subject to experimental manipulations. Instead, the investigator controls the presentation of stimuli and instructions, with the hope that these circumstances will provoke the subject to enter a certain cognitive state and *no other*. Careful consideration reveals how this assumption might fail. For instance, although

cooperative, the subject may unwittingly engage in confounding cognitive processes in addition to that intended by the experimenter, or alternatively, may fail to differentially engage the process (i.e., it may already be on). It is therefore not possible to know if observed changes in neural activity in a brain region are the result of the evocation of the cognitive process of interest or an unintended, confounding process. Negative results (even in the face of arbitrarily high statistical power) are also not conclusive, both because of the failure of cognitive control and because of the possibility that the neuroimaging method employed might not be sensitive to the critical change in metabolic activity (e.g., pattern of neuronal firing as opposed to bulk, integrated synaptic activity) (PUCE et al. 1997).

Alternatively, neuroimaging techniques can provide inferentially strong tests regarding *implementation* hypotheses. This improvement in inference derives from a shift in the structure of the hypothesis. An implementation hypothesis *begins* with the assumption that a cortical region is involved in a particular cognitive process. The experiment then examines the activity in that area during different experimental conditions to: (1) characterize the nature of the computation that underlies the involvement of the region, or (2) learn something about a behavioral state. For example, the first type of implementation study might ask if the bulk rate of neuronal activity in area MT is monotonically related to a behavioral measure of motion perception. In the second type of implementation experiment, one might use activity in area MT as an index of motion perception, and then test hypotheses regarding how unrelated distractor tasks affect motion perception (REES et al. 1997). Neuroimaging can provide for stronger inference for these types of hypotheses because there is an isomorphic relationship between the independent variables that the experimenter manipulates and the subject of inference. Of course, the value of these types of experiments is itself dependent upon the veracity of the assumptions employed.

References

- Aguirre GK, Zarahn E, D'Esposito M (1997) Empirical analyses of BOLD fMRI statistics. II. Spatially smoothed data collected under null-hypothesis and experimental conditions. *NeuroImage* 2:199-212
- Aguirre GK, Zarahn E, D'Esposito M (1998a) A critique of the use of the Kolmogorov-Smirnov statistic for the analysis of BOLD fMRI. *Magn Reson Med* 39:500-505
- Aguirre GK, Zarahn E, D'Esposito M (1998b) The variability of human BOLD hemodynamic responses. *NeuroImage* 8:360-369
- Aguirre GK, Zarahn E, D'Esposito M (1998c) An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron* 21:373-383
- Boynton G, Engel S, Glover G, Heeger D (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* 16:4207-4221
- Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC (1997) A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage* 5:49-62
- Dale AM, Buckner RL (1997) Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapping* 5:329-340
- D'Esposito M, Zarahn E, Aguirre GK, Shin RK, Auerbach P, Alsop DC, Detre JA (1997) The effect of pacing of experimental stimuli on observed functional MRI activity. *NeuroImage* 6:113-121
- Engel S, Zhang X, Wandell B (1997) Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature* 388:68-71
- Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky EJ, Shadlen MN (1994) fMRI of human visual cortex [letter] [published erratum appears in *Nature* 1994 Jul 14;370(6485):106]. *Nature* 369:525
- Friston KJ, Price CJ, Fletcher P, Moore C, Frackowiak RSJ, Dolan RJ (1996) The trouble with cognitive subtraction. *NeuroImage* 4:97-104
- Holmes AP, Friston KJ (1998) Generalisability, random effects and population inference. *NeuroImage* 7:5754
- Johnson MK, Nolde SF, Mather M, Kounios J et al. (1997) The similarity of brain activity associated with true and false recognition memory depends on test format. *Psychol Sci* 8:250-257
- Jonides J, Smith EE, Koeppe RA, Awh E, Minoshima S, Mintun MA (1993) Spatial working memory in humans as revealed by PET. *Nature* 363:623-625
- Josephs O, Turner R, Friston KJ (1997) Event-related fMRI. *Hum Brain Mapping* 5:243-248
- Kanwisher N, Chun MM, McDermott J, Ledden PJ (1996) Functional imaging of human visual recognition. *Cogn Brain Res* 5:55-67
- Ono M, Kubik S, Abernathy CD (1990) *Atlas of cerebral sulci*. Thieme, New York
- Platt JR (1964) Strong inference. *Science* 146:347-353
- Posner MI, Petersen SE, Fox PT, Raichle ME (1988) Localization of cognitive operations in the human brain. *Science* 24:1627-1631
- Price CJ, Friston KJ (1997) Cognitive conjunctions: a new experimental design for fMRI. *NeuroImage* 5:261-270
- Puce A, Allison T, Spencer SS, Spencer DD, McCarthy G (1997) Comparison of cortical activation evoked by faces measured by intracranial field potentials and functional MRI: Two case studies. *Hum Brain Mapping* 5:298-305
- Rees G, Frith CD, Lavie N (1997) Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science* 278:1616-1619
- Sarter M, Berntson G, Cacioppo J (1996) Brain imaging and cognitive neuroscience: toward strong inference in attributing function to structure. *Am Psychol* 51:13-21
- Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR et al. (1995) Borders of multiple visual

- areas in humans revealed by functional magnetic resonance imaging. *Science* 268:889-893
- Sternberg S (1969) The discovery of processing stages: extensions of Donder's method. *Acta Psychol* 30:276-315
- Vasquez AL, Noll DC. Non-linear temporal aspects of the BOLD response in fMRI. *Proc Int Soc Magn Reson Med*, New York, NY, 1996, pp 1765
- Woods RP (1996) Modeling for intergroup comparisons of imaging data. *NeuroImage* 4:S84-S94
- Zarahn E (1998) The neural correlates of spatial mnemonic processing. PhD dissertation., University of Pennsylvania
- Zarahn E, Aguirre GK, D'Esposito M (1997a) Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage* 5:179-197
- Zarahn E, Aguirre GK, D'Esposito M (1997b) A trial-based experimental design for fMRI. *NeuroImage* 6:122-138