EXPERIMENTAL DESIGN

Hugh Garavan [1] & Kevin Murphy [2]

[1] School of Psychology and Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

[2] Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, NIH, USA.

1. Overview

Non-invasive functional neuroimaging techniques enable researchers to study the neurobiological substrates of psychological processes. The increasingly large body of neuroimaging research has two fundamental purposes. The first is to identify the brain regions that underlie a particular psychological process while the second seeks to identify differential responses of these regions to various stimuli or task challenges. The latter focus yields insights into both how the brain accommodates varying task demands and how differences between individuals or between clinical and healthy comparison groups might be explained by differences in neurobiological functioning. To achieve these goals, it is essential that one be able to isolate the psychological process of interest and how best to do so, with particular regard to experimental design, is the focus of this chapter. Part II will describe issues particular to psychological experimental design, that is, experimental control over the cognitive or emotional process of interest. Part III focuses on data analysis with emphasis on optimizing and isolating the neuroimaging signal in the activated brain regions. Part III also addresses a number of practical matters that confront all researchers when designing their experiments.

The distinction between isolating the psychological process of interest and isolating the signal associated with that process is made for pedagogical purposes. In practice, the two considerations are closely intertwined in that the experiment must be designed with a view to how the data are to be analysed. In brief, a typical analysis decomposes the time-series data into their contributing sources of variance. These sources of variance, which are generally assumed to be linearly additive, can include signals of interest such as task-induced brain activity as well as nuisance signals such as those created by headmovement, scanner signal drift or the intrusion of extraneous psychological processes. The most common method for analysing these data is a linear decomposition of the various signal sources using for example, a multiple regression in which separate regressors and planned contrasts between regressors capture both the unwanted variance and the variance of interest. Clearly, the design of the experiment needs to take into consideration what regressors and what contrasts of interest will be included in the analyses in order to ensure that the final brain activation map can be attributed to the psychological process of interest.

2. Task Design

As fMRI data are inherently noisy, it is important to induce as strong a signal as possible. This serves to maximize the contrast between the active task state and a comparison state (e.g., between a cognitive task and a visuomotor control task). In addition to maximizing contrast within an individual, it is also important to maximize contrast between individuals (e.g., between a clinical group and healthy controls) or between two times of testing (e.g., a pre-post comparison of treatment effects). To maximize the contrast between groups, it is advisable to isolate the psychological process that best discriminates the two groups. In this regard, neuroimaging researchers would be well-served by grounding their experimental methods in the relevant psychological literature that identifies the key functions that distinguish the clinical and control groups and that provides a wealth of research methods detailing how to isolate those functions experimentally.

Experimental designs in fMRI can be categorized into block, event-related and a third, broader category, labelled participant-response dependent, in which a continuous measure obtained from the participant provides a regressor for probing brain activity. The block design averages brain activation over a sustained period of time (20 to 30 seconds would be typical durations) and contrasts this with similar periods of either a resting state or a comparison task which is typically chosen to contain all task demands bar the psychological process of interest. Brain regions that differ between these conditions may then be attributed to the psychological process. In an effort to exclude signals associated with confounding physiological processes (described in detail below), aperiodic block durations, in which the alternating ON and OFF periods vary in durations, may be advisable.

This standard block design can be supplemented by including gradations of task challenge. This type of parametric manipulation can be quite advantageous: whereas the standard two-condition comparison (e.g., task A vs. task B or task A vs. rest) is open to the criticism of pure insertion (i.e., whether it is possible to selectively include and exclude a psychological function without affecting other task-related processes), the parametric manipulation assumes that the process is always present but to varying degrees in accordance with the demands placed on that process. Examples would include presenting various intensity levels of a sensory stimulus (1) or manipulating the number of memoranda in a working memory task (2). Block designs can also be enhanced by a sort of psychological triangulation in which the conjunction of distinct block design contrasts allows one to isolate a psychological process that can be separated from irrelevant surface features of the tasks (3). For example, if one wishes to isolate the neuroanatomy of the mental rehearsal component of verbal working memory, one might design an experiment using quite distinct classes of stimuli with each class accompanied by its own control comparison. One task might require participants to store a list of common nouns over a rehearsal period and recall the words after that rehearsal period. A reasonable control condition for this task might be one in which the word list remains on-screen for the duration of the rehearsal period and participants read, rather than recall, the words after the rehearsal period. The second task might present a list of nonsense syllables through earphones. At the end of the rehearsal period a single nonsense syllable is presented and participants report, using a button box, if the single item was one of the rehearsed items. A control condition for this second task might simply prompt the participant to make a predetermined button press response at the end of a delay period that was of similar duration to the rehearsal period. The conjunction between the two activation maps, in which activation for each task is first subtracted from its control condition, may be argued to represent core regions responsible for verbal working memory for which the influence of extraneous task features (e.g., linguistic stimulus properties, response modalities, recall vs. recognition) are minimized. This strength of the conjunction approach, however, may often need to be balanced against the time costs involved in testing all the required conditions.

In circumstances in which a psychological process can be isolated temporally then event-related designs are particularly useful. Here, brain activation time-locked to the events of interest can be selectively averaged enabling the researcher to embed trials of interest amidst other control trials and to categorize the trials after the participant has completed the experiment. Error trials can be excluded (or averaged separately) and events can be coded by whether a participant detected a target or not, responded relatively fast or not, produced a subsequent behaviour or not and so on (4). This affords the researcher increased flexibility in probing the dataset and has obvious advantages over the block design in circumstances in which the psychological process cannot be presented in blocks as in, for example, an oddball paradigm in which the nature of the phenomenon mandates that events are infrequent and unpredictable. The block and event-related designs can also be combined such that events of interest during an active task period can be isolated while the task period itself can be simultaneously contrasted against a control period (5). This type of mixed design provides additional information in that one can determine the inter-relationships between tonic activity levels (e.g., sustained attention or an induced emotional state) and the processing of a discrete trial (e.g., detection of a fearful face).

A final category of experimental design is what we have labelled participant-response dependent. Here, the participant provides a continuous measure that can, for example, be used to generate a regressor to correlate against brain activity measures. Despite a loss of experimental control over the participant's behaviour, this category of design affords much flexibility when the phenomenon of interest is either not strictly task-dependent or is difficult to experimentally induce. Examples include resting state acquisitions (in which correlated patterns of brain activity can be detected while the participant simply rests) (6), biofeedback (in which, for example, a participant learns to control their level of brain activity) (7), passive viewing of a movie clip (in which there may be multiple sources of stimulation with each varying with a different time-course) (8) or in which performance varies in an unpredictable manner (9). Performance modulations for which one could assess brain activation changes can be quite wide-ranging including response times or response time variability on a continuous performance task (10), frequently sampled self-report measures of mood (11) and physiological measures such as heart rate or pupil-diameter (12). In these examples the discrete measurements can be interpolated to provide a continuous time-series that can be correlated with the brain activation time-series data.

An important consideration permeating all experimental designs is the choice of baseline against which activation is contrasted. These baselines can be explicit as in the block design in which specific blocks are chosen for comparison or implicit as in the event-related and participant-response dependent designs in which the baseline is all task-related activity that is not accommodated by regressors in the data analysis. The choice of baseline determines the interpretation of what processes are captured in an activation map and requires very careful consideration by the experimenter.

2.1 Choosing an Experimental Design

The choice of which design to employ will be dictated by the particulars of the psychological process to be investigated and how easy it is to isolate. Block designs can be employed if the psychological function is easy to isolate or if it is of particular interest to compare two tasks. A simple example would be a contrast between unilateral and bilateral finger movements. Here, blocks of finger movement in just one hand could be alternated with blocks of finger movements in two hands. Rest periods might also be included in order to provide a low-level baseline against which any task-related activity could be assessed. The inclusion of a resting state baseline is generally advantageous as contrasts between two task-active periods can often be ambiguous in that greater activation in condition A relative to condition B could result from either more positive activation in A or a greater deactivation in B. A resting state baseline allows one to resolve this ambiguity by showing if activation increases or decreases in any one condition relative to the resting baseline.

If the psychological function is not easily isolated then a conjunction analysis may be useful. As can be seen in the verbal working memory example given above, the conjunction design enables the researcher to identify the core functional neuroanatomy that is common across different operationalizations of a psychological process. In addition, it can also reveal task-specific activations enabling, for example, one to determine how verbal working memory rehearsal for linguistic information differs to that of non-linguistic information. An alternative approach may parametrically manipulate verbal working memory demands by asking participants to rehearse items of varying set sizes. The presumption here is that more items will engage verbal working memory rehearsal to a greater extent resulting in changes in activation corresponding to the increased memory loads.

Although block designs suffer from an inability to isolate cognitive events that are temporally proximal by virtue of averaging over a prolonged duration, and may provide activation measures contaminated by extraneous tonic processes or isolated events (e.g., errors) they nonetheless have some advantages. For example, if the psychological process of interest by its very nature exists over a prolonged duration (e.g., sustained attention) or does not exist as a temporally discrete event (e.g., an emotional reaction) then it may be assayed best by a block design.

Conversely, the event-related design is particularly useful if one's goal is to isolate distinct cognitive events. In between-group comparisons (or time 1 vs. time 2 comparisons) the event-related design has the added advantage of being able to equate performance levels by comparing the groups on correct trials only. That is, one can compare correct performance trials of one group against the correct performance trials of the second group even if the absolute numbers of correct trials differ between the groups. In this regard, contamination from activity specific to error-related processes will not confound the between-group comparison (13). In a similar manner, selective averaging of trials may make it possible to eliminate other group differences (e.g., response speed) assuming that there are sufficient numbers of trials for this type of a matched-trial analysis. This is a particularly welcome feature as activation differences between groups that one may wish to attribute to a psychological difference can be confounded by secondary behavioural or performance differences (14). Indeed, the relationship between performance and activation levels is not straightforward. Often, researchers wish to ensure that the task produces performance differences between groups (or within a group following some experimental manipulation) in order to justify that choice of task or the focus on the psychological process engaged by the task; why study the neurobiology of attention between healthy controls and children with attention deficit hyperactivity disorder (ADHD) if the latter are not shown to be worse on the attention task? However, this can be a double-edged sword in that performance differences and knock-on effects such as differences in frustration or anxiety levels can confound interpretation of activation levels. One proposed solution is to administer a task that is within the level of competence of all participants and which, therefore, may not produce group differences in performance. Such a task can be considered a probe of the neurocognitive functioning of the groups and substantial empirical evidence shows that brain activation differences are often observed in the absence of performance differences. The typical interpretation of activation differences when there are no performance differences is that reduced activation reflects better neural efficiency and less "effort." This interpretation is supported by studies showing greater levels of activity as task difficulty increases or those that show reduced activation following practice of a psychological process (15). Finally, the participant dependent response design may be a sensible choice when one can obtain a continuous measurement from the participant but cannot exercise full experimental control over behaviour. For example, although emotional states are difficult to induce (and extinguish) experimentally, a physiological, self-report or task-induced measure can provide a time-course of that emotional state which can be used to detect correlated brain regions.

3. Optimizing Experimental Task Designs

The key issue when optimizing experimental task design in fMRI is statistical power. There are many important basic variables to be chosen which, if selected wisely, will lead to high power and thus robust and reliable results. Too often these variables are chosen arbitrarily leading to poor experimental designs that fail to yield the expected outcomes. When designing a task one needs to consider practical issues such as the number of participants or events required to give reliable results along with more analytic issues such as the estimation efficiency of the task design. These pragmatics are often dictated by feasibility constraints such as the availability of participants (e.g., how much access to the clinical population under study does one have?), the cost of scan time or the amount of available time in which the participant will remain comfortable and compliant. When it comes to the practical issues, the real question researchers have is not “How many participants/events does my study require?”, but “How few can I get away with?” Despite the ubiquity of these concerns, surprisingly few studies have addressed them and, instead, more emphasis has been placed on the analytic issues of presentation rate, duty cycles, sampling procedures, detectability of activation and efficiency of response estimation. These analytic issues relate the task that will be performed to the analysis methods that will be employed and provide guidance on the design details of an experiment. It is important, however, that analytic considerations are not allowed to dictate the design of a task such that it is no longer appropriate for measuring/engaging the psychological process under study.

3.1 Practical Issues

Only a handful of studies have addressed how many participants are required to yield stable activation maps. The first paper addressing this issue showed that conjunction analysis with a fixed-effect model is sufficient to make inferences about population characteristics thus reducing the number of participants required to infer differences between populations (16). Although quite useful, this conclusion does not give a clear indication of the number of participants required. By estimating the mean differences and variability between two block conditions, Desmond and Glover were able to perform simulation experiments generating power curves from which they could calculate the required number of participants (17). They found that for a liberal threshold of p=0.05, 12 participants were required to yield 80% power in a single voxel for typical block design activation levels. However, in fMRI the multiple comparisons problem and the associated potential for high levels of false positives requires us to go to stricter thresholds where they demonstrated that twice the number of participants would be needed to maintain the same level of statistical power. This recommended number of participants is higher than the vast majority of fMRI studies but is similar to independent assessments based on empirical data from a visual/audio/motor task (18) and from an event-related cognitive task (19). The Murphy and Garavan study (19) found that statistical power is surprisingly low at typical sample sizes (n<20) but that voxels that were significantly active from these smaller sample sizes tended to be true positives. Although voxelwise overlap may be poor in tests of reproducibility, the locations of activated areas provide some optimism for studies with typical sample sizes. It was found that the similarity between centres-of-mass for activated regions does not increase after more than 20 participants are included in the statistics. The conclusion can be drawn from this paper that a study with fewer numbers of participants than Desmond and Glover propose is not necessarily inaccurate but it is incomplete: activated areas are likely to be true positives but there will be a sizable number of false negatives. Needless to say, the required number of participants is influenced by the effect size which, in turn, is affected by the sensitivity of the experiment (e.g., the strength of the experimental manipulation, the quality of the data acquisition and the accuracy of the data analyses). These considerations may be even more important if one's intention is to detect what is likely to be an even smaller effect size of a between-group comparison.

Little research has addressed the optimal number of scans/events needed for a successful fMRI study. A simple reason for this is that there is no standard metric for determining the required number of scans/events and no gold standard for determining when the optimal number of events has been reached. One metric that has been utilized is the spatial extent of activation under the assumption that as more scans/events are included in the analysis, the spatial extent of activation will increase until all activated cortex is deemed above significance. When this occurs the spatial extent should asymptote providing an estimate of the required number of scans/events. Using this approach in a block design experiment, Saad and colleagues demonstrated that the spatial extent of activation increased monotonically with the number of scans included in the analysis and failed to asymptote after twenty-two 200s long scans (20). Similarly, Huettel and McCarthy found that the spatial extent of activation failed to asymptote even after 150 events in an event-related design (21). However, this failure to asymptote may be a consequence of the analysis method employed (22). The correlation method does not asymptote because the goodness-of-fit to the regressor continues to rise with increasing degrees-of-freedom (df), which implies that the correlation measure will never plateau by adding more time points. The Huettel and McCarthy result (21) was replicated by Murphy and Garavan (22) but they also demonstrated that when using a standard general linear modelling (GLM) analysis rather than a correlation, the spatial extent of activation asymptotes after roughly 25 events in a properly jittered event-related design. This is certainly a more attainable number of events in the available scan time of standard fMRI studies. It can be assumed that at least 25 of each type of event are needed if there is more than one psychological process under study. Also, these results have been derived from primary sensorimotor processes in the brain so it is unclear whether they will still hold for more subtle cognitive activations. Again, differences in activation have not been addressed either: it is quite possible that many more events would be required to distinguish two processes with slightly varying activation levels since these differences could be dominated by noise.

A related concern is the optimal duration of a scan. How long a scan should last is obviously dependent on how densely the required number of events can be distributed. For example, a GO/NOGO task must sparsely distribute NOGO events due to the need to build up a prepotency to respond while a simple motor response task can present the events more frequently. Other issues that limit how long a scan can last include participant comfort and ability to stay engaged in the task along with technical concerns such as throughput of data and image reconstruction times. For these reasons and more, it is common to split a scanning session into separate scans lasting 5 to 10 minutes each after which they can be concatenated into one single time-series and treated as a single scan in the analyses. However, breaks in scanning reduce the efficiency of any temporal filtering that is used and can also introduce unwanted session effects. If the goal is to detect activation then a block design is the most efficient approach (see below). In this case, the length of the scan is dependent on the amount of noise in the time series (which can be measured by calculating the temporal signal-to-noise ratio (TSNR) defined as the mean of the time series divided by its standard deviation), the size of the effect to be measured (*eff*) and the significance level (*P*) at which the activation is to be detected (23). These authors derive an equation that determines how long one needs to scan to detect activation with a block design for volumes with high spatial resolution and suggest how this can be extended to an event-related design:

$$N_G = 8\left[1.5\left(1+e^{\log_{10}P/2}\right)\left(\frac{erfc^{-1}(P)}{(TSNR)(eff)}\right)\right]^2 \qquad (1)$$

where N_G is the number of time points required for activation detection. Estimates of the size of the effect can be obtained from previous studies and TSNR measurements can be made using a short resting scan. Since these variables differ widely across types of task, brain regions and scanners, it is impractical to suggest an optimal scan duration here.

3.2 Analytic Issues

The purpose of an experimental design is to alter neural activity, and hence the blood oxygen level dependent (BOLD) signal, as effectively as possible and in a predicted way thereby enabling the researcher to detect the resulting brain changes. Using this prediction, one looks for corresponding patterns in the fMRI time series to determine which voxels were engaged in the task. A simple reference time series can be produced by convolving the stimulus timing function (which is equal to 0 when no stimulus is applied and 1 when a stimulus is presented) with a haemodynamic response function (HRF) that accurately represents the shape of the BOLD response after a single event. The gamma variate function, $y(\mathrm{t}) = t^r\, e^{-t/b}$, has been shown to effectively model the haemodynamic response to brief stimuli (24), with parameters $r = 8.6$ and $c = 0.51$, and is a popular choice for modelling the haemodynamic shape. The difference between two gamma-variates is also used in order to model the post-stimulus undershoot. It is important that the chosen HRF model accurately reflects the true shape of the response. If, for some reason, the haemodynamic shape of a participant is atypical (e.g. following treatment with a substance that directly affects the vasculature or a patient group with vascular damage), then the results of the analysis could be confounded by this difference in shape. (It should be noted that there are more advanced approaches to reference time-series formation, such as ones that use basis functions rather than a predetermined HRF shape and these are addressed in a later chapter).

The simplest type of analysis is a linear least squares regression of the equation:

$$y(t) = \beta \cdot x(t) + \alpha + \varepsilon(t) \qquad (2)$$

where $y(t)$ is the voxel time-series data, $x(t)$ is the reference function (i.e., the expected BOLD response to the stimulus) with β its scaling factor, α is a constant and $\varepsilon(t)$ is a random Gaussian white-noise term. Both β and α are unknown parameters that are fit by the linear regression method. This equation can be extended to include extra regressors to remove unwanted trends in the data, such as baseline drift, whilst simultaneously computing the scaling factor. This scaling factor, β, can then be used as an activation measure for each voxel.

Multiple reference waveforms can easily be included in this type of analysis, denoted by the term *multiple linear regression*. In an experiment with two active conditions (1 and 2) the equation:

$$y(t) = \beta_1 \cdot x_1(t) + \beta_2 \cdot x_2(t) + \alpha + \delta \cdot t + \varepsilon(t) \qquad (3)$$

is fitted to the data. For this model, four parameters are estimated, the two scaling factors β_1 and β_2 and the baseline α and also a baseline drift rate δ which accommodates for linear changes in the baseline over time. It is possible to investigate whether β_1 or β_2 are nonzero and whether β_1 is different from β_2 with statistical significance calculated using F-tests. This method allows one to identify active areas in the brain and calculate if an area is more active in one condition than another, thereby satisfying the two primary purposes of fMRI. This equation can be further generalized to $\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is a column vector of the voxel's time-series data, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is a column vector of scaling factors and $\boldsymbol{\varepsilon}$ is a column vector of Gaussian white noise terms (25, 26). This equation is called the *General Linear Model* (GLM) and is the basis for most fMRI analytic techniques. The columns of the design matrix $\mathbf{X}$ model the effects of interest and also confounding variables and are, in essence, the reference waveforms mentioned above.

When designing an experimental task, one is essentially specifying these reference waveforms/regressors. To maximize statistical power, these regressors must be chosen wisely. For example, F-tests are used to determine if there are significant differences between the regressors. To increase statistical power, one can increase the df by lengthening the task (for long TRs, each additional timepoint adds a new df). It might seem like a good idea to use extremely short TRs to increase the number of time points and hence the statistical power. However, to gain an extra df for each additional timepoint, each timepoint must be statistically independent from every other.

Unfortunately, due to autocorrelations introduced into the fMRI data by physiological noise and scanner drifts, this is not the case. This example demonstrates that knowledge of the underlying mechanisms of fMRI along with the analysis methods is required when choosing even the simplest variables (such as the number of time points and the TR) for experimental design.

Efficiency of a task design is a measure of how accurately the GLM can estimate the β weights for each of the regressors, that is, how small the predicted variance of the β estimates will be. For example, assume a block design task that induces exactly a 2% signal change in hundreds of voxels, all with different noise properties but with the same noise variance. If a GLM analysis is performed, the β estimate for every voxel will be approximately 2% for all voxels with very little deviation. Since the variance of the estimates does not differ widely with noise distributions, this would be considered an efficient task. However, matters become more complicated when there is more than one regressor. Assume that there are two block conditions, A and B, where A and B are identical with the exception that B is delayed with respect to A by one TR. These two regressors are highly correlated so if a voxel responds only to condition A, it will be extremely difficult for the GLM to distinguish this from a voxel that responds only to B. For this reason, the β estimates for each of the conditions will vary substantially and this would be considered an inefficient design. If the conditions are designed so that they have zero correlation (this is achieved by delaying B by half a block length relative to A), it would be very easy to distinguish voxels that respond to each of the conditions individually or both of the conditions together. Therefore, the variance of the β estimates would be quite small and so the design is efficient. These simple examples show that efficient task designs come from regressors that are not correlated with each other. This can be slightly complicated by the contrasts of interest. For example, say we have a jittered event-related design where conditions A and B are randomly presented. If we want to find voxels that respond only to A (i.e., a contrast matrix of C=[1 0]), only to B (i.e., C=[0 1]) or differ in their response from A to B (i.e., C=[1 -1]), this design is very efficient. However, if we want to determine voxels that respond equally to both A and B (i.e., C=[1 1]), then the design is very inefficient because such a voxel will always have an elevated activation level and therefore will be indistinguishable from a voxel that does not respond to either task. The simple idea that regressors must be minimally correlated becomes more complicated when multiple conditions, nuisance regressors and contrasts are placed into a GLM analysis. The efficiency of a task is related to the covariance of the design matrix X (i.e., all regressors expressed as columns of a matrix) and is given the formula:

$$e = \mathrm{trace}(\ C' * (X'X)^{-1} * C)^{\ -1} \qquad (4)$$

where C is the matrix of contrast weights and ' denotes the transpose of a matrix. Efficiency calculations should be carried out on all experimental designs before scanning to check that the regressors are sufficiently independent. A paper by Smith and colleagues argues against this efficiency calculation since it relates to computational precision rather than image noise (27). This paper formulates the standard efficiency equations in terms of the required BOLD effect which takes into account the strength and smoothness of the time-series noise.)

The question “how do we design a good fMRI task?” is really asking “what experimental timing will produce the most efficient design?”. There are two variables under our control, the stimulus duration (SD: defined as the length of time the stimulus is displayed) and the interstimulus interval (ISI: defined as the length of time between the offset of one event and the onset of another). Another common term is stimulus onset asynchrony (SOA) defined as SOA=SD+ISI. (Sometimes, ISI is used to mean SOA so care must be taken to understand the true meaning when reading the literature.) To maximize efficiency (i.e., minimize correlations between regressors by ensuring a clear temporal separation between the event types) one can use either a fixed ISI but vary the order of events from different conditions or one can fix the order of the conditions and vary the ISI. For example, if an event from either condition A or condition B is to be presented every TR, it is very inefficient to present the events in an alternating fashion A,B,A, … However, efficiency is increased if the order is randomized. On the other hand, if B must follow A (e.g., A is a picture of an object and the participant must respond to B, a word, deciding whether it matches the object or not), then randomizing the order is not possible. Therefore, we must vary the ISI between successive As and Bs to increase the efficiency of the design.

The issue of experimental timing is very important in fMRI tasks due to the relatively poor temporal resolution of the technique. Bandettini and Cox have shown that with a 2s SD the optimal ISI is 12 to 14 secs when the ISI is kept constant (28). At this optimal ISI, the experimentally determined functional contrast (i.e., the ability to detect activation) of an event-related task is only 35% lower that that of a block-design (which, as explained below, is the most efficient design). Simulations assuming a linear system showed that this should be 65% lower suggesting the HRF is a non-linear system. Most techniques in event-related fMRI analysis assume that the haemodynamic shape of the BOLD signal is linearly additive. It has also been shown that when the ISI is allowed to vary, the haemodynamic response shows a 17-25% reduction in amplitude when trial onsets are spaced (on average) 5 secs apart compared to those spaced 20 secs apart (29). However, power analysis indicated that the increased number of trials at fast rates outweighs this decrease in amplitude if statistically reliable response detection is the goal. So, despite the HRF being non-linear at fast presentation rates, the mismatch with the regressor is compensated by the increase in trial numbers. Dale also demonstrated that if the ISI varies, the statistical efficiency improves monotonically with decreasing mean ISI and that the efficiency can be up to 10 times greater than that of a fixed ISI design (30). These lessons on stimulus timing suggest that even though the HRF is non-linear at short ISIs, closely packed, randomly presented events produce highly efficient designs.

There are two fundamentally different goals when analysing event-related fMRI tasks: detection of signal change (which has been the focus thus far) and estimation of the HRF. Detection of the signal change involves determining one variable: the amplitude of the haemodynamic response. More information can be gleaned by estimating the HRF (e.g., time to onset, rise time, fall time, area under the curve) which can be used to determine subtle differences between groups or conditions that may not show up in an amplitude measure. However, this information comes at a cost: the experimental task can be optimized for either detection or estimation but not both. Birn and colleagues showed that the estimation of the HRF is optimized when stimuli are frequently alternated between task and control states, whereas detection of activated areas is optimized by block designs

(31). Liu and colleagues have developed a method that can simultaneously achieve the estimation efficiency of randomized designs and the detection power of block designs at a cost of increasing the length of the experiment by less than a factor of two (32). There are many programs that allow one to randomly (or not so randomly) generate thousands of task designs in order to choose the most efficient for the task at hand, be it detection or estimation. Genetic algorithms (optimization algorithms that code different designs like chromosomes and allow them to "crossover" and "point mutate" as they "replicate") that can produce designs that outperform random designs on estimation efficiency, detection efficiency and design counterbalancing have also been developed (33). Further work has also shown that using advanced mathematical techniques, block designs, rapid event-related designs, m-sequence designs (reference time series with an autocorrelation of zero) and mixed designs can nearly achieve their theoretically predicted efficiency and can be used in practice to obtain advantageous trade-offs between efficiency and detection power (34). It is important when using programs to design experiments to realize that they may converge on a structure that may be problematic for the psychological process under investigation (e.g. the most efficient task for detecting activation is a block design, however, if we want to design an oddball study the oddball events of interest should not occur in a block).

When designing a task, one must also consider the frequencies at which the events of interest are presented. Analysis packages often perform high pass filtering to remove low-frequency drifts from the data. If all frequencies below the limit of, say, 0.01 Hz are removed then the activation to a task with a block lasting greater than 100 secs will also be removed. Similarly, this would be true for event-related tasks if the events were presented at the same low frequency. Other frequencies exist in the data that one must be aware of. It is possible to remove the influence of physiological noise from fMRI data using techniques such as RETROICOR (35). These physiological noise sources are known to produce fluctuations in the data at the cardiac frequency ~1.1Hz, at the respiration frequency ~0.3 Hz and also at the respiration volume variation frequency ~0.03 Hz (36). If these techniques are to be used and the task predominantly displays power at one of these frequencies (e.g. blocks lasting 33 s have a frequency of 0.03 Hz), then the correction techniques may remove the activations of interest and not just the fluctuations due to unwanted physiological processes. Conversely, if these corrections are not used, then the GLM may denote these physiological fluctuations as activations (if the phase of the fluctuations matches the phase of the task). One must also bear in mind that when using a long TR, all frequencies will be aliased into a narrow frequency band (e.g. with a TR of 2 s all frequencies above 0.25 Hz will be aliased into the range of 0-0.25 Hz). This means that although the frequencies may seem far apart, the task and the physiological noise may alias to the same frequency (e.g., for a TR of 2s, the respiratory frequency 0.3 Hz will be aliased to 0.2 Hz as will a task frequency of 0.7 Hz, that is, one event every 1.4 sec). To avoid this problem it is best not to have the events regularly spaced so they reside at one frequency but to have random ISIs thus spreading the power to different frequencies. The most efficient tasks are ones whose power is spread widely across the whole available frequency spectrum.

4. Conclusions

Designing fMRI tasks can be difficult with logistical constraints (e.g., how many participants and how much time per participant can one afford) obliging the experimenter to optimize the study design. The emphasis here has been on the experimental and analytic means of isolating a psychological process and its associated fMRI signal. Both considerations are central: optimal efficiency is of little comfort if one measures the wrong thing but there is little to be gained from an inaccurate measurement of a robust psychological phenomenon. General recommendations include the importance of grounding one's experiment in the appropriate theoretical framework and using appropriate experimental methods, generating designs that are tested for their efficiency prior to data collection, ensuring that a sufficiently large sample is tested and being clear on whether one's goal is the detection of a response or the estimation of that response.

References

1. Helmchen C, Mohr C, Erdmann C, Binkofski F, Buchel C. Neural activity related to self- versus externally generated painful stimuli reveals distinct differences in the lateral pain system in a parametric fMRI study. Hum Brain Mapp 2006;27:755-765.

2. Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC. A parametric study of prefrontal cortex involvement in human working memory. Neuroimage 1997;5:49-62.

3. Price CJ, Friston KJ. Cognitive conjunction: a new approach to brain activation experiments. Neuroimage 1997;5:261-270.

4. Garavan H, Ross TJ, Murphy K, Roche RA, Stein EA. Dissociable executive functions in the dynamic control of behavior: inhibition, error detection, and correction. Neuroimage 2002;17:1820-1829.

5. Donaldson DI, Petersen SE, Ollinger JM, Buckner RL. Dissociating state and item components of recognition memory using fMRI. Neuroimage 2001;13:129-142.

6. Margulies DS, Kelly AM, Uddin LQ, Biswal BB, Castellanos FX, Milham MP. Mapping the functional connectivity of anterior cingulate cortex. Neuroimage 2007;37:579-588.

7. Weiskopf N, Veit R, Erb M, et al. Physiological self-regulation of regional brain activity using real-time functional magnetic resonance imaging (fMRI): methodology and exemplary data. Neuroimage 2003;19:577-586.

8. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. Science 2004;303:1634-1640.

9. Slotnick SD, Yantis S. Common neural substrates for the control and effects of visual attention and perceptual bistability. Brain Res Cogn Brain Res 2005;24:97-108.

10. Hahn B, Ross TJ, Stein EA. Cingulate activation increases dynamically with response speed under stimulus unpredictability. Cereb Cortex 2007;17:1664-1671.

11. Risinger RC, Salmeron BJ, Ross TJ, et al. Neural correlates of high and craving during cocaine self-administration using BOLD fMRI. Neuroimage 2005;26:1097-1108.

12. Kampe KK, Frith CD, Frith U. "Hey John": signals conveying communicative intention toward the self activate brain regions associated with "mentalizing," regardless of modality. J Neurosci 2003;23:5258-5263.

13. Murphy K, Garavan H. Artifactual fMRI group and condition differences driven by performance confounds. Neuroimage 2004;21:219-228.

14. Poldrack RA. Imaging brain plasticity: conceptual and methodological issues--a theoretical review. Neuroimage 2000;12:1-13.

15. Kelly AM, Garavan H. Human functional neuroimaging of brain changes associated with practice. Cereb Cortex 2005;15:1089-1102.

16. Friston KJ, Holmes AP, Worsley KJ. How many subjects constitute a study? Neuroimage 1999;10:1-5.

17. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J Neurosci Methods 2002;118:115-128.

18. Thirion B, Pinel P, Meriaux S, Roche A, Dehaene S, Poline JB. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. Neuroimage 2007;35:105-120.

19. Murphy K, Garavan H. An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage 2004;22:879-885.

20. Saad ZS, Ropella KM, DeYoe EA, Bandettini PA. The spatial extent of the BOLD response. Neuroimage 2003;19:132-144.

21. Huettel SA, McCarthy G. The effects of single-trial averaging upon the spatial extent of fMRI activation. Neuroreport 2001;12:2411-2416.

22. Murphy K, Garavan H. Deriving the optimal number of events for an event-related fMRI study based on the spatial extent of activation. Neuroimage 2005;27:771-777.

23. Murphy K, Bodurka J, Bandettini PA. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. Neuroimage 2007;34:565-574.

24. Cohen MS. Parametric analysis of fMRI data using linear systems methods. Neuroimage 1997;6:93-103.

25. Friston KJ, Holmes AP, Poline JB, et al. Analysis of fMRI time-series revisited. Neuroimage 1995;2:45-53.

26. Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited--again. Neuroimage 1995;2:173-181.

27. Smith S, Jenkinson M, Beckmann C, Miller K, Woolrich M. Meaningful design and contrast estimability in FMRI. Neuroimage 2007;34:127-136.

28. Bandettini PA, Cox RW. Event-related fMRI contrast when using constant interstimulus interval: theory and experiment. Magn Reson Med 2000;43:540-548.

29. Miezin FM, Maccotta L, Ollinger JM, Petersen SE, Buckner RL. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the

possibility of ordering brain activity based on relative timing. Neuroimage 2000;11:735-759.

30. Dale AM. Optimal experimental design for event-related fMRI. Hum Brain Mapp 1999;8:109-114.

31. Birn RM, Cox RW, Bandettini PA. Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. Neuroimage 2002;15:252-264.

32. Liu TT, Frank LR, Wong EC, Buxton RB. Detection power, estimation efficiency, and predictability in event-related fMRI. Neuroimage 2001;13:759-773.

33. Wager TD, Nichols TE. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. Neuroimage 2003;18:293-309.

34. Liu TT. Efficiency, power, and entropy in event-related fMRI with multiple trial types. Part II: design of experiments. Neuroimage 2004;21:401-413.

35. Glover GH, Li TQ, Ress D. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn Reson Med 2000;44:162-167.

36. Birn RM, Diamond JB, Smith MA, Bandettini PA. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. Neuroimage 2006;31:1536-1548.