ELEMENTS OF FUNCTIONAL NEUROIMAGING

Tor D. Wager[1*]
Luis Hernandez[2]
John Jonides[3]
Martin Lindquist[4]

[1]Columbia University, Department of Psychology
[2]The University of Michigan, Department of Engineering
[3]The University of Michigan, Department of Psychology
[4]Columbia University, Department of Statistics

Citation:

Wager, T. D., Hernandez, L., Jonides, J., & Lindquist, M. (2007). Elements of functional neuroimaging. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (4th ed., pp. 19-55). Cambridge: Cambridge University Press.

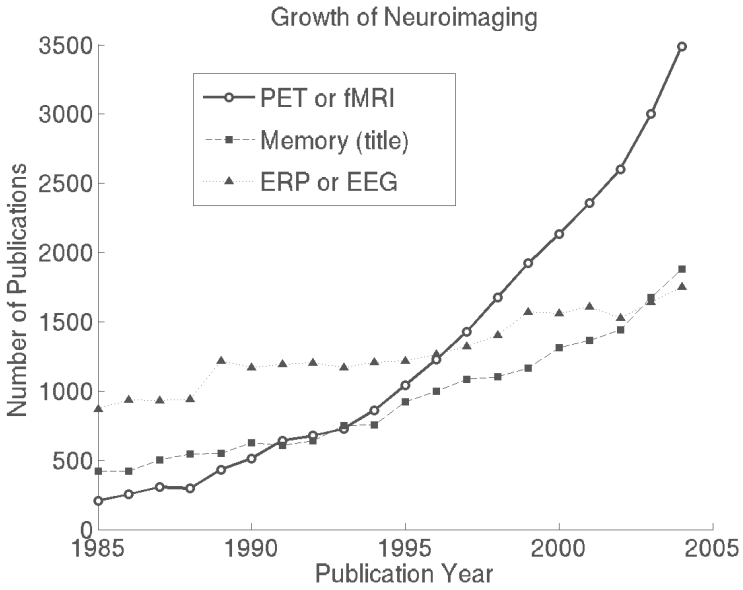
Acknowledgements

We would like to thank Dr. Doug Noll for providing Figure 9, and Brent Hughes and Dr. Matthew Keller for helpful comments on the manuscript.

There has been explosive interest in the use of brain imaging to study cognitive and affective processes in recent years. Examine Figure 1, for example, to see the dramatic rise in numbers of publications using positron emission tomography (PET) and functional Magnetic Resonance Imaging (fMRI) from 1985 to 2004. A recent surge in integrative empirical work that combines data from human performance, neuroimaging, neuropsychology, and psychophysiology provides a more comprehensive, but more complex, view of the human brain-mind than ever before. Because the palette of evidence from which researchers draw is larger, there is an increasing need to for cross-disciplinary integration and education. Our goal in this chapter is to provide an introduction to the growing field of neuroimaging research, including a brief survey of important issues and new directions.

The many aspects of PET and fMRI methodology are organized here into three sections that describe the physical, social, and inferential contexts in which imaging studies are conducted. The first section covers the physical basis of PET and fMRI imaging. This section describes the physics of each technique, what each measures, aspects of data processing, current limits of resolution, and a comparison of the relative advantages and disadvantages of these two techniques. The second section concerns aspects of neuroimaging related to social issues. In this section, we explore the kinds of questions that might be fruitfully addressed using imaging, human factors considerations when designing imaging studies, and a "road map" of an imaging experiment. The third section deals with inference in neuroimaging. It contains a review of types of experimental designs, analysis strategies and statistics, and localization of neuroimaging results in the brain. The statistical part of the section reviews the General Linear Model (GLM; the most commonly used analysis framework), hierarchical and robust extensions to the GLM that are increasingly applied to neuroimaging data, and the most commonly used multivariate analyses. In this section we also address group analyses and multiple comparisons, and pitfalls in the use of the various analysis techniques.

Although we review the physics underlying PET and fMRI here, we would like to emphasize that much of the material in the remainder of the chapter can stand on its own; the reader need not have a thorough grasp of the physics before proceeding to other sections. An outline of the major topics covered is as follows:



I. PHYSICAL CONTEXT

Imaging methods for human studies include a number of alternatives: fMRI, PET, single positron emission computerized tomography (SPECT), event-related potentials (ERP), electroencephalography (EEG), magnetoencephalography (MEG), and near-infrared spectroscopy. A number of other brain-imaging techniques are available for use in animals using radiolabeling, histological, or optical imaging techniques. Each of these techniques has advantages and disadvantages, and provides a unique perspective on the functions of mind and brain.

We focus chiefly on PET and fMRI because of their popularity and because of their potential for combining with psychological methods to unravel aspects of mind and brain. PET and fMRI offer a balance between spatial resolution and temporal resolution—in contrast to EEG and MEG, which provide millisecond timing information but with uncertain spatial localization. Finally, PET and fMRI can be used to create dynamic (over time) images of the whole brain, including deep subcortical structures whose activity is largely undetectable with EEG and MEG. This last feature offers a great potential for synergy with animal research: animal electrophysiology and lesion experiments often focus on isolated brain regions or pathways, whereas imaging can assess global function and interactions across diverse brain systems.

What PET and fMRI can measure

The number of techniques for imaging brain processes with PET and fMRI is growing. Although a thorough discussion of all of these is beyond the scope of this chapter, we provide a brief summary of them here. In the remainder of the chapter, we focus most intensively on measures of regional brain "activation" or "deactivation". We will use these terms to refer to a local increase or decrease, respectively, in signal linked to local blood flow and/or oxygen concentration. Activation and deactivation in PET and fMRI reflect changes in neural activity

only indirectly. Table 1 shows a summary of the various methods available using PET and fMRI as measurement tools. Following is a brief description of each method. In addition, a summary of the relative advantages and disadvantages of fMRI and PET is provided in Table 2.

[Insert Table 2 about here.]

Structural scans. While we are largely concerned here with functional studies, we note that MRI by itself can provide detailed anatomical scans of gray and white matter with resolution well below 1 mm^3. This can be useful if one expects structural differences between two populations of individuals, such as schizophrenics versus normal controls (Andreasen et al., 1994), or changes in gross brain structure with practice or some other variable. An example is a recent study that reported larger posterior hippocampi in London taxi drivers who had extensive training in spatial navigation (Maguire et al., 2000).

Anatomical connectivity. Another structural scanning technique is diffusion tensor imaging. This technique allows one to identify white-matter tracts (such as corpus callosum) in the human brain and changes in these structures as a function of some variable, such as age or training. Although not a technique to image brain function, "tractography," most often performed using diffusion tensor imaging (DTI) in MRI, is a technique that allows the investigator to map the white matter tracts that connect regions of the brain and hence determine the physical connectivity network underlying brain activity (Peled, Gudbjartsson, Westin, Kikinis, & Jolesz, 1998). MR images can be made sensitive to the spontaneous diffusion of water. Near a white matter tract, water diffuses most easily along the tract, producing a diffusion tensor (a generalization of a vector) that is large along the axis of the tract and small in the other dimensions. In the published literature, diffusion tensor images are usually labeled with different colors for the x, y, and z components of motion; a solid block of one color indicates fiber tracts running along either the x, y, or z-axis of the image. [1]

Table 1: Summary of PET and fMRI Methods

What is imaged	PET	fMRI
Brain structure		Structural T1 and T2 scans
Regional brain activation	Blood flow (^{15}O) Glucose metabolism (^{18}FDG) Oxygen consumption	BOLD (T2*) Arterial spin labeling (ASL) FAIR
Anatomical connectivity		Diffusion tensor imaging
Receptor binding and regional chemical distribution	Benzodiazapines, dopamine, acetylcholine, opioids, other neurochemicals Kinetic modeling	MR spectroscopy
Gene expression	Various radiolabeling compounds	MR spectroscopy with kinetic modeling

[1] A standard convention, which we adopt throughout this chapter, is to refer to locations in the brain according to their relative position from the anterior commissure. Coordinates are recorded in three dimensions: lateral (negative x values are in the left hemisphere, positive in the right), rostrocaudal (positive is anterior), and dorsal-ventral (positive is superior), referred to as x, y, and z, respectively.

DTI can be used to study the structure of fiber tracts in healthy or patient populations, or they can be used in combination with functional imaging studies. For example, a recent implementation of this approach used DTI to define adjacent regions of the anterior cingulate that receive different projections from other brain regions (Johansen-Berg et al., 2004). Subsequent

Table 2: Relative advantages of PET and fMRI

PET	FMRI
- Mapping of receptors and other neuroactive agents	- Repeated scanning
- Direct measurement of glucose metabolism	- Single subject analyses possible
- No magnetic susceptibility artifacts; better signal around brain sinuses	- Higher spatial resolution
- Quiet environment for auditory tasks	- Higher temporal resolution
- Easily combined with ERP and other measurements because there is no magnetic field	- Single trial designs; image events within trials
- Quantitative baseline possible for measuring resting-state metabolism, comparing scans across days, comparing across long time periods within a session	- Estimation of hemodynamic response and separation of stimulus and task set related variables
	- Measure dynamic connectivity (correlations) among brain regions
	- Lower cost

fMRI imaging showed that these adjacent subregions responded to different psychological tasks.

Regional brain activation. Perhaps the most frequent use of both PET and fMRI is the study of metabolic and vascular changes that accompany changes in neural activity. With PET, one may separately measure glucose metabolism, oxygen consumption, and regional cerebral blood flow (rCBF). Each of these techniques allows one to make inferences about the localization of neural activity based on the assumption that neural activity is accompanied by a change in metabolism, in oxygen consumption, or in blood flow.

Functional MRI using the Blood Oxygen Level Dependent method (BOLD) measures the ratio of oxygenated to deoxygenated hemoglobin in the blood across regions of the brain. The rationale is that a) more oxygenated blood in an area causes a decrease in BOLD signal, and b) oxygen consumption is followed by an overcompensatory increase in blood flow, which dilutes the concentration of deoxygenated hemoglobin and produces a relative increase in signal (Hoge et al., 1999); (Kwong et al., 1992); (Kwong et al., 1992). The BOLD effect is a complex interplay between oxygen consumption, blood flow and blood volume, and it is described in more detail below.

Cerebral perfusion can be measured using MRI rapidly and non-invasively by using arterial spin labeling (ASL). These measurements are a more direct, and quantitative measure of brain activity (Duong et al., 2002) but they present some technical challenges that will be

discussed in later sections.

Receptor Binding. The affinity of particular chemicals for specific types of neurotransmitter receptors offers researchers a leverage point for investigating the functional neurochemistry of the human brain. Radioactive labels are attached to compounds that bind to receptors in the brain. Labeled compounds are injected into the arteries of a subject by either a bolus (a single injection) or continuous infusion of the substance until the brain concentrations reach a steady state. This method can be used to image the density of a specific type of receptor throughout the brain. It can also be used to image the amount of binding to a particular type of receptor that accompanies performance of a task, as it was used in one study of dopamine binding during video game playing (Koepp, 1998).

The most common radioligands and transmitter systems studied are dopamine (particularly D2 receptors) using [^{11}C]raclopride or [^{123}I]iodobenzamide, muscarinic cholinergic receptors using [^{11}C]scopolamine, and benzodiazepines using [^{11}C]flumazenil. In addition, radioactive compounds that bind to serotonin, opioid, and several other receptors have been developed. Because the dynamics of radioligands are complex, a special class of mathematical models, called kinetic models, have been developed to explain the dynamic action of the labels. Kinetic modeling can allow a researcher to estimate how much of the radiolabeled compound is in the vasculature as opposed to in the brain, how much is freely circulating in brain tissue, how much is bound to the specific receptor-type under investigation, and how much is bound to nonspecific sites in the brain. Estimation of these parameters requires a detailed knowledge of the properties of the specific substances used, and how they act in the brain over time.

Having provided a brief summary of these techniques, we shall now concentrate on PET and fMRI as they are used to measure changes in blood flow and oxygenation. In the sections below, we describe the basics of what PET and fMRI measure, and how these measurements are made.

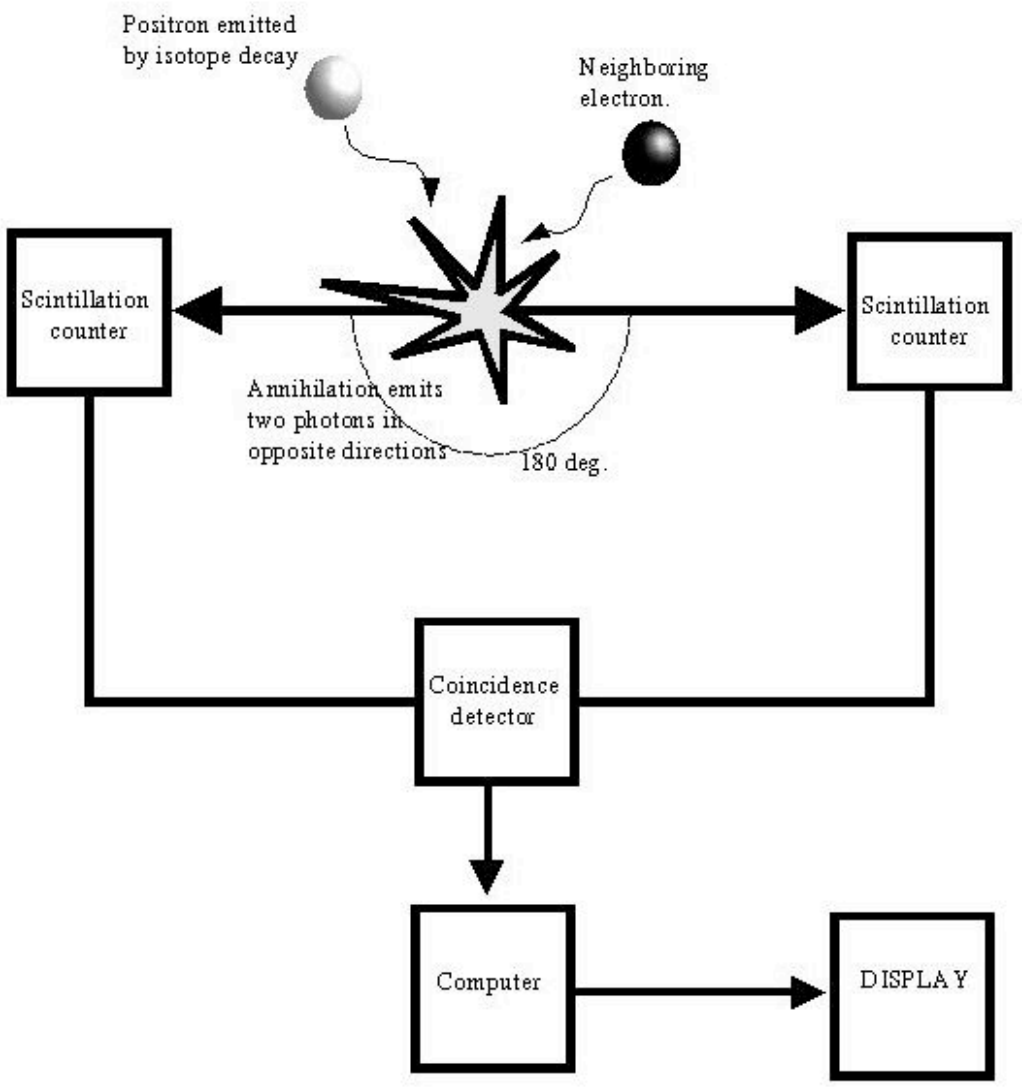
PET physics

Positron Emission Tomography provides a three-dimensional image of blood flow, glucose consumption, or neurotransmitter receptor binding by detecting positrons emitted by a radioactive tracer. Positrons are subatomic particles having the same mass but opposite charge as an electron -- they are "anti-matter electrons." The most common radioactive tracers are ^{15}O, “oxygen-15,” commonly used in blood-flow studies, ^{18}F (fluorine), used in deoxyglucose mapping, and ^{13}C (carbon) or ^{123}I (iodine), used to label raclopride and other receptor agonists and antagonists. The decay rate of such isotopes is quite fast, and their half-lives vary from a couple of minutes to a few hours, which means that a cyclotron must be available nearby in order to synthesize the radioactive tracer minutes before each PET scan.

The tracer is injected into the subject’s bloodstream in either a bolus or a constant infusion that produces a steady-state concentration of tracer in the brain. As the tracer decays within the blood vessels and tissue of the brain, positrons are emitted. The positrons collide with nearby electrons (being oppositely charged, they attract), annihilating both particles and emitting two photons that shoot off in opposite directions from one another. The photons are detected by photoreceptive cells positioned in an array around the participant’s head. The fact that matched pairs of photons travel in exactly opposite directions and reach the detectors simultaneously are important for the tomographic reconstruction of the 3-D locations where the particles were annihilated. Note that the scanner does not directly detect the positrons themselves; it detects the energy that results from their annihilation.

Depending on the design, most PET scanners are made up of an array of detectors that are arranged in a circle around the patient's head, or in two separate flat arrays that are rotated around the patient's head by a gantry. To detect simultaneously occurring pairs of photons, each pair of detectors on opposite sides of the participant’s head must be wired to a "coincidence detector" circuit, as illustrated in Figure 2**.** Small tubes (called "septa" or "collimators") are placed around the detectors to shield them from radiation from the sides and help prevent coincidences

Figure 2



due to background radiation.

The injected tracer will be distributed throughout the blood vessels and tissue of the brain (indeed, throughout the rest of the body as well). The goal of *image reconstruction* is to determine the density of radioactive counts, and thus the amount of radioactively labeled substance, in each location within the brain. Here, we describe this process for a single 2-D slice through the brain. Using mathematical notation, we use the letter **R** to refer to the location in 2-D space within the slice, and the density of tracer in each location as a 2-D function **D(R)**. The goal of reconstruction is to find **D(R)**. Each pair of coincidence detectors counts the number of positrons **P** emitted (plus error) at a particular angle throughout the entire brain slice. Positrons annihilated at a subset of locations **R** will yield coincidence counts on a particular pair of detectors. In mathematical terms, the positron count **P** at each detector is equal to the sum of the positron densities over all locations lying between the detectors:

$$P(\theta) = \sum_{r} D(r) \cdot \Delta r \qquad \textbf{(1)}$$

It's useful to think of P as a one-dimensional projection, or shadow, of the 2-D densities

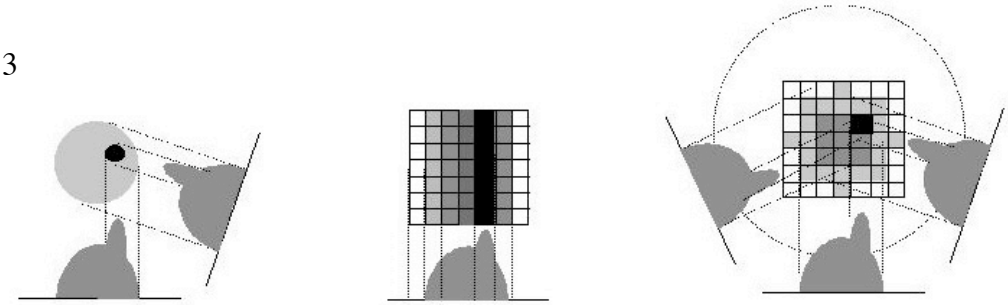
in the brain slice. The counts at each pair of detectors around the head are a projection at a different angle. For instance, in the image in Figure 3A there is a low-intensity gray circle that represents the brain, and a higher density dark circle that represents a tracer-rich area in the brain. The two detectors each contain a projection of the image (imagine a light placed behind the "brain" in Figure 3A) at a different angle. A simple way to think of reconstruction is to take each detector and assign the number of counts found at that detector to each location **r** that could have produced the counts, as shown for one detector in Figure 3B. Summing the reconstructed image over detectors at different angles (Figure 3C) yields an estimate of the original densities **D(R)**. As the number of projections is not infinite, and neither is the number of pixels in the image, some severe artifacts will occur in the image, and they must be compensated for by applying different filters to the data. This method is referred to as filtered backprojection.

In practice, this procedure is usually implemented by treating the projections as one-dimensional representations of the spatial frequency of the image. Spatial frequencies can be translated into image intensities using the inverse Fourier transform. A Fourier transform is an operation that re-expresses a signal (data collected over time or space) in terms of the power in the signal at different frequencies (spatial or temporal), and it plays an important role in both PET and fMRI data analysis. The inverse Fourier transform, FT^{-1}, expresses frequency-domain signals back in the spatial (or temporal) domain. Thus, the tracer density D(R) can be estimated by taking the inverse two-dimensional Fourier transform of the count data:

$$D(R) = FT^{-1}\left\{R \cdot P(\theta)\right\} \quad \textbf{(2)}$$

A more complete explanation of filtered backprojection and other methods can be found

Figure 3



A B C

in several good texts (Bendriem, 1998; Sandler, 2003).

What do PET counts reflect? The answer depends, of course, on what molecule the label is attached to and where that molecule goes in the brain. Ideally, for ^{15}O PET, counts reflect the rate of water uptake into tissue. 18-fluorodeoxyglucose (FDG) PET measures glucose uptake, whereas ^{13}C Raclopride PET measures dopamine binding. However, in practice the observed level of signal depends on a number of factors, including the concentration of the radiolabeled substance in the blood, the blood flow and volume, the binding affinity of the substance to receptors, the presence of other endogenous chemicals that compete with the labeled substance, the rate of dissociation of the substance from receptors, and the rate at which the substance is broken down by endogenous chemicals.

Estimation of all these parameters requires detailed knowledge of the properties of the specific substances and how they act in the brain over time. *Kinetic models* have been developed to estimate how much tracer is contained in different categories, or *compartments*, of blood and tissue. Different forms of kinetic modeling have different numbers of compartments; for example, a two-compartment model estimates how much of the radiolabeled compound is in the

vasculature as opposed to in the brain. A three-compartment model used in receptor binding studies estimates tracer quantities in blood, 'free' tracer in tissue, and label bound to receptors. Often a reference region with few or no receptors (i.e., the cerebellum for dopamine) is used to model the separation of free from bound tracer; this requires the assumption that none of the signal in the reference region comes from 'bound' tracer. A four-compartment model additionally separates tracer bound to receptors of a specific type (called specific binding) from those bound to other receptors (called nonspecific binding).

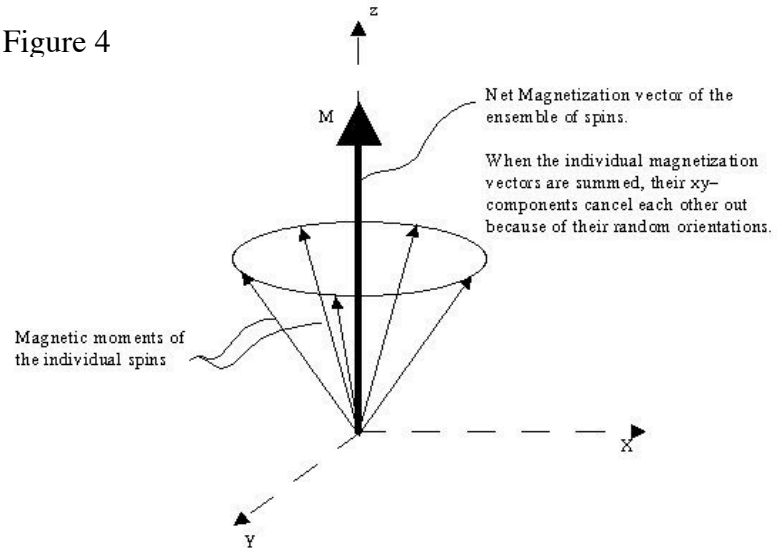
MRI physics

The raw signals in both NMR and MRI are produced the same way. As we will soon explain in more detail, a sample (e.g., a brain) is placed in a strong magnetic field and radiated with a radiofrequency (RF) electromagnetic field pulse. The nuclei absorb the energy only at a particular frequency, which is dependent on their electromagnetic environment, and then return it at the same frequency. The returned energy is detected by the same antenna that produced the RF field. Pulse sequences, or particular patterns of manipulations of the RF pulse and other magnetic fields, are used to acquire data that can be localized in the 3-D space of the brain.

The human body is made mostly of water, whose two hydrogen atoms are each made up of a single proton and a single electron. Every proton has its own magnetic dipole moment, represented mathematically by a vector in 3-D space. The magnetic moment is the amount of "magnetization" of an object, and it determines how strongly it interacts with magnetic or electric fields (a bar magnet is a dipole, and a very strong one would have a very large dipole moment).

The main magnetic field of an MRI scanner, usually labeled $\mathbf{B_0}$, is extremely powerful—anywhere from 1.5 Tesla (the standard magnet used for clinical scans, 15,000 time stronger than the magnetic field of the earth) to 8 T. To achieve such high magnetic fields requires very high electric currents. Thus, MRI magnets are typically *superconducting* closed loops of wire that carry a large current. The magnet's current is loaded ("ramped") when it is initially installed, and then it is simply allowed to flow along the closed loop perpetually. Due to the superconductive state of the coils, the magnet is always on, although no electricity is being applied to it. The drawback of this is that the magnet requires liquid helium to maintain the superconductive state, and that the perpetual field requires that the operator be constantly vigilant for safety hazards that might arise from inappropriate materials being in the environment of the magnet. Note that the potential for hazard is present even when the rest of the scanner hardware and software is turned off, because the magnetic field is always turned on. The field is most homogenous, or uniform, in the **bore** of the magnet, the hollow core in which the participant's head is situated.

Figure 4



When they are placed in the $\mathbf{B_0}$ field, the precession, or "spins," of a portion of the protons will align with or against the magnetic field. Being aligned with the magnetic field takes less energy than being aligned against it, so a greater number of the spins will be aligned in the direction of the field. The overall net magnetization of the spins in a piece of brain tissue is the *net magnetization vector*. We shall call the net magnetization vector **M**. The larger the magnetic field, the greater the proportion of spins that are aligned, and the easier the magnetization vector is to detect.

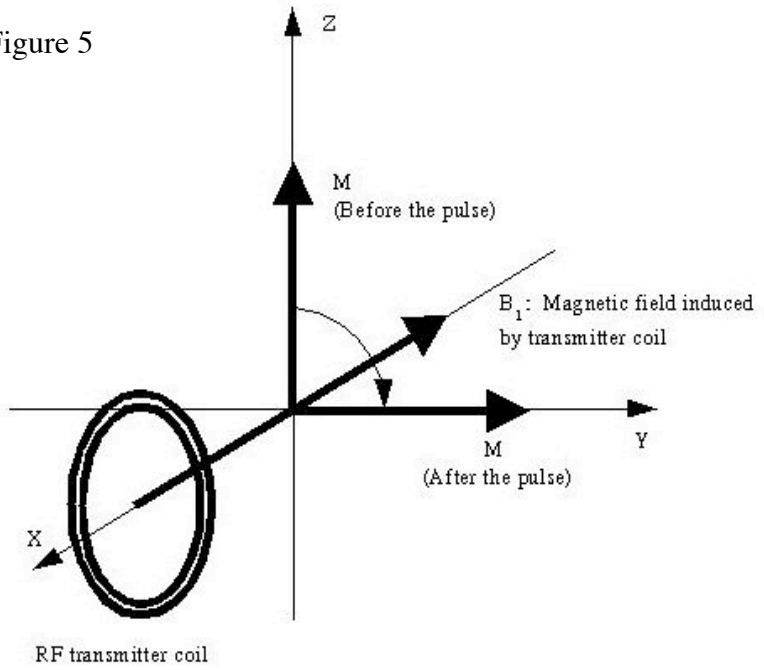
The net magnetization vector *precesses* around the axis of the $\mathbf{B_0}$ field, parallel to the long axis of the bore of the magnet. Precession is a movement that looks like a spinning top: The origin of the vector stays fixed, while the end of the vector, describing the net magnetization of the spins in 3-D space, describes a circle around the axis of the $\mathbf{B_0}$ field, as shown in Figure 4. One of the most important equations in NMR physics is that describing the rate or angular velocity of precession, also called the *resonance frequency*, ω_0 (pronounced omega-naught):

$$\omega_0 = \gamma B_0 \qquad \textbf{(3)}$$

The rate of precession equals the magnetic field $\mathbf{B_0}$, multiplied by a constant, γ (gamma), called the gyromagnetic ratio. The gyromagnetic ratio is specific for the nucleus in question; γ for hydrogen nuclei is 42.58 MHz/Tesla. By introducing systematic local variations in the magnetic field, called *gradients*, the spins precess at different frequencies at different spatial locations. This information is critical to localizing an NMR signal in space. NMR spectroscopy takes advantage of the fact that the number of electrons present, the proximity of other nuclei, and other factors can change the local $\mathbf{B_0}$ field, thus altering ω_0**.** Certain molecules can be identified by the pattern, or spectrum, of resonance frequencies in a piece of tissue.

The MRI signal is generated by introducing a second magnetic field, which we refer to as $\mathbf{B_1}$. This second field is applied by an antenna and it rotates at the same rate of precession as the magnetization vector. This rate is typically in the radio frequency range, so these pulses are also referred to as "RF pulses". The RF pulse is applied in a direction perpendicular to the main magnetic field.

Figure 5



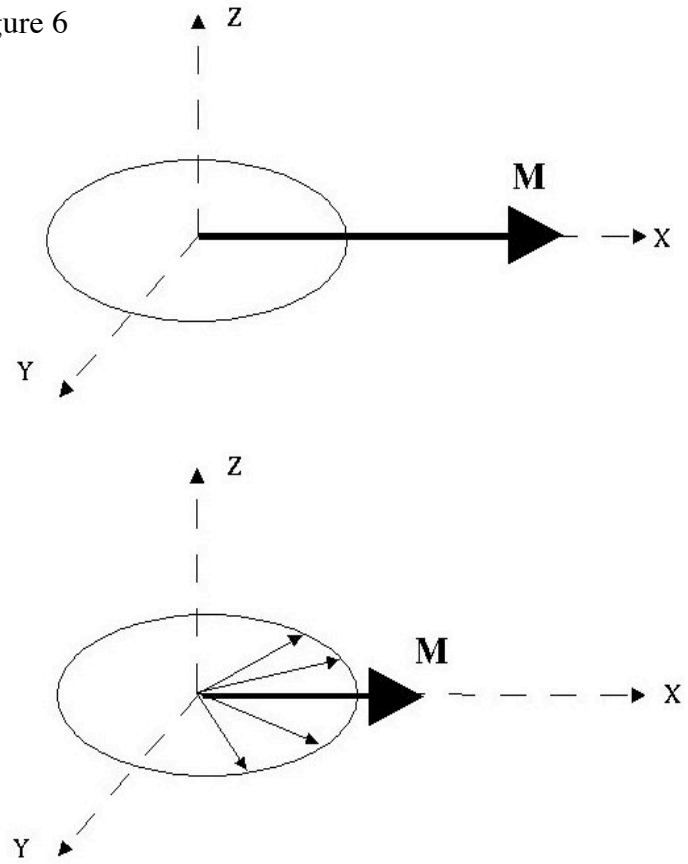
Physicists typically describe precession by referring to a "rotating frame of reference." This means that the vectors are considered from the point of view of an observer who is rotating along with the $\mathbf{B_1}$ field. To such an observer, the net magnetization vector (**M**) and the RF field ($\mathbf{B_1}$) are stationary and perpendicular to each other, but the rest of the world is spinning. In this frame of reference, the effect of a magnetic field on a magnetic dipole is a rotation of the dipole toward the transverse plane. Thus,

by applying the pulse at the frequency ω_0, the pulse tips the net magnetization vector, increasing the "wobble" in the precessing vector. The degree of tipping is called the flip angle. If the flip angle is 90 degrees, as is shown in Figure 5, the net magnetization vector will be rotating in a plane transverse to the direction of $\mathbf{B_0}$. We'll call the $\mathbf{B_0}$ direction **z**, and the axes of the transverse plane **x** and **y**.

Changes in a magnetic field will induce electrical currents in a wire coil. The antenna used for transmission of the RF pulse is such a coil, and the rotation of **M** through the plane of the antenna coil induces a current in the coil. This current induced in the coil is the NMR signal that we observe. The current oscillates at the resonance frequency, and the power (amplitude) of the signal is proportional to the degree of magnetization in the transverse plane.

When the RF pulse is turned off, the magnetization vector will "relax" back to its equilibrium position in line with $\mathbf{B_0}$. This relaxation happens through several mechanisms: 'Spin-lattice' relaxation occurs as the spins give away their energy, and they return to their original quantum state. The time this takes is called the T1 relaxation time, and it depends on the density of the microscopic environment of the water protons (hence the protons in hydrogen) in the imaged tissue. This, in turn, will be reflected in differences in brain structure such as gray matter, which has a high water content versus white matter which has a lower content. Pulse sequences sensitive to differences in T1 relaxation time are called "T1-weighted" images, and they are commonly used for high-resolution structural scans.

Figure 6



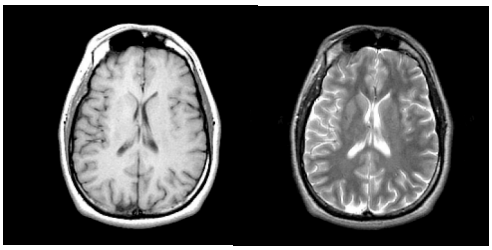
Spin-spin relaxation happens along the transverse (i.e., on the *x-y* plane) component of the magnetization vector. Recall that the magnetization vector that gives rise to the signal is made up of the sum of an ensemble of dipoles that precess at a given rate. Spin-spin relaxation is due to some of the spins rotating faster than others in the transverse plane. When this happens, the ensemble of spins get out of phase with each other and thus decrease magnetic coherence and reduce the net magnetization vector, as illustrated in Figure 6. The driving mechanism for this kind of relaxation is collisions between molecules that cause them to get out of phase with each other. The rate at which this happens is called the T2 relaxation rate. Structural images that are sensitive to the differences in T2 relaxation rates among different tissue types can also be acquired; these are called "T2-weighted" images.

Another kind of relaxation is caused by local inhomogeneities in the magnetic field. These variations cause some protons to precess faster than others. Over the course of a few

milliseconds, the protons will fall out of phase with each other, and the transverse-plane component of **M** will shrink faster than expected due to simple T_2. This change is referred to as T_2^* (pronounced 'T2 star'). A major reason for differences in T_2^* signal over the brain is variation in the local ratio of oxygenated to deoxygenated hemoglobin. A major cause of this variation is differences in regional brain metabolic processes, a signal of interest to neuroscientists. Thus, T_2^* weighted images are a reflection of changes in brain metabolism, and they are the "functional" brain images collected in fMRI BOLD (Blood Oxygen Level Dependent) imaging.

An example of the same slice of tissue imaged with T_1 and T_2 weighting can be seen in Figure 7. The images look strikingly different. Changing the contrast mechanism can be very useful in differentiating brain structures or lesions, since some structures will be apparent in some kind of images but not in others. For example, multiple sclerosis lesions are virtually invisible in T_1 weighted images, but appear very brightly in T_2 weighted images.

Figure 7



The differences between T_1, T_2, and T_2^* weighted images lie in the construction of the pulse sequence, the pattern of RF excitations and data collection periods. A thorough discussion of pulse sequences is beyond the scope of this chapter because few neuroscientists or psychologists will ever program one. Rather, this is the province of physicists and bioengineers, and this division of labor is a key indication of how the discipline of neuroimaging is a truly interdisciplinary venture. For a more detailed introduction to types of pulse sequences, we refer the reader to Huettel, Song, & McCarthy (Huettel, 2004).

Now that we have a rough idea of how a signal is produced, let us take a look at how we can extract spatial information from it so that we can form an image. Localization of the signal in 3-D space in the brain is no easy task: it requires producing changes in the signal that vary systematically with location in the x, y, and z direction. MRI and fMRI images are usually acquired slice-by-slice in the z direction; thus, localization in the z direction is handled with *slice selection*, the process of selectively exciting one slice of brain tissue at a time. Signal is localized within a slice, in the x and y directions, by *frequency*- and *phase-encoding*, respectively. These are described below.

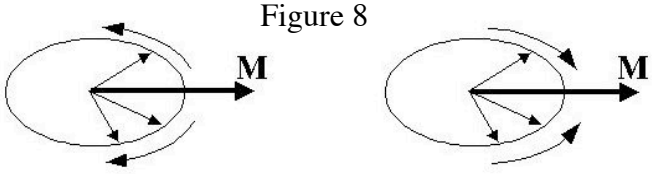
We mentioned above that the precession frequency of the spins (and thus their resonance frequency) was proportional to the strength of the magnetic field. Now, consider what happens when we apply a *gradient*—another magnetic field in the direction of $\mathbf{B_0}$, except that this field varies linearly in intensity with location along the *x*-axis (This is called the x-gradient). Since the magnetic field strength varies with position in space, so does the rotational frequency of the magnetization vector, and consequently, the frequency of the detected signal. Hence, the power of the signal collected by the RF antenna at each frequency tells us the power of the signal source at each location along the direction of the applied gradient. This technique is called 'frequency encoding.' As with PET, the Fourier transform plays a role in reconstructing the image intensities from a frequency encoded signal. In MR image reconstruction, a Fourier transform of the frequency-encoded signal will result in an intensity image. The intensity is the amount of power, or 'signal,' at each position along the direction of the frequency-encoding gradient.

In reality, things are a bit more complex. Since the spins at different locations along the x-axis are precessing at different rates in the presence of the gradient, their magnetization vectors get out of phase with each other, causing the net transverse magnetization (the overall signal) to decay quickly. The spins can be refocused in two different ways. One could apply a gradient of equal but opposite intensity to un-do the phase gain caused by the first gradient, such that the spins will regain their phase coherence and form what's called a "gradient-echo". Alternatively, one could also apply another, "refocusing" RF pulse to rotate the magnetization 180 degrees, and

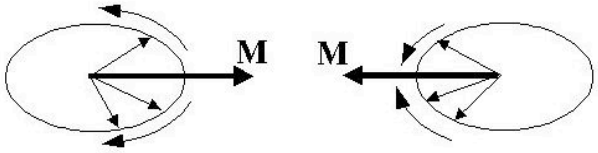
then re-apply the original x gradient so that the spins regain their phase coherence. This technique is called "spin-echo," and both techniques are illustrated in Figure 8. Most BOLD pulse sequences employ the gradient echo technique.

To encode the spatial location along the y-axis, we *very briefly* apply a gradient along the y direction prior to applying the frequency encoding gradients. The result of this gradient is a brief change in the spins' precession rate, and they end up a little bit ahead or behind the others depending on where they are along the *y*-axis. In other words, they gain or lose a little bit of phase relative to the vectors rotating at the resonant frequency, depending on their position along the y-axis and the strength of the applied gradient. This is called 'phase encoding', since the phase gained by the signal is determined by its position in space. This procedure is repeated over a range of gradient amplitudes, so that we obtain a set of signals with a *distribution* of phases along one axis and a distribution of frequencies along the other axis.

Figure 8



Gradient Echo technique: A gradient in the magnetic field causes the spins to lose phase coherence. Reversal of the gradient causes them to regain it.



Spin Echo technique: spins are dephased by the gradient. After application of a 180 degree pulse, all the spins are rotated about the y–axis, and the application of the same gradient causes the spins to regain coherence along the negative x–axis.

It is important (albeit difficult to visualize) to notice that obtaining a distribution of phase gains along the y-direction (phase encoding) is equivalent to frequency encoding, and vice versa. Effectively, there are now two spatial frequency axes, commonly referred to as k_x and k_y . You may find yourself wondering why one doesn't just apply simultaneous frequency encoding gradients along x and y directions. The answer is that doing so would result in a gradient that runs along the diagonal instead because gradients add linearly as vectors.

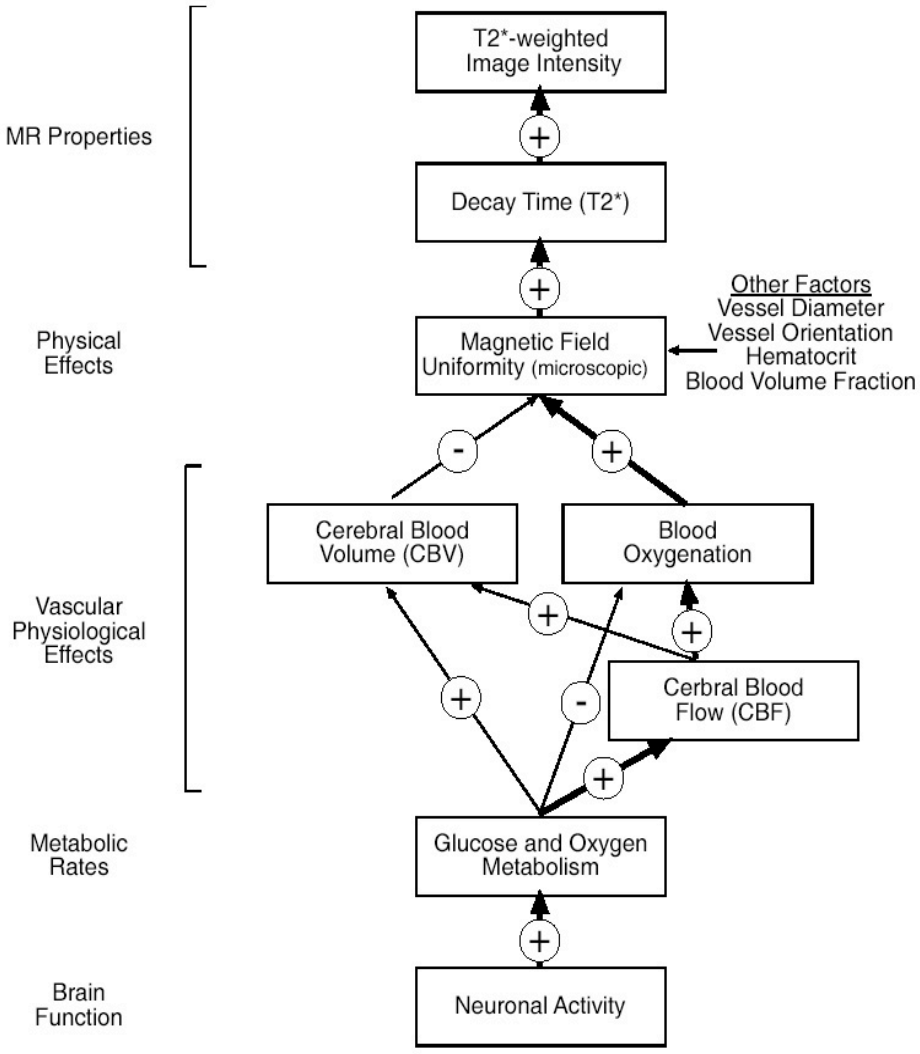
K_x and k_y data are arranged in "k-space," a 2-D matrix whose rows and columns contain the power in the signal at each frequency. The two-dimensional Fourier transform of the k-space data is an image of the signal contribution from each location in the slice.

Up to this point, we have given a brief description of one of many techniques for generating an MR image. For more in-depth information, we refer the reader to a very approachable text by Elster (Elster, 1994).

We mentioned earlier that BOLD imaging uses a T_2^*-weighted signal that depends on the oxygenation of hemoglobin. As neural activity increases, so does metabolic demand for oxygen and nutrients. Capillaries in the brain containing oxygen and nutrient-rich blood are separated from brain tissue by a lining of endothelial cells, which are connected to astroglia, a major type of glial cell that provides metabolic and neurochemical-recycling support for neurons. Neural firing signals the extraction of oxygen from hemoglobin in the blood, likely through glial processing pathways (Shulman, Rothman, Behar, & Hyder, 2004; Sibson et al., 1997). As oxygen is

extracted from the blood, the hemoglobin becomes paramagnetic—iron atoms are more exposed to the surrounding water—which creates small distortions in the $\mathbf{B_0}$ field that cause the T_2^* effect. Increases in deoxyhemoglobin can lead to a decrease in BOLD signal, often referred to as the "initial dip." The initial decrease in signal (whose existence is controversial) is followed by an

Figure 9



Courtesy Dr. Doug Noll

increase, due to an over-compensation in blood flow that tips the balance towards oxygenated hemoglobin (and less signal loss due to dephasing). It is this that leads to a higher BOLD signal. Initially, fMRI was performed by injection of contrast agents (such as iron) with paramagnetic properties, but the discovery that the T_2^* relaxation rate of oxygenated hemoglobin was longer than that of deoxygenated hemoglobin led to BOLD imaging as it is currently used with humans, without contrast agents (Kwong et al., 1992; Ogawa, Lee, Kay, & Tank, 1990).

BOLD physiology

How well does BOLD signal reflect increases in neural firing? The answer to this important question is complex, and understanding the physiological basis of the BOLD response is currently a topic of intense research (Buxton & Frank, 1997; Buxton, Uludag, Dubowitz, &

Liu, 2004; Heeger & Ress, 2002; Vazquez & Noll, 1998). Some relationships among factors that contribute to BOLD signal are summarized in Figure 9.

Essentially, the BOLD signal corresponds relatively closely to the local electrical field potential surrounding a group of cells—which is itself likely to reflect changes in post-synaptic activity—under many conditions. Demonstrations by Logothetis and colleagues have shown that high-field BOLD activity closely tracks the position of neural firing and local field potentials in cat visual cortex, even to the locations of specific columns of cells responding to particular line orientations (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001). However, under other conditions, neural activity and BOLD signal may become decoupled (Disbrow, Slutsky, Roberts, & Krubitzer, 2000). Thus, for these reasons and others, BOLD signal is only likely to reflect a portion of the changes in neural activity in response to a task or psychological state. Many regions may show changes in neural activity that is missed because they do not change the net metabolic demand of the region.

Another important question is whether BOLD signal increases reflect neural excitation or inhibition. Some research supports the idea that much of the glucose and oxygen extraction from the blood is driven by glutamate metabolism, a major (usually) excitatory transmitter in the brain. Shulman and Rothman (Shulman & Rothman, 1998) suggest that increased glucose uptake is controlled by astrocytes, whose end-feet contact the endothelial cells lining the walls of blood vessels. Glutamate, the primary excitatory neurotransmitter in the brain, is released by some 60-90% of the brain's neurons. When glutamate is released into synapses, it is taken up by astrocytes and transformed into glutamine. When glutamate activates the uptake transporters in an astrocyte, it may signal the astrocyte to increase glucose uptake from the blood vessels. Although it remains plausible that some metabolic (and BOLD) increases could be caused by increased *inhibition* of a region, in many tasks where both BOLD studies and neuronal recordings have been made, BOLD increases are found in regions in which many cells increase their activity. This is true in studies of eye movements, task switching, working memory, food reward, pain, and other domains.

Arterial spin labeling (ASL)

Blood Oxygenation Level Dependent (BOLD) fMRI is currently the dominant technique for functional imaging and has produced a wealth of information about the brain's cognitive and affective functions. However, BOLD signal is difficult to quantify in a physically meaningful way because it is a non-linear function of many physiological parameters as well as the scanner's own characteristics (Fig. 9). An alternative is to measure cerebral perfusion (or cerebral blood flow, CBF), a quantifiable physiological measure that may be more directly related to neuronal metabolism than BOLD. Recent animal imaging studies have indicated that CBF changes are more localized to gray matter than are BOLD changes, consistent with the notion that the BOLD effect is more weighted toward draining veins (Duong et al., 2002; Pfeuffer et al., 2002).

Perfusion imaging, unlike BOLD, may be used to study baseline activity (i.e., the resting state) without comparison to an active state. This type of measurement is particularly desirable for studies concerned with pathological states and/or testing the effects and specificity of different drugs, and in longitudinal studies. Even in non-quantitative studies (relative CBF), it has been shown that the variance of ASL perfusion studies across scanning sessions is dramatically less than that of BOLD, making studies that span days or even months feasible (Wang, Aguirre, Kimberg, & Detre, 2003).

In the pursuit of a practical cerebral perfusion measurement, there has been extensive work toward the development of arterial spin labeling techniques (Alsop & Detre, 1996; Kim, 1995; Wong, Buxton, & Frank, 1997) which employ magnetically labeled arterial water as an endogenous tracer. In typical ASL experiments, two images are collected: one following a tagging period, and the other after a control period during which the tag clears and the system is allowed to return to equilibrium. Subtraction of these two images yields a signal that is proportional to the amount of tag that is in the tissue, and can be used to quantify the perfusion

rate.

There are numerous ASL schemes, but they all rely on the same principles and can be categorized as one of two approaches: the label is applied in either a "pulsed" or a "continuous" manner. The pulsed approach is carried out by applying a fast magnetic inversion pulse over a region outside the tissue of interest (e.g., the neck), thus creating a bolus of inverted spins that are allowed to flow into the tissue of interest, where they are detected. The continuous approach, on the other hand, uses a long RF pulse in combination with a magnetic field gradient along the direction of flow that results in the inversion of the spins that flow through the inversion plane. The inversion pulse is usually applied until the tissue of interest is saturated with tagged spins (Williams, Detre, Leigh, & Koretsky, 1992). Whereas the pulsed approach only requires 1-2 seconds for the label to reach its maximum concentration in the tissue, the continuous approach requires approximately 3 seconds before a steady state is reached. However, because of the longer inversion times of the continuous labeling scheme, the amount of tag to be detected is much larger and thus, the SNR of the method is also larger (Wong, Buxton, & Frank, 1998).

One issue plaguing the BOLD effect technique is that, since the BOLD effect is based on sensitivity to local changes in magnetic susceptibility, artifacts due to susceptibility gradients are also greatly exacerbated. These artifacts are especially problematic in areas of the brain that lie near air-spaces, such as the roof of the mouth, nose and ear canals, as well as the sinuses. Such air-tissue interfaces create local field gradients that accelerate T_2^* dephasing (Bandettini, Wong, Hinks, Tikofsky, & Hyde, 1992); (Kwong et al., 1992); (Ogawa et al., 1990). Arterial spin labeling techniques do not require T_2^* weighting, so one can use standard Spin Echo imaging techniques to collect image data and avoid susceptibility artifacts. Functional imaging studies of the basal forebrain and orbitofrontal cortex would, thus, benefit from such techniques.
BOLD effect imaging remains the dominant technique for fMRI because arterial spin labeling techniques pose a number of challenges, too. The techniques suffer from low SNR, since less than 10% of the water in a given voxel is contributed by blood (Pawlik, Rackl, & Bing, 1981) and the label decays at a quick rate. This problem can be alleviated by increasing the amount of label that is introduced into the tissue, i.e., using longer RF labeling pulses until the tissue of interest becomes saturated with labeled blood (this approach is referred to as "continuous arterial spin labeling", or CASL). Collecting multi-slice data can also be challenging because the imaging RF pulses can interfere with the inversion label of the arterial water (Silva, Zhang, Williams, & Koretsky, 1995). The technical development of ASL techniques for use in functional MRI is an active field of research, and an increasing number of studies are being carried out with ASL.

Limitations of PET and fMRI

Spatial limitations

There are limitations restricting what both PET and fMRI can measure. Neither technique is good for imaging small subcortical structures or fine-grained analysis of cortical activations. The spatial resolution of PET, on the order of 1-1.5 cc^3, precludes experiments testing for neural activity in focused areas of the brain (e.g., mapping receptive fields of cells in visual cortex). FMRI has a greater spatial resolution, as low as 1 mm^3 but often on the order of 3 mm^3 for most whole-brain functional studies. Careful work in individual participants has demonstrated the imaging of ocular dominance columns in humans (Cheng, Waggoner, & Tanaka, 2001).

However, the high potential resolution is limited by several factors. First, fMRI researchers typically smooth (blur) the data before analysis, and some inferential methods (i.e., Gaussian Random Field Theory used in SPM, discussed later) require a high degree of blurring to be valid. Second, making inferences about populations of subjects requires analyzing groups of individuals, each with a different brain. The *normalization* procedures often used to warp individual brains into a common "reference space" introduce spatial imprecision and blurring in the group data. Third, larger activated areas are commonly taken as stronger evidence for

activation—an assumption that may be misleading for small anatomical structures! Larger areas of activation often also mean less precise localization. Thus, the effective spatial resolution of both PET and fMRI is likely to be limited not only by the technology itself, but by these other considerations. One impact of spatial limitations and intentional data-blurring is that activation in small structures, particularly compact subcortical nuclei, may be missed entirely or mislocalized. A fruitful approach for those interested in particular subcortical structures is to optimize both the physical acquisition and analysis to detect signal in these regions.

Acquisition artifacts

Artifactual activations (i.e., patterns that appear to be activation arising from non-neural sources) may arise from a number of sources, some unexpected. An early study, for example, found a prominent PET activation related to anticipation of a painful electric shock in the temporal pole (Reiman, Fusselman, Fox, & Raichle, 1989). However, it was discovered some time later that this temporal activation was actually located in the *jaw* – the subjects were clenching their teeth in anticipation of the shock!

Functional MRI signals are especially susceptible to artifacts near air and fluid sinuses and at the edges of the brain. Testing of hypotheses related to activity in brain regions near these sinuses—particularly orbitofrontal cortex, inferior temporal cortex, hypothalamus and basal forebrain, and amygdala—is problematic using fMRI, though a number of groups have used optimized acquisition and analysis schemes to deal with susceptibility. Spiral imaging sequences generally lead to signal loss (dropout) in susceptible regions, causing signal loss, whereas echo planar sequences (EPI) may lead to spatial distortion of the image and mislocalization of signal. The essential problem is that the fluid spaces cause distortions, or inhomogeneities, in the local magnetic field that are different for each participant. Some approaches are to use "z-shimming" during acquisition to make the field more homogenous (Constable & Spencer, 1999), improved reconstruction algorithms (Noll, Fessler, & Sutton, 2005), "unwarping" algorithms that measure EPI distortion by measuring inhomogeneity in the magnetic field and attempting to correct it (Andersson, Hutton, Ashburner, Turner, & Friston, 2001), and use of magnetic inserts in the mouths of participants that act as shims (Wilson & Jezzard, 2003).

Functional MRI also contains more sources of signal variation due to noise than does PET, including a substantial slow drift of the signal in time and higher frequency changes in the signal due to physiological processes accompanying heart rate and respiration (the high frequency noise is especially troublesome for imaging the brainstem). The low-frequency noise component in fMRI can obscure results related to a psychological process of interest and it can produce false positive results, so it is usually removed statistically prior to analysis.

A consequence of slow drift is that it is often practically unfeasible to use fMRI for designs in which a process of interest only happens once or unfolds slowly over time, such as drug highs or the experience of strong emotions (though there are some examples of such studies). The vast majority of fMRI designs use discrete events that can be repeated many times over the course of the experiment—for example, the most common method for studying "emotion" in fMRI is to repeatedly present pictures with emotional content.

Temporal resolution and trial structure

Another important limitation of scanning with PET and fMRI is the temporal resolution of data acquisition. The details of this are discussed in subsequent sections, but it is important to note here that PET and fMRI measure very different things, over different time scales. Because PET computes the amount of radioactivity emitted from a brain region, at least 30 seconds of scanning must pass before a sufficient sample of radioactive counts is collected. This limits the temporal resolution to blocks of time of at least 30 seconds, well longer than the temporal resolution of most cognitive processes. For glucose imaging (FDG) and receptor mapping using radiolabeled ligands, the period of data collection for a single condition is much longer, on the order of 30-40 minutes.

Functional MRI has its own temporal limitation due largely to the latency and duration of

the hemodynamic response to a neural event. Typically, changes in blood flow do not reach their peak until several seconds after local neuronal and metabolic activity has occurred. Thus, the locking of neural events to the vascular response is not very tight. Because of this limitation, a promising current direction is the estimation of the onset and peak latency of fMRI responses, and other parameters, averaged over many trials (Menon, Luknowsky, & Gati, 1998). We provide a more thorough discussion of this and related issues in the General Linear Modeling section below.

II. SOCIAL AND PROCEDURAL CONTEXT

Uses of data from functional neuroimaging

A fundamental question in neuroimaging research is what one hopes to achieve with the chosen method. We begin our discussion of the social context of neuroimaging with a discussion of the types of questions that neuroimaging is and is suited to answer. Embarking on neuroimaging research requires a solid grasp of what kinds of imaging results would constitute evidence for a psychological or physiological theory, and a grounded understanding of what kinds of results are likely to be obtainable. Following this discussion, we provide a 'road map' of the stages of an fMRI experiment and some of the human factors issues involved in conducting an fMRI study.

Brain mapping: Learning about the brain

Perhaps the most obvious rationale for conducting functional neuroimaging experiments is to correlate structure with function. Through a combination of animal, human patient, neuroimaging, and neurophysiology, we now know that there is substantial localization of many functions in the neural tissue of the brain.

The majority of neuroimaging studies to date may be classified as brain mapping studies, in which investigators are exploring the patterns of activity elicited by a particular psychological process or condition. Brain mapping has been used to investigate virtually every field of study in human psychology and psychiatry, including (for example) attention, perception, memory, learning, emotion, reward, depression and other mood disorders, psychopathy, and Parkinson's and other neurological disorders. Recently, this trend has broadened to include mapping of social processes such as representations of the "self" and of others' intentions, economic principles such as expected utility and risk, and emotional self-regulation. Another approach particularly relevant here is the mapping of brain regions that correspond with changes in the autonomic and endocrine systems as measured by heart rate (Critchley et al., 2003), electrodermal responses (Critchley, Elliott, Mathias, & Dolan, 2000), pupil dilation (Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003), and cortisol release (Dedovic et al., 2005; Oswald et al., 2005).

Although data are typically analyzed and described in terms of activation changes corresponding to a psychological process region-by-region in the brain (e.g., the "role of the insula in process X"), a more comprehensive view is that psychological processes arise from interactions among distributed networks of neurons. It is quite possible that patterns of *functional connections* among different brain regions may best characterize tasks, and multivariate brain mapping is likely to become more prevalent in the future (Beckmann & Smith, 2005; Calhoun, Adali, Stevens, Kiehl, & Pekar, 2005; W. D. Penny, Stephan, Mechelli, & Friston, 2004; Roebroeck, Formisano, & Goebel, 2005).

Overall, the sort of behavioral neurology that is provided by studies of functional neuroimaging is quite helpful on several fronts. A detailed mapping of the functions of various brain structures will give us solid evidence about the primitive psychological processes of the brain. It will also provide detailed information for neurosurgical planning. Thus, if there were no other reason to conduct studies that use functional neuroimaging, mapping the brain would be sufficient reason. However, there *are* other reasons as well. Two major ones include psychological inference and prediction of future brain and psychological states.

Psychological inference: Learning about psychology

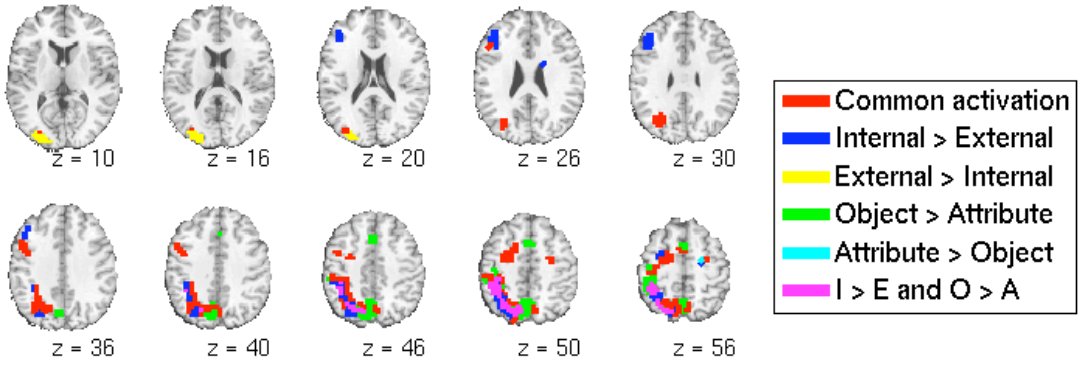
For reasons that we discuss more extensively below, brain mapping can teach us much about the gross organization of the brain, but it isn't likely in and of itself to be very informative about *psychological mechanisms*. For example, researchers may want to know *how* people make decisions, whether there are *different kinds* of cognitive control processes and how they are organized, or whether there are *distinct mechanisms* for different kinds of memories. Much of cognitive psychology has focused on identifying the component parts of systems such as decision-making, attentional control, and memory, among other topics. This has been approached in two major ways: by trying to identify behavioral dissociations in performance between different processes, and by trying to identify the similarities in performance between different processes. The logic of both approaches can also be applied to measures of brain activation.

Studies of dissociations and associations include assessment of two or more tasks studying the same experiment. In a dissociation design, researchers test whether particular brain regions are more active in one condition than in another. A double dissociation occurs when manipulation A affects Task 1 more than Task 2, while manipulation B affects Task 2 more than Task 1. In a brain imaging study attempting to establish a double dissociation, investigators might look for two regions, one of which is more active in Task 1 than Task 2, and the other of which is more active in Task 2 than Task 1. Of course, a region need not be limited to one coherent piece of brain tissue; it might well consist of a network of brain areas. Finding a dissociation of this sort is typically considered evidence for separability of the two tasks, in that they each engage component processes differentially.

A recent study in our laboratory illustrates this approach. We sought to discover whether different kinds of shifts of attention activate different brain areas. One kind of attention-shift is required when one must change which of several objects is currently relevant for behavior. For example, an airline pilot may be monitoring the cockpit windshield, and shift attention to the radar display to check for other nearby objects. Another type of attention-shift is between attributes of objects: a pilot may switch attention from the location of objects on the radar to their apparent velocity. In addition to the object and attribute switching described above, the study also manipulated whether attention-switching was performed on items stored in working memory ("internal" representations) or whether they were available on-screen ("external" representations). Figure 10 shows results from different types of attention shift in a group of 38 participants.

In a double-dissociation design, investigators might want to know whether there are areas that respond more to object-switching than to attribute-switching, or vice versa. For instance, in Figure 10, the region in left inferior frontal gyrus shown in blue responds more to switches between internal than external representations. The extrastriate region shown in yellow, by contrast, responds more to switches of external representations than internal ones. We may

Figure 10



conclude from this finding that the two types of attention-switches involve some different processes. There are limitations to this kind of inference, and we discuss them below.

An alternative way to learn about the relationships among different tasks from brain imaging is to examine the similarity of activation patterns across tasks. The logic is that the more similar the activation maps are for a pair of tasks or states, the more components the tasks have in common. This logic has been used in studies of individual differences in performance (Miyake et al., 2000). For example, if math scores and reading scores are highly correlated across individuals, then math and reading are assumed to share some common underlying performance components. In the brain imaging study on attention-switching, one might examine the activation patterns for different kinds of task-switching and quantify the degree of overlap among them. In this case, the substantial overlap in activations among all four types of attention-switching suggests that common mechanisms are shared across all the tasks (red in Fig. 10). Qualitative analyses of the similarities among activation patterns across tasks are common in the literature. What is critical is to have quantitative analyses of those similarities so that one can determine just how much overlap there is in brain activation between two tasks.

A third approach to use neuroimaging data is to study a psychological process by studying a physiological *marker* for that process. This approach has been taken in the psychophysiological literature—e.g., using cardiovascular responses as a marker for discrete or diffuse affective processes—although many researchers in that field are well aware of the problems in equating a physiological process with a psychological one. We discuss issues with this kind of psychological inference more extensively below.

Human factors in functional neuroimaging

Many types of studies are not easily adaptable to the neuroimaging environment. Here, we briefly discuss human factors limitations and some safety issues to consider when embarking on a neuroimaging study. A complete review of safety concerns is beyond the scope of this chapter (see (Elster, Link, & Carr, 1994) for more detail), but we cover a few of the basics here.

First, the MR environment is highly magnetic, and the magnetic field in the scanner is always on (barring unusual events that may require a shutdown). The strength of pull of ferromagnetic objects toward the bore of the magnet increases with the *square* of the closeness to the magnet, and the force increases with the mass of the object. Metal objects can fly out of hands or pockets as one moves closer to the bore, and larger objects pose a serious health hazard because they become missiles in the MR environment. Metallic objects in and around the participant's body can become dislodged in the intense magnetic field and cause serious injury or death. Because of the intensity of the magnetic field, participants who have implants that have any chance of being metallic (e.g., pacemakers or any electronic implants) are to be kept out of the magnetic environment. In addition, the presence of metal otherwise in the body (as in metal workers who may have tiny metallic shreds in their eyes) is a strict contraindication for scanning.

Electromagnetic fields can cause heating of tissue if applied with sufficient intensity and over enough time. This is typically not an issue, as most MRI scanners have built-in safeguards to prevent too much RF power deposition into the subject. However, should there be any metal conductors inside the RF coil, they can become quite hot because of induced currents, and they cause burns even at RF power levels that would otherwise be harmless to the participant. This is the same principle that underlies the production of sparks when you put metal objects in the microwave oven. Third, in some rare instances, changes in the magnetic field produced by the gradient coils can induce electric currents in long nerves causing them to depolarize and produce mild twitching. This is referred to as peripheral nerve stimulation (PNS) and occurs only rarely during fast imaging sequences but is more likely at higher field strength.

A safety concern particular to the PET environment is radiation exposure, which is limited by FDA regulations of 5 rem per session or 15 rem annually. NIH guidelines are 3 rem within 13 weeks or 5 rem annually (http://www.cc.nih.gov/ccc/protomechanics/chap_6.html)

In addition to safety precautions, there are other considerations that one must address to protect the integrity of the acquired data. For both PET and fMRI, the participant must remain

motionless for the duration of the session, but particularly while imaging data are being collected. PET tolerances for head movement are generally higher, but task-correlated head movement can be a serious confound in either modality. In fMRI, head motion is particularly problematic, as it induces changes in the local magnetic field. While linear motion-correction algorithms generally do a reasonable job at correcting for gross displacement of the head, they do not correct for the more complex artifacts created with movement. Participants' heads are usually restrained with a vacuum bag (a soft pillow that becomes hard when air is pumped out of it), a forehead strap and foam pads, a bite bar, or some combination of these restraints. Participants who move too much are excluded from analysis.

The restrictiveness of the scanning environment means that it is not generally advisable to use neuroimaging for tasks in which head movement is unavoidable (e.g., studies involving overt speech or pain studies involving sudden-onset electric shocks). Researchers have partially circumvented the problem and collected vocal responses during tasks by pausing data collection during verbalization or including a set of regression predictors to account for motion artifacts (Frackowiak, 1997), but the latter in particular may be only a partial solution.

The enclosed MR environment can also be a problem for individuals with claustrophobia. It is a good idea for groups working with special populations—children and individuals with psychiatric disorders—to familiarize participants in a mock-scanning environment (including an enclosed bore and simulated scanner noise) before they enter the magnet proper.

Functional MRI scanning creates repeated loud tapping and buzzing noises (approximately 100dB at the patient's location), which makes auditory presentation of stimuli more difficult than presentation in other sensory modalities. The noise can be reduced with earplugs and shielded earphones, and so some kinds of auditory studies are possible. However, earphones, like other electrical and electronic devices that may be present in the scanner room, may cause magnetic susceptibility artifacts in the images.

For visual stimulation participants in PET and MRI scanners typically view a visual display either projected by an LCD display onto a screen in the magnet room, on a shielded in-scanner LCD display, or projected onto each eye with fiber optics or LCD screens mounted in goggles. The screen is often projected onto participants' retinas through small mirrors mounted in the head coil (in MRI). The visual angle of presentation is often limited to about 15 degrees. The contrast and display image quality should be assessed before imaging, particularly for tasks that require the viewing of photographs or other fine-grained visual discriminations.

Response devices vary from scanner to scanner, but often the options are limited because in-scanner devices must be adequately RF-shielded. Responses are often limited to pressing buttons, making eye movements (several manufacturers provide scanner-compatible eye trackers), and moving joysticks or track-balls.

In general, any device that has wires or ferromagnetic parts can induce artifacts into the images because they distort the magnetic field. If severe, these artifacts may be visible as stripes or distortions in the structural (T1 or T2) images. However, T_2^* contrast is much more sensitive to artifacts, so distortions may not be visible in the structural scans but present in the functional data. Such artifacts have been observed (in our experience) with electrodermal response (EDR) leads, earphones, joysticks, mice, and keyboards in the scanning room. Some of these were not designed for use in the scanning environment, so there is no surprise that these result in artifacts. However, other commercial products intended to be RF-shielded for fMRI use also may lead to artifacts in images. In general, it is advisable to look carefully at structural and functional scans (and do a complete analysis of a simple paradigm such as visual stimulation) with and without any new piece of equipment in the room.

RF artifacts can also be caused by improper shielding of electrical cables running through the wall from the control room into the scanner room, and even by the use of hair products and makeup that contain ferrous material (not uncommon!) The closer a metallic object is to the patient's head, the greater the potential for artifacts. Jewelry, watches, and credit cards should

never be taken into the MR environment (electronics are likely to be destroyed). Small metallic objects far from the participant's head, such as buttons on jeans, haven't presented a problem in our experience.

Another concern is that the neuroimaging environment itself may change performance of the task and other physiological measures. Changes may be due to anxiety about the scanning environment, changes in temperature (many scanner rooms are chilly), changes in posture that induce physiological changes (lying down reduces orthostatic load), practice, a response to the medical context presented by imaging suites, or other variables. It is advisable to test paradigms outside the scanner, and use objective measures of performance and other behaviors whenever possible.

Data Preprocessing

Neuroimaging, and particularly fMRI, data undergo substantial processing before data analysis. There are several conditions about the fMRI images that must be met in order to carry out a successful data analysis. Most analyses are based on the assumption that all the voxels in any given image were acquired at the same time. Second, it is assumed that each data point in the time series from a given voxel were collected from that voxel only (i.e., that the participant did not move). Third, it is assumed that the residual variance will be constant over time and have a Gaussian distribution. Additionally, when carrying out analyses across different subjects, we assume that each voxel occupies the same location within the brain for all the subjects in the study. Without any pre-processing, none of these assumptions are entirely true, and the assumptions will introduce errors in the results. Here, we describe preprocessing for an fMRI experiment; other types of neuroimaging will require different steps.

Reconstruction. Images are reconstructed from data in k-space and transformed into image space. Raw and reconstructed data are stored in a variety of formats, but reconstructed images are generally composed of a 3-D matrix of data, containing the signal intensity at each "voxel" or cube of brain tissue sampled in an evenly-spaced grid, and a *header* that contains information about the dimensionality, voxel size, and other image parameters. A popular format is Analyze, also known as AVW, which uses a separate header file and image file for each brain volume acquired. Other formats, such as AFNI, are also gaining popularity. A series of images describes the pattern of activity over the course of the experiment. It is also common to store images in a 4-D matrix, where the fourth dimension is time.

Slice Timing. Statistical analysis assumes that all the voxels in an image are acquired at the same time. In reality, the data from different slices are shifted in time relative to each other—because most BOLD pulse sequences collect data slice-by-slice, some slices are collected later during the volume acquisition than others. Thus, we need to back calculate what the signal intensity of all the slices would have been at the same moment in the acquisition period. This is done by interpolating the signal intensity at the chosen time point from the same voxel in previous and subsequent acquisitions. A number of interpolation techniques exist, from bilinear to sinc interpolations, with varying degrees of accuracy and speed. Sinc interpolation is the slowest, but generally the most accurate.

Realignment. A major problem in most time-series experiments is movement of the subject's head during acquisition of the time series. When this happens, the image voxels' signal intensity gets "contaminated" by the signal from its neighbors. Thus, one must rotate and translate each individual image to compensate for the subject's movements.

The coordinates of a point in space can be expressed as a vector. It can be shown that the coordinates of a given point in space after any given translation, rotation, or combination of both, can be calculated by multiplying a matrix by the original vector. Such a matrix is called an *affine transformation* matrix. Multiplying all the voxel coordinates of an image by the same matrix will rotate and translate the entire image. Thus, in order to undo the rotation and translation of the head, we begin with a reference image (popular choices are the first image or the mean image) and transform all the other images in the time series to match it. For each image in the times

series, we calculate the elements in a six-parameter affine transformation matrix that corresponds to the displacement (x, y, z) and rotation (roll, pitch, yaw) of the head relative to the reference image. Usually, this is done by a least squares approximation that will minimize the difference between the image to be corrected and the reference. Multiplying by the matrix of best-fitting affine parameters applies the transformation and adjusts the image so that it matches the reference. Realignment corrects adequately for small movements of the head, but it does not correct for the more complex spin-history artifacts created by the motion. The parameters at each time point are saved for later inspection and are often included in the analysis as covariates of no interest.

Smoothing. Currently, many investigators apply a spatial smoothing kernel to the functional data, blurring the image intensities in space. This is ironic, given the push for higher spatial resolutions and smaller voxels—so why does anyone do it? One reason is that in a group analysis, voxels are assumed to occupy the same brain space across individuals. Smoothing can help minimize errors in inter-subject registration and normalization (see below). A second reason is that a popular choice for correcting for multiple comparisons, Gaussian Random Field Theory, assumes that the signal over space is a continuous field, and that the images have normally distributed noise. This is not the case in most experiments, since the signal is often correlated among different voxels, especially in fMRI experiments. In order to make the noise in the images meet the assumption, the images are convolved with a Gaussian kernel (a 3-D normal probability density function), which gives the noise a more Gaussian distribution. The kernel is often described by the full width of the kernel at half its maximum height ("FWHM") in mm. One estimate of the amount of smoothing required to meet the assumption is a FWHM of 3 times the voxel size (e.g., 9 mm for 3 mm voxels).

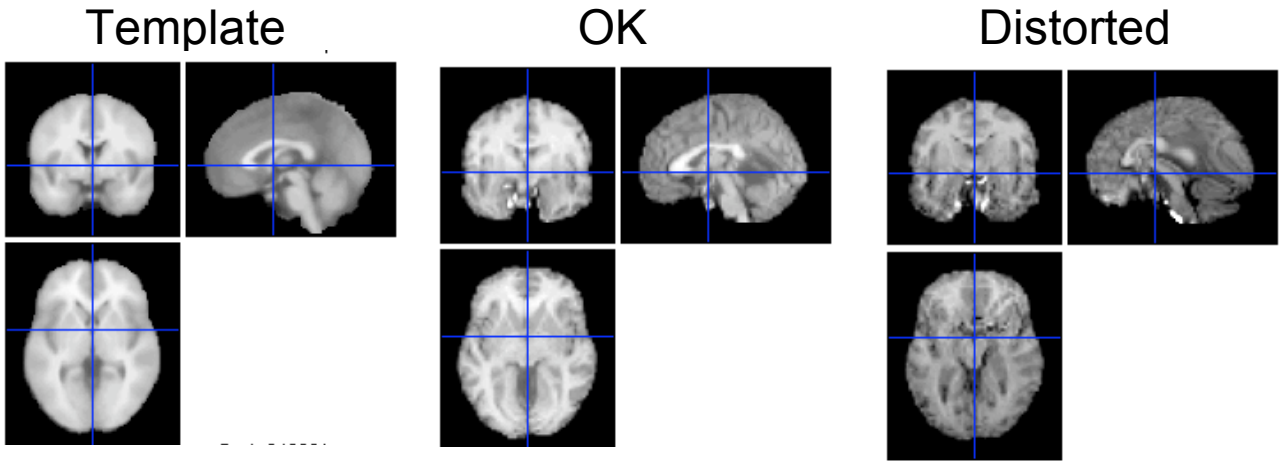
Acquiring an image with large voxels and acquiring with small voxels and smoothing an image are not the same thing. The signal-to-noise ratio during acquisition increases as the square of the voxel volume, so acquiring small voxels means that signal is lost that can never be recovered. Thus, it is optimal from a signal-detection point of view to acquire voxels at the desired resolution and *not* smooth the images, using other methods (e.g., the nonparametric methods described below) in the statistical analysis.

Coregistration. Because the high-resolution structural images contain so much more anatomical information than the functional images, it is useful for localization and inter-subject normalization to make sure that the image locations correspond to the same brain regions in both structural and functional images. A problem is that the functional and structural images are collected with different sequences and in different image dimensions. This means that though the images correspond closely, the intensity values between the two images do not map in a monotonically increasing fashion. For example, the brightness (intensity) of gray, white, and ventricular tissue types may be ordered W – G – V in the T_2^* images, and V – G – W in a T2 image. Minimizing squared errors is inappropriate in this case, but an affine transformation matrix is often estimated by maximizing the *mutual information* among the two images, or the degree that knowing the intensity of one can be used to predict the intensity of the other (Cover & Thomas, 1991). Often, a single structural image is co-registered to the first or mean functional image. Other schemes co-register a high-resolution structural image to an in-plane structural image (in the same image space as the functional images) acquired immediately before acquisition of the functional images.

Normalization or warping. In order to make quantitative comparisons across subjects, the corresponding brain structures must have the same spatial coordinates in the images. Of course, this is usually not the case, since each individual has a uniquely shaped brain. We can however, stretch and compress the images in different directions so that the brain structures are in approximately the same locations. Usually we normalize all the brain images so that they will match a standard *template* (or target) brain (e.g., the Montreal Neurological Institute brain templates).

Whereas the realignment and co-registration procedures perform a *rigid body* rotation, normalization can stretch and shrink different regions of the image to achieve the closest match.

Figure 11



The warping is described by a set of cosine basis functions, whose coefficients must be estimated by a least squares error-minimization approach. How closely the algorithm attempts to match the local features of the template depends on the number and spatial frequency of basis functions used. Often, warping that is too flexible (many basis functions) can produce gross distortions in the brain, as local features are matched at the expense of getting the right overall shape, as shown in Figure 11. This happens essentially because the problem space is too complex, and the algorithm can settle into a "local minimum" solution that is not close to the global optimal solution (Frackowiak, 1997).

Following preprocessing, statistical analysis can proceed. The next section describes task design and statistical analysis in more detail. Together, these comprise the inferential context for neuroimaging.

III. INFERENTIAL CONTEXT

The inferential context for neuroimaging studies includes how tasks are designed (including special considerations for neuroimaging studies), how data analysis is conducted, and how results are interpreted in light of other findings—sometimes converging ones from other methodologies. We begin with a discussion of the logic and limitations of psychological inference from brain imaging studies, and we then return to issues of task design and analysis. We believe it is important to begin at the end, so to speak, because limitations in the kinds of inferences that can be made are a major constraint on choices of designs and analyses.

Forward and reverse inference

Once brain mapping has revealed a set of regions that are consistently activated by a task, it is tempting for scientists and lay people alike to believe that this is a brain network that implements the psychological processes involved. Two kinds of inferences emerge from this belief. First, researchers may conclude from a set of data that some brain regions are activated by a task. This is *forward inference*, inference about the brain given a particular psychological manipulation. The second inference is that activation of some regions implies that a particular psychological process has occurred. This is *reverse inference*, inference about psychology given brain activation.

Reverse inference treats a brain region--the anterior cingulate for pain or the caudate for reward--as a marker for a psychological process. That is, if one observes activity in a region, then the corresponding psychological process is assumed to have been engaged. A number of theoretical and practical problems with making reverse inferences about psychology have been

described (Sarter, Berntson, & Cacioppo, 1996). However, the temptation to make premature inferences about psychology from brain activations is strong, because the major motivation for conducting research on the brain and behavior is to learn about psychological processes.

Researchers have inferred that romantic love and retribution involve "reward system" activation because these conditions activate the caudate nucleus (Aron et al., 2005; de Quervain et al., 2004), that social rejection is like physical pain because it activates the anterior cingulate (Eisenberger, Lieberman, & Williams, 2003), among countless similar conclusions. The trouble is that both these regions are involved in motor control and cognitive planning and flexibility in a wide range of tasks, including basic shifting of attention, working memory, and inhibition of simple motor responses (Bush, Luu, & Posner, 2000; Kastner & Ungerleider, 2000; Paus, 2001; Wager, Jonides, & Reading, 2004; Wager, Jonides, Smith, & Nichols, 2005). One meta-analytic review concluded that cingulate activity was related most reliably to "task difficulty" In a large range of tasks (Paus, Koski, Caramanos, & Westbury, 1998). Thus, the assumption that one can make reverse inferences in this case is seriously flawed.

These kinds of inferential difficulties are not unique to the brain imaging community. Current issues raised in interpreting brain imaging studies seem eerily similar to the debates over electrical brain stimulation in the 1950's. In a famous case, Jose Delgado claimed that he had identified the aggression centers of the brain (also, incidentally, in the caudate nucleus), and by remote-control stimulation of an implanted electrode could make a raging bull as passive as a gentle lamb. Delgado demonstrated his findings by placing himself in a ring with a Spanish bull and using his device to pacify the animal in mid-charge (Valenstein, 1973). The experiment worked: the bull stopped—but not likely because he was truly pacified. Valenstein (1973) recounts that the electrodes were placed in motor-control regions of the caudate that forced the bull to turn in one direction (and probably confused the animal as well!).

It is useful here to analyze what went wrong in formal terms. What Delgado observed in the previous example is the disruption of aggressive behavior (B), which is "consistent with" the hypothesis that brain circuits for aggression were stimulated (A). That is, there is a high probability of observing B, given the truth of hypothesis A, or P(B | A) -> 1. But "consistent with" is not good enough. He made an erroneous inference about the probability of hypothesis A given the data B, and concluded that P(A | B) is high. According to probability theory, these two statements are not equivalent, but rather P(A | B) depends on P(B | A), and also on the base rates (prior probabilities) of A and B, as stated by Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \textbf{(4)}$$

In ignoring what else besides A might have disrupted behavior, they ignored the base rate of B, and reached a false conclusion.

In our brain imaging example, if the base rate of anterior cingulate activation across studies is high--if it's activated in every study--then its activation tells us very little about the engagement of any particular psychological process.

Types of experimental designs

Neuroimaging designs require making tradeoffs. One basic tradeoff is between experimental power and the ability to make strong inferences from results. Some types of designs, such as the simple blocked design, often yield high experimental power, but they provide imprecise information about the particular psychological processes that activate a brain region. Event-related designs allow brain activation to be related more precisely to particular cognitive processes engaged in particular types of trials, but they are often less powerful and require additional assumptions about the underlying hemodynamics of the brain. Researchers may also choose to focus intensively on testing one comparison of interest, maximizing power to detect an

effect, or they may test multiple conditions in order to draw inferences about the generality of a brain region's involvement in a class of similar psychological processes. We describe some types of designs and the uses to which they are best suited below.

Blocked designs

Because long intervals of time (30 seconds or more) are required to collect sufficient data in a PET experiment to yield a good image, the standard experimental design used in PET-activation studies is the blocked design. A blocked design is one in which different conditions in the experiment are presented as separate blocks of trials, with each block representing one scan during an experiment. To image a briefly occurring psychological process (e.g., the activation due to attention switching) using a blocked design, one might repeat the process of interest during an experimental block (A) and not during a conrol block (B). The A – B comparison is a simple *contrast* across experimental conditions; contrasts of various types are used in all kinds of experimental designs. Given the temporal limitation of this technique, PET is not well suited to examining the fine time course of brain activity that may change within seconds or fractions of a second.

The blocked structure of PET designs (and blocked fMRI designs) imposes limitations on the interpretability of results. Activations related to slowly changing factors such as task-set or general motivation are captured well by blocked designs. However, if one wishes to image the neural responses to individual stimuli, blocked designs are not well suited to that goal. In addition, the A – B contrast does not permit the researchers to infer if a region is activated in A but not B, deactivated in B but not A, or shows a combination of both effects. The use of multiple controls and comparison conditions can ameliorate this problem to some degree.

One advantage to a blocked design is that it offers more statistical power to detect a change. Under ideal conditions, blocked designs can be over 6 times as efficient as randomized event-related designs (Wager & Nichols, 2003). Generally, theory and simulations designed to assess experimental power in fMRI designs point to a 16-18 s task / 16-18 s control alternating-block design as optimally statistically powerful (Liu, 2004; Skudlarski, Constable, & Gore, 1999; Wager & Nichols, 2003). However, it is worth noting that this is not always true; the relative power of a blocked design depends on whether a) the target mental process is engaged relatively continuously in A and not at all in B, and b) imposing a block structure changes the nature of the task. On this latter point, it is easy to imagine that blocking some variable may fundamentally alter the strategy that subjects apply to the task.

Event-related fMRI

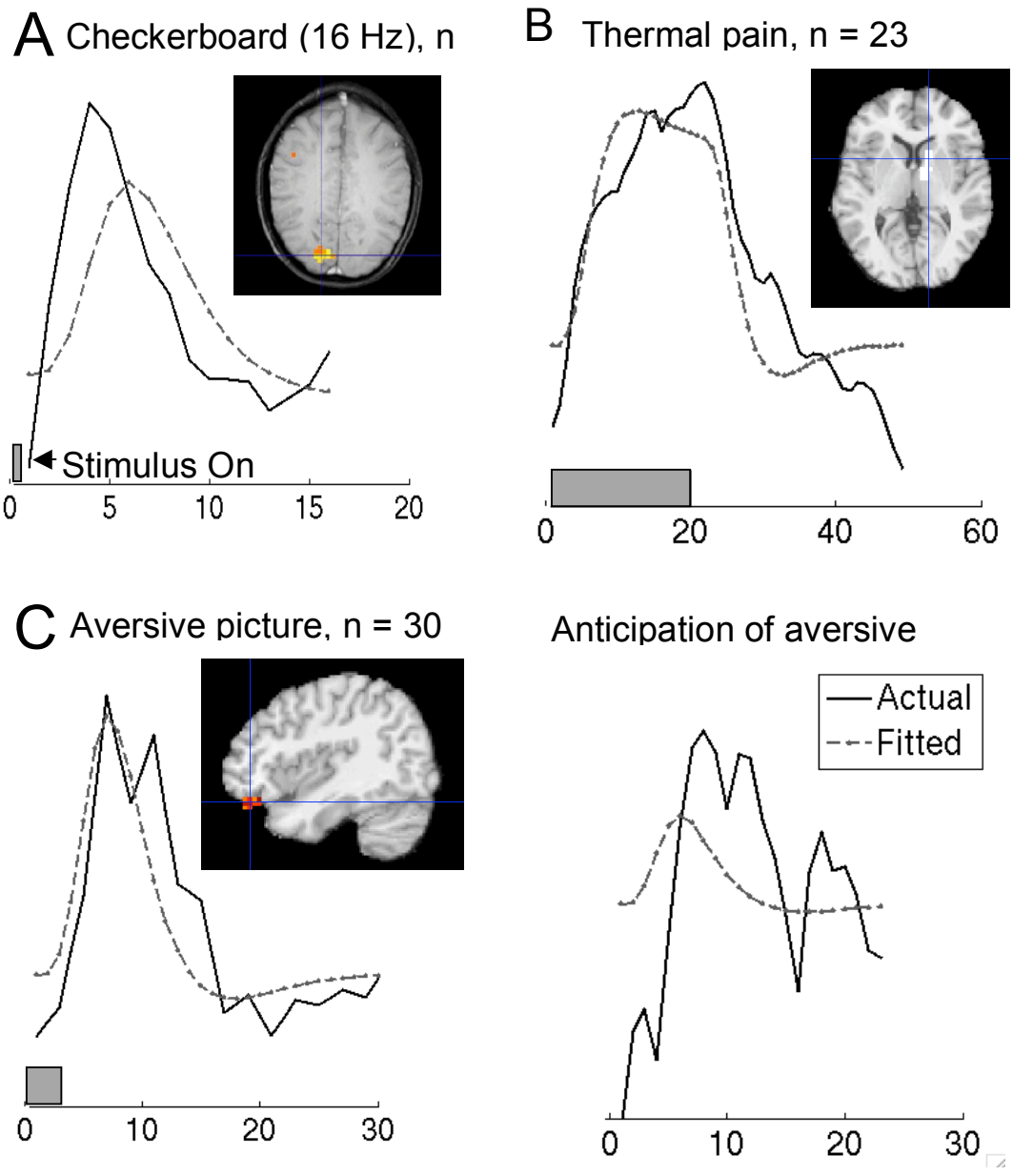
To take advantage of the rapid data-acquisition capabilities of fMRI, event-related fMRI designs and analyses have been developed. These designs, when employed judiciously, allow one to estimate the fMRI response evoked by specific stimuli or cognitive events within a trial (Rosen, Buckner, & Dale, 1998). A sample of the MRI signal in the whole brain can be obtained in 2-3 seconds on average (the "TR", or repetition time of image acquisition), depending on the way the data are acquired and depending on the required spatial resolution of the voxels that are imaged. The limiting factor in the temporal resolution of fMRI is generally not the speed of data acquisition, but the speed of the underlying evoked hemodynamic response to a neural event (or the hemodynamic response function, HRF), which begins within a second after neural activity occurs and peaks 5-8 seconds after that neural activity has peaked (Aguirre, Zarahn, & D'Esposito, 1998; Friston, Frith, Turner, & Frackowiak, 1995).

While event-related designs are attractive because of their flexibility and the information they provide about the individual responses, they require more caution in the design and analysis of the experiment. It is quite common to assume an ideal or "canonical" hemodynamic response (HRF) in order to generate linear models for statistical analyses (the fitted curves shown by the dashed lines in Figure 12). This procedure requires a realistic model of the shape and timing of the HRF. Notably, the canonical estimates typically used come from studies of brief visual and motor events. In practice, however, the timing and shape of the HRF are quite variable across the

brain within an individual and across individuals (Aguirre et al., 1998; Schacter, Buckner, Koutstaal, Dale, & Rosen, 1997). Part of the variability among shapes of the HRF is due to the underlying configuration of the vascular bed, which may cause differences in the HRF across brain regions in the same task for purely physiological reasons. Another source of variability in the shape of the HRF is differences in the pattern of evoked neural activity in regions performing different functions related to the same task.

Figure 12A shows an example of an average response in left visual cortex, identified on an individual basis in 10 participants (scanning time was approximately 60 minutes per participant, with one stimulus every 30 s) (Wager, Vazquez, Hernandez, & Noll, 2005). The figure shows that the HRF peaks earlier than the canonical model, and the magnitude of activation is thus underestimated by about 30% in this group of subjects. Blocked designs are less sensitive to the variability of the HRF because they are dependent on the total activation caused by a *train* of stimulus events (though this depends on the density of psychological activity during the block (Price, Veltman, Ashburner, Josephs, & Friston, 1999)!). For example, Figure 12B shows the responses to a 20 s epoch of thermal pain in the right thalamus and caudate, averaged over 23 subjects (Wager, Rilling et al., 2004). The predicted model of this longer

Figure 12



period of stimulation is also based on the canonical HRF. Specifically, the model is generated by convolution of the HRF with a 20-s long step-function representing the stimulation period. The canonical model fits reasonably well in this case.

On the other hand, Figure 12 C shows a problematic case. The orbitofrontal cortex response to a 3 s presentation of an aversive visual stimulus is somewhat prolonged relative to the model, but the estimated magnitude is reasonably accurate (unpublished data). The right panel, on the other hand, shows the response in the same brain region to anticipation of viewing an aversive picture. In this case, the timing and duration of the cognitive activity are unknown—the activity could occur at the onset of the cue, throughout the anticipation epoch, or with increases proportional to the proximity of the picture. The model fits poorly in this case, illustrating the importance of knowing the onset and duration of the stimulation for linear model analyses.

In a single-trial event-related design, events are spaced far apart in time (every 20 – 30 s is considered sufficient). FMRI signal can be observed on single trials if the eliciting stimulus is very strong (Duann et al., 2002), permitting the possibility of fitting models at the level of an individual trial (Rissman, Gazzaley, & D'Esposito, 2004). This promising technique enables the testing of relationships between brain activity and trial-level performance measures such as reaction time and emotion ratings for particular stimuli (Phan et al., 2004).

Early studies frequently employed selective averaging of activity following onsets of a particular type (Aguirre, Singh, & D'Esposito, 1999; Buckner et al., 1998)(Menon, Luknowsky, & Gati, 1998)). However, even brief events (e.g., a 125 ms visual checkerboard display) have been shown to affect fMRI signal more than 30 s later (Wager, Vazquez et al., 2005). Because the selective averaging procedure does not take stimulus history into account, it must be used with caution when responses to different events may overlap in time. Because of this, the majority of analyses, including those that estimate the shapes of HRFs, are currently done within the GLM framework (described below).

Reports that the fMRI BOLD response is linear with respect to stimulus history (Boynton, Engel, Glover, & Heeger, 1996) encouraged the use of more rapidly-paced trials (Zarahn, Aguirre, & D'Esposito, 1997), spaced less than 1 s apart in extreme cases (Burock, Buckner, Woldorff, Rosen, & Dale, 1998; Dale & Buckner, 1997). Linearity in this context means that the magnitude and shape of the HRF does not change depending on the preceding stimuli. Studies have found that nonlinear effects in rapid sequences (1 or 2 s) can be quite large (Vazquez & Noll, 1998)(Birn, Saad, & Bandettini, 2001; Friston, Mechelli, Turner, & Price, 2000; Wager, Vazquez et al., 2005), but that responses are roughly linear if events are spaced 4 – 5 s apart (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000).

To get an intuition about how rapid designs allow one to discriminate the effects of different conditions, consider this: With a randomized and jittered design, sometimes several trials of a single type will occur in a row, and because the hemodynamic response to closely spaced events sums in a roughly linear fashion, the expected response to that trial-type will build to a high peak. Introducing longer delays between some trials and shorter ones between others allows peaks and valleys in activation to develop that are specific to particular experimental conditions. Thus, it is critical either to systematically intermix events of different types in a rapid event-related design or to vary ('jitter') the inter-stimulus interval (ISI) between trials.

Suppose, for example, you have a rapid sequence with two types of trials—say, attention switch trials (S) and no-switch trials (N) as in task switching experiment described above (Fig. 10). Randomly intermixing the trials with an ISI of 2 s will allow you to compare responses to S with those to N; that is, you will be able to estimate the difference S – N. However, you will not be able to tell if S and N 'activate' or 'deactivate' relative to some other baseline. If you vary the inter-stimulus intervals randomly between 2 and 16 s, you'll be able to compare S – N (albeit with less power because there will be fewer trials), but you'll also be able to test whether S and N show positive or negative activation responses. This ability comes from the inclusion of inter-trial rest intervals against which to compare S and N, and the relatively unique signature of

predicted responses to both S and N afforded by the random variation in ISIs.

The advantages of rapid pacing— including trial pacing comparable to other experiments, reduced boredom, and sometimes increased statistical efficiency—must be weighed against potential problems with nonlinearity, multicolinearity, and model mis-fitting when very rapid designs are used. A current popular choice is to use 'jittered' designs with inter-stimulus intervals of at least 4 s, with exponentially decreasing frequencies of delays up to 16 s.

Techniques for contrasting experimental conditions

Thus far, we have alluded to a simple kind of contrast between two conditions, the subtraction of a control condition (B) from an experimental one (A). Such contrasts are critical because any task, performed alone, produces activation in huge portions of the brain. To associate changes in brain activation with a particular cognitive process requires that we isolate changes related to that process from changes related to other processes.

The simple contrast discussed above was the first method used to make inferences about psychological processes from neuroimaging data (Petersen, Fox, Posner, Mintun, & Raichle, 1988; Posner, Petersen, Fox, & Raichle, 1988). It is called the 'subtraction method,' the logic of which is this: If one tests two experimental conditions that differ by only one process, then a subtraction of the activations of one condition from those of the other should reveal the brain regions associated with the target process. This subtraction is accomplished one voxel at a time throughout brain regions of interest. Together, the results of the voxel-wise subtractions yield a three-dimensional matrix of the difference in activation between the two conditions throughout the scanned regions of the brain, or *contrast images*. T-tests can be performed for each voxel to discover where in the brain the difference is reliable. The resulting parametric map of the t-values for each voxel shows the reliability of the difference between the two conditions throughout the brain. Figures of 'activation maps' in publications, such as those shown in Fig. 10, generally show images of voxels whose t-values or comparable statistics (z or F) exceed a statistical threshold for significance.

Subtraction logic rests on a critical assumption, what has been called the assumption of "pure insertion" (Sternberg, 1969). According to this assumption, changing one process does not change the way other processes are performed. Thus, by this assumption, the process of interest may be 'purely' inserted into the sequence of operations without altering any other processes. Although violations of subtraction logic have been demonstrated experimentally (Zarahn et al., 1997), the logic is still widely used because it greatly simplifies the inference-making process. However, if the two tasks studied differ in more than one overt (e.g., visual stimulation) or covert process (e.g., attention, error checking, subvocalization, attention shifts), the results will be ambiguous. With only two tasks or conditions, this is very often the case.

One way to constrain the interpretation of brain activity and strengthen the credibility of subtraction logic is to incrementally vary a parameter of interest across several levels—essentially performing multiple subtractions on a single variable. An example is a study of the Tower of London task (Dagher, Owen, Boecker, & Brooks, 1999), which requires subjects to make a sequence of moves to transfer a stack of colored balls from one post to another in the correct order. The experimenters varied the number of moves incrementally from 1 to 6. Their results showed linear increases in activity in dorsolateral prefrontal cortex across all 6 conditions, suggesting that this area served the planning operations critical for good performance. The contrast tested in this condition is a *linear contrast*. We provide a mathematical explanation of how exactly this and other contrasts are performed in the section on linear modeling below. Other contrasts, such as quadratic or monotonic contrasts, can also be specified.

Another extension of subtraction logic is the factorial design. The study of task switching presented in the introduction to this chapter serves as an example. Consider the comparison in this study of two types of switching, each varied independently: switching among objects and switching among attributes of objects. This design is a simple 2 X 2 factorial, with 2 types of trials (switch vs. no switch) crossed with 2 types of judgments (object/attribute). This

design permits the testing of three contrasts: a) a main effect of switch vs. no switch; b) a main effect of task type; and c) the interaction between the two, which tests whether the switch vs. non-switch difference is larger for one task-type than the other. Factors whose measurements and statistical comparisons are made within subjects, as are those described above, are *within-subjects* factors, and those whose levels contain data from different individuals (e.g., depressed patients vs. controls) are *between-subjects* factors. Within-subjects factors generally offer substantially more power and have fewer confounding issues (e.g., differences in brain structure and HRF shapes) than between-subjects factors.

Factorial designs allow one to investigate the effects of several variables on brain activations. They also permit a more detailed characterization of the range of processes that activate a particular brain region – e.g., attention switching in general, or switching more for one task-type than the other. Factorial designs also permit one to discover double dissociations of functions within a single experiment. In our example (Fig. 10), a factorial design was required in order to infer that a manipulation (e.g. object-switching) affected dorsolateral prefrontal cortex, but a second manipulation (e.g. attribute switching) did not.

Factorial designs also provide some ability to test for brain regions activated in common by a set of related tasks or uniquely by only certain tasks (Fan, Flombaum, McCandliss, Thomas, & Posner, 2003; Wager, Jonides et al., 2005; Wager, Sylvester et al., 2005). An important note in this regard is that inferring that an area responds, say, to 'task switching in general' requires more than just a significant main effect of switching, but rather a *conjunction* across object and attribute switching. That is, each effect must be independently significant to conclude that both effects were present in a region (T. Nichols, Brett, Andersson, Wager, & Poline, 2005).

Also of great interest to psychologists and neuroscientists is the *specificity* of regional activation to one particular psychological process. An activated region or pattern of activations specific to a process is activated by *only* that process. If such specificity exists, then that pattern may serve as a marker for the psychological process, and reverse inference becomes possible. In general, if such inferences can be made at all, it seems practical to compare across a great many studies using a meta-analysis (Wager, Jonides et al., 2004)—because establishing a marker requires determining which tasks do not activate the region. However, a problem with meta-analyses is their poor spatial resolution compared with individual studies.

One productive line of research attempting to address specificity within individual studies has been pursued by Kanwisher and colleagues in the study of face recognition (Kanwisher, McDermott, & Chun, 1997). In this study and many subsequent ones, they identified an area in the fusiform gyrus that responded to pictures of faces and drawings of faces, but not to houses, scrambled faces, partial faces, facial features, animal faces, and other control stimuli. By presenting a large number of control stimuli of various types, Kanwisher et al. were able to infer that the brain area they studied, which they called the Fusiform Face Area (FFA), was specific to the perception of faces. They tested both for the types stimuli that elicited an FFA response and for those that did *not* elicit a response. In addition, because this region lies in slightly different locations across different participants, they performed separate "localizer" scans to locate the functionally face-selective region within individual participants.

The general linear model (GLM) in neuroimaging

The techniques described above can be understood mathematically in terms of the GLM framework. GLM is a linear analysis method that subsumes many basic analysis techniques, including t-tests, ANOVA, and multiple regression. In neuroimaging designs, analyses of both blocked and event-related designs are most frequently carried out in the GLM framework, though with some extensions that we describe below. The GLM can be used to estimate whether the brain responds to a single type of event, to compare different types of events, to assess correlations between brain activity and behavioral performance or other psychological variables, and for other tests.

The GLM is appropriate when multiple predictor variables—which together constitute a

simplified *model* of the sources of variability in a set of data—are used to explain variability in a single, continuously distributed outcome variable. In a typical neuroimaging experiment, the predictors are related to psychological events, and the outcome variable is signal in a brain voxel or region of interest. In analysis with the 'massively univariate' GLM approach commonly used in brain imaging, one performs a separate regression analysis at every voxel in the brain.

In a single-subject fMRI analysis, the GLM assumes that the observed signal time series in a voxel is the sum of the activity evoked by a number of independent processes. Some of these processes are of interest, but others are nuisance contributions to the overall signal (e.g., apparent activation due to movement of the head). Activity evoked by each process is modeled with by constructing a predictor, or estimate of the predicted brain response, for each process. The amplitudes (weights) of the predictors, which reflect the magnitudes of the evoked activity for each process, are unknown. The GLM estimates these magnitudes and determines whether they are significantly different from zero at each voxel in the brain.

By convention, predictors are arranged in columns. The *design matrix*, X, describes the model. It contains a row for each of *n* observations collected (subjects or samples) and a column for each of *k* predictors. The data, *y*, are modeled as a linear combination of the predictors (i.e., a weighted sum) plus error, ε. The GLM framework is described by the equation:

$$y = X\beta + \varepsilon \quad \textbf{(5)}$$

where β is a *k* x 1 vector containing the predictor weights (also called regression slopes or parameter estimates). The equation is in matrix notation. Thus, *X* is an *n* x *k* model matrix, *y* is an *n* x 1 vector containing the observed data, and ε is an *n* x 1 vector of unexplained error values. Error values are assumed to be independent and to follow a normal distribution with mean 0 and standard deviation σ. The β weights in a neuroimaging experiment correspond to the *estimated magnitude of activation* for each psychological condition described in the columns of X.

The model is fit to the data by finding the weights (β) that minimize the squared distance between the vector of fitted values, $X\beta$, and the data. One of the advantages of the GLM is that there is an algebraic solution for the βs that minimizes the squared error:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \textbf{(6)}$$

where $\hat{\beta}$ are the estimates of the true regression slopes. In algebraic terms, the GLM projects data (y) in an n-dimensional space (n independent data observations) onto a k-dimensional model subspace (k predictors). The matrix product $(X^T X)^{-1} X^T$ is the matrix of the orthogonal projection onto the reduced-dimensional space of *X*.

Inference is generally conducted by comparing the $\hat{\beta}$s with their standard errors and using classical inferential procedures to estimate the likelihood of the estimates under the null hypothesis. The standard errors of the estimates are the diagonal elements of the matrix:

$$se(\hat{\beta}) = (X^T X)^{-1} \hat{\sigma} \quad \textbf{(7)}$$

The ratio of betas to their standard errors follows a t-distribution with n – k degrees of freedom. Notably, the error term is composed of two separate terms from different sources. σ is the error variance, and $(X^T X)^{-1}$ depends on the design matrix itself. If the matrix contains high values, the estimates will be less precise, and statistical power will decrease. If X is scaled appropriately, $X^T X$ is the variance-covariance matrix of X. Low values on the diagonal of $X^T X$ (low predictor variability) and high off-diagonal elements (high colinearity) will both cause $(X^T X)^{-1}$, and the variability of the estimates, to grow.

Contrasts

Contrasts across conditions can be easily handled within the GLM framework. Mathematically, a contrast is a linear combination of predictors. The contrast (e.g., A – B in a simple comparison, or A + B – C – D for a main effect in a 2 x 2 factorial design) is coded as a *k* x 1 vector of contrast weights, which we denote with the letter *c*. For example, the contrast weights for a simple subtraction are $c = [1\ -1]^T$, where T indicates the transpose operator. A single contrast for a linear effect across four conditions might be $c = [-3\ -1\ 1\ 3]^T$. A set of contrasts can be simultaneously tested by concatenating the contrasts into a matrix. Thus, the main effects and interaction contrasts in a 2 x 2 factorial design can be specified with the following matrix:

$$C = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix};$$

For the contrast to be orthogonal to the intercept, contrast weights must sum to zero. If the weights do not sum to zero, then the contrast values partially reflect overall scanner signal intensity. It is worth going into this level of detail here because contrast weights in statistics packages used with imaging data are often specified by the analyst, rather than being specified automatically as they might be in standard statistical programs such as SPSS or SAS. The true contrast values $C^T\beta$ can be estimated using $C^T\hat{\beta}$, where $\hat{\beta}$ is obtained using Eq. (6). The standard errors of each contrast are the diagonals of:

$$se(C^T\hat{\beta}) = C^T(X^TX)^{-1}C\hat{\sigma} \quad \textbf{(8)}$$

Most imaging statistics packages write a series of images to disk containing the betas for each condition throughout the brain, and another set of contrast images containing the values of $C^T\hat{\beta}$ throughout the brain. As the latter images contain estimates of activation differences across conditions, these are typically used in a group analysis. A third set of images contain t-statistics, or the ratio of contrast estimates to their standard errors.

Assumptions

The model-fitting procedure assumes that the effects due to each of the predictors add linearly and do not change over time (i.e., a linear, time-invariant system). The inferential process assumes that the observations are independent, that they all come from the same distribution, and that the residuals are distributed normally and with equal variance across the range of predicted values. All of these assumptions are violated to a degree in at least some brain regions in a typical imaging experiment, which has prompted the development of a number of important extensions. Violations of the assumptions are not merely a theoretical nuisance. They can make the difference between a valid finding and a false positive result, or between finding meaningful activations in the brain and wasting substantial time and money.

Diagnostic tools have been developed for exploring the data, looking for artifacts, and checking a number of assumptions about the data and model (Luo & Nichols, 2003), and like many tools developed by members of the neuroimaging community, they are freely available on the internet. The quantity of data—e.g., 100,000 separate regressions on 1000 data points per subject x 20 subjects—and the software and data structures that support its analysis makes it very difficult to examine assumptions and check the data, which makes such diagnostic tools all the more important.

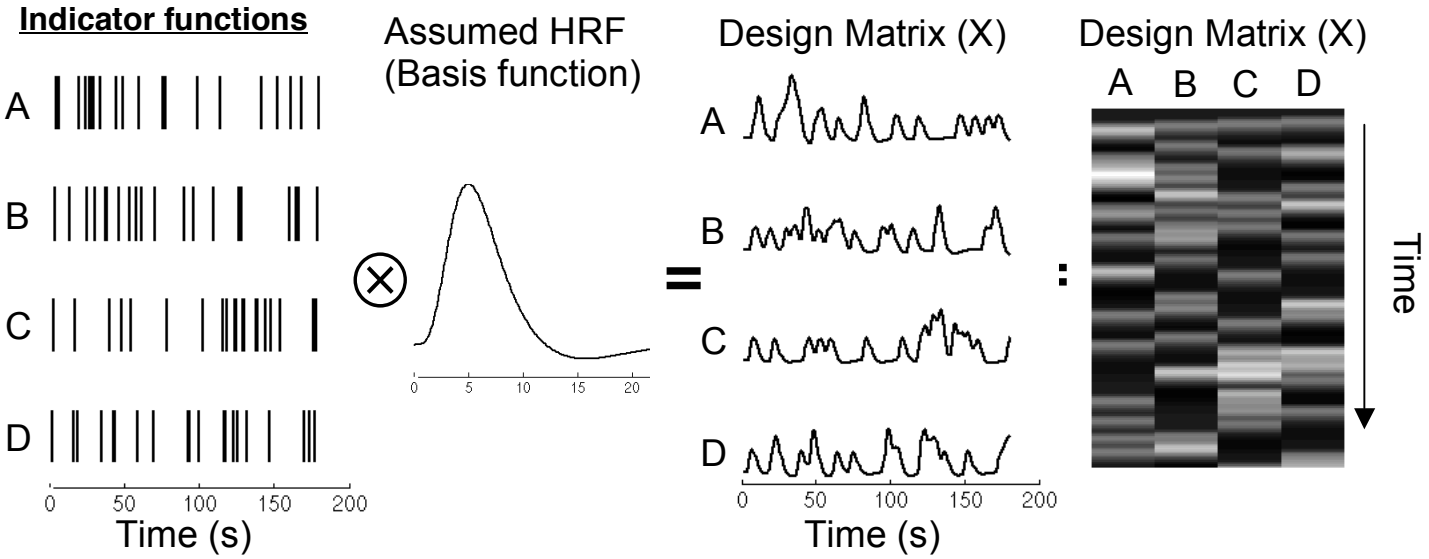
Another active area of research concerns strategies for dealing with some known violations of assumptions, described below. Violations of independence can be handled in a limited way using generalized least squares. Violations of equality and normality can be dealt with by using nonparametric permutation tests to make statistical inferences (T. E. Nichols & Holmes, 2002), or, if they result from the presence of outliers, by robust regression techniques

(Wager, Keller, Lacey, & Jonides, 2005). Free implementations of each of these extensions are available, and we return to a description of them after providing additional background on application of the GLM to imaging experiments.

GLM model-building in fMRI

Perhaps the most challenging task in linear regression analysis is the creation of a *realistic* model of the signal. This model constitutes the matrix X in the general linear model

Figure 13



equation. In a neuroimaging study, researchers typically build a model of the predicted brain response to each psychological event-type or condition. A stimulus or psychological *event* that elicits a brief burst of neural activity in an area typically produces a prolonged HRF that peaks 5-6 s later (but may vary across brain regions, as described above!) PET images integrate across many such events, and each brain image (an image containing one data observation at every voxel) reflects overall activity in a particular condition. Thus, much of the remainder of this section does not apply to PET analyses.

Accounting for the delayed hemodynamic response

A popular method of forming a prediction about BOLD activity is to assume that the response to a brief event will follow a canonical shape, such as that shown in Figure 12A and the center panel of Figure 13, for every event-type and every voxel in the brain. To build the model, researchers start with an 'indicator' vector representing the neuronal activity for each condition sampled at the resolution of the fMRI experiment. This vector has zero value except during activation periods, when the signal is assigned a unit value. To form the predicted response in a condition, the indicator vector for that condition is convolved with the assumed HRF, and the result forms a column of the design matrix.

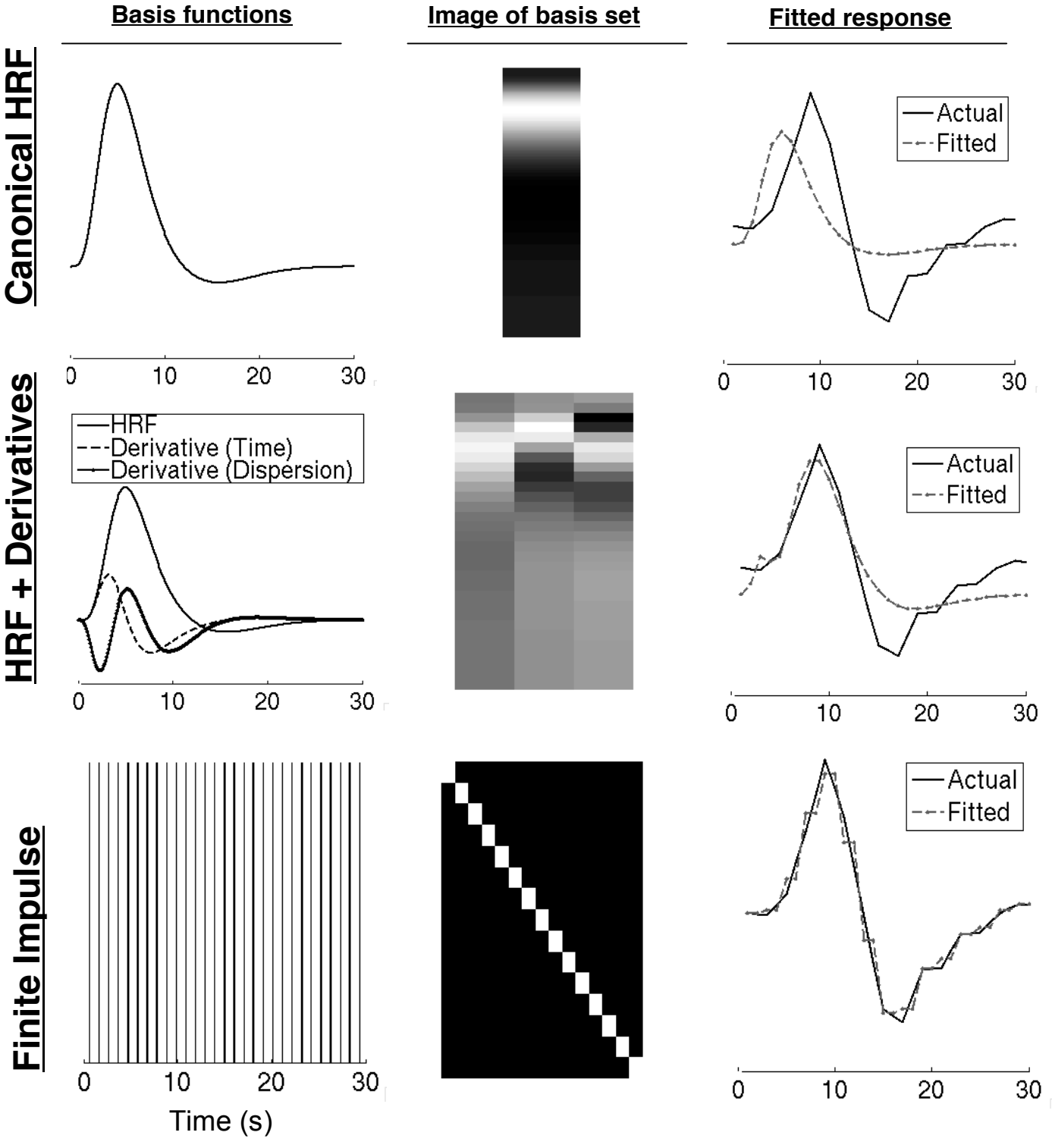
The process, shown in Figure 13 for an event-related design with four trial types (A – D), is similar for both blocked and event-related fMRI designs. The left panel shows the indicators, with rows of the plot corresponding to the four conditions. Each of the indicator vectors is convolved with the canonical HRF, shown in the center panel, to yield the predictors shown in the rows of the next panel. The image of the design matrix, a commonly used presentation format in imaging experiments, is shown in the rightmost panel.

If the assumed HRF does not fit, there is at best a drop in power to detect a response. At worst, model mis-specification can produce false positive results. Say, for example, the HRF peaks at the expected time in condition B, but later in A. Since the shape is fixed, the amplitude

of the fit for B will be greater than A. Without some additional diagnostic tests, one might falsely infer that B activates the brain region more than A.

Comparing groups of individuals (e.g., older versus younger adults, or patients and normal controls) can also be especially problematic. If one finds differences between Task A and Task B in evoked response magnitude, are those differences caused by differences in neural activity, or by differences in how well subjects' responses fit the canonical, assumed shape? Elderly subjects, for example, have reduced and more variable shapes of their HRF's compared to younger subjects (D'Esposito, Zarahn, Aguirre, & Rypma, 1999).

Figure 14



One approach that has been used to avoid this problem is the measurement of hemodynamic responses in visual and motor cortex for each individual subject (Aguirre et al., 1998). This approach may work for brain regions whose HRF shapes match those in sensorimotor cortex. An alternative approach is to use a more flexible model of the HRF, which we describe next.

Basis sets

In the previous discussion, responses to each psychological event-type or condition are modeled by only one linear regressor, which allows one to estimate only the fitted amplitude of the response. If more than one regressor is used to model each event-type, then the regressors can model different components of the response. The same onsets are convolved with different canonical functions that together can model a range of different HRF shapes. These canonical functions are called basis functions, and the group of basis functions chosen to model the response is called a basis set. Fitting a basis set at each voxel means that the fitted shape of the HRF is allowed to vary across brain regions.

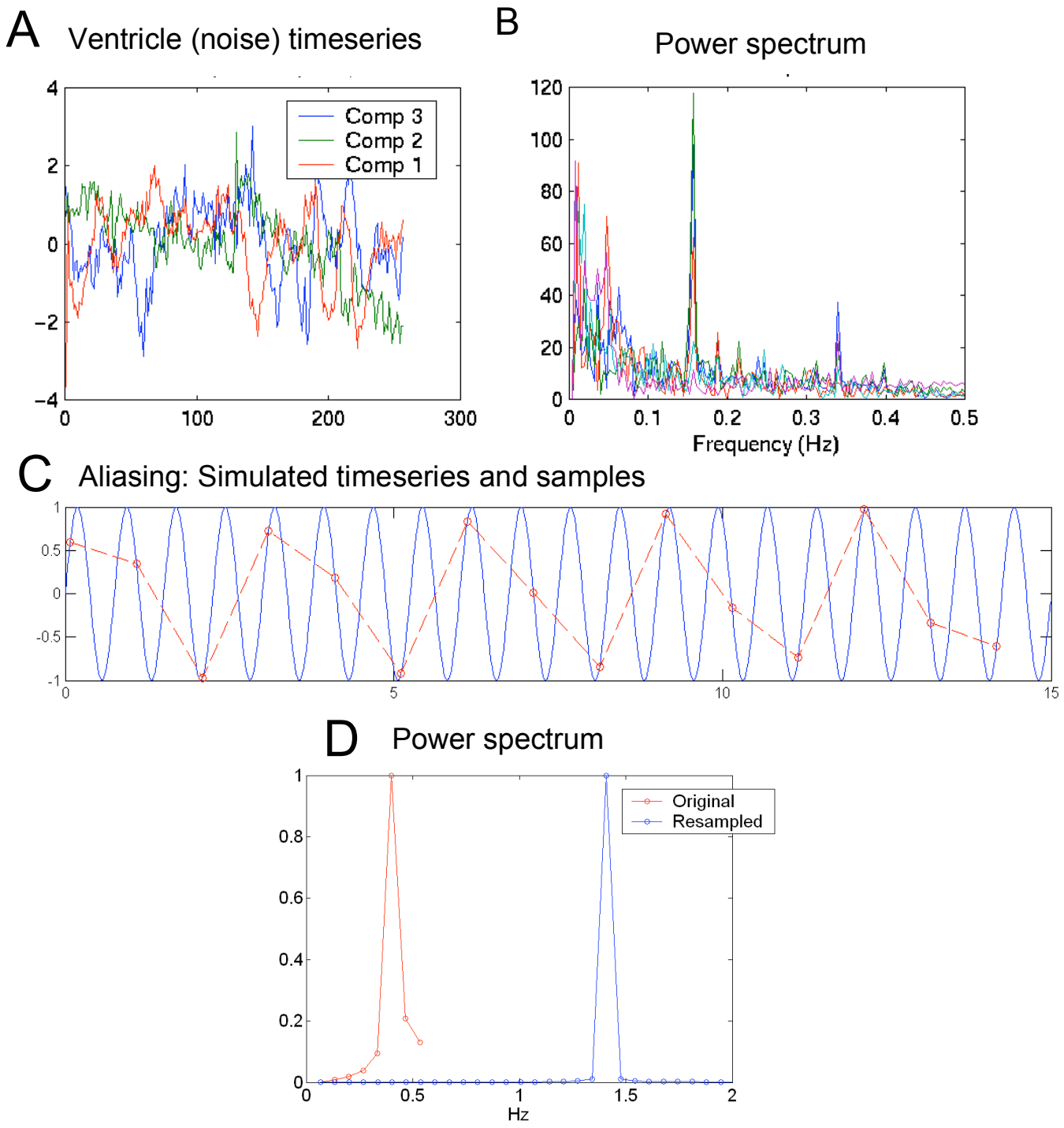
Basis sets vary in the number of parameters estimated per event-type and in the shapes of the basis functions. A flexible basis set will be able to model more different HRF shapes, but generally at the cost of statistical power (more parameters means more variability in parameter estimates). A popular choice is to use a canonical HRF and its derivatives with respect to time and dispersion (we use TD to denote this hereafter (Friston, Glaser et al., 2002; Friston, Josephs, Rees, & Turner, 1998). Other choices include basis sets composed of principal components (Aguirre et al., 1998; Woolrich, Behrens, & Smith, 2004), cosine functions (Zarahn, 2002), radial basis functions (Riera et al., 2004), and spectral basis sets (Liao et al., 2002).

One of the most flexible models, a finite impulse response (FIR) basis set, contains one free parameter for every time-point following stimulation in every cognitive event-type that is modeled (Glover, 1999; Goutte, Nielsen, & Hansen, 2000; Ollinger, Shulman, & Corbetta, 2001). Using such a model makes minimal assumptions about the shape of the HRF—in fact, the set of parameter estimates (betas) from this model constitute an estimate of what the shape of the HRF looks like when sampled at the frequency of data collection (the TR).

Figure 14 shows fits to some empirical data from our laboratory using three popular basis sets with different degrees of flexibility. The "actual" HRF in these plots (solid lines in the right panels) comes from a region of the right thalamus that responds to viewing aversive emotional pictures. They are estimates using an FIR model with 1 s time bins, so they are the most assumption-free estimates we can obtain of the true response in this region. The SPM2 canonial HRF, composed of two summed gamma functions, is shown in the top left panel. The image of the basis set is shown in the center panel, and the fit of the basis set to the data is shown in the rightmost panel. As the figure shows, the fit leaves much to be desired, and the height of the response is substantially underestimated, because the actual HRF peaks later and shows a much larger undershoot than the model. The second row of panels shows a plot of the basis set of the SPM2 canonical HRF with its derivatives with respect to time and dispersion (left), the image of the basis set (center), and the fitted response (right). The fit is substantially better, and would probably be adequate in this case to capture the positive-going lobe of the evoked activation. An FIR model is a linear model too, and the bottom panels show the same results for an FIR model with one estimate every 2 s.

Whatever basis set is used, statistical comparisons across event types must be made. The advantage of assuming a canonical shape is that because beta values represent response amplitudes for different conditions, the differences between betas represent differences in activation across psychological tasks or states. It is thus relatively straightforward to assess contrasts in a subtraction, parametric, or factorial design.

Figure 15



With more flexible basis sets, the parameters for each event-type combine—sometimes in complex ways—to model the HRF, and the betas for one condition cannot simply be subtracted from another. An omnibus F-test in repeated-measures ANOVA can be used to look for parameter x condition interactions (implemented in some neuroimaging statistics packages, such as SPM2), if care is taken to account for the correlations among parameter estimates. However, this approach carries a cost in power and a problem in interpretability: A significant test statistic implies that the conditions differ in some way, but does not specify in what way (amplitude,

latency, shape) they are different. For this reason, one-parameter canonical models remain popular despite their problems—but more sophisticated approaches, such as measuring the height and peak/onset delay from estimated HRFs, have been developed and may be implemented in popular packages soon.

Physiological noise and covariates of no interest

In both PET and fMRI designs, additional predictors are typically added to account for known sources of noise in the data. These are covariates of no interest, and they are included to reduce noise and to prevent signal changes related to head movement and physiological (e.g., respiration) artifacts from influencing the contrast estimates. In PET, a common covariate is the global (whole-brain) mean signal value for each subject, included to control for differences in amount of radioactive tracer in circulation.

In fMRI, the signal can drift slowly over time, as shown in Figure 15A. These signals were extracted from the ventricles during an fMRI scan, where there is no tissue, and thus no task-related activation. Though there is noise at all frequencies, the noise has greater amplitude at low frequencies, as shown in the spectral plot of the Fourier transform in Figure 15B. The Fourier transform of fMRI noise shows an inverse relationship with the noise frequency, and can be approximated by a $1/f$ (1/frequency) function. Also apparent in these data are spikes of high-power at about 0.15 Hz and 0.34 Hz. These frequencies may correspond approximately to average heart rate and respiration rate aliased back into the task frequencies. Aliasing occurs when a true signal (blue in Figure 15C) occurs at more than twice the sampling frequency (red circles in Figure 15C), and the apparent frequency (e.g., of the red curve) is 'reflected' back to a lower value. The power spectra of the original and aliased signals are shown in Figure 15D. In fact, much of the autocorrelated noise in fMRI may come from aliased physiological artifacts (Lund, Madsen, Sidaros, Luo, & Nichols, 2005).

Because of slow drift, it is advantageous to filter out low-frequency noise with a *high-pass filter*. This filtering is often performed in the GLM, by adding covariates of no interest (e.g. low-frequency cosines). Such filtering precludes using designs that vary the task condition at low frequencies. To the degree that the task frequencies overlap with the filtered frequencies, the filter will remove the signal of interest from the data! A good practice is to set the lowest period of the cosine functions modeling the drift to twice that of the experimental period. Other types of commonly used covariates of no interest include the estimated movement parameters and their derivatives/squared values, as well as signals from pulse and respiration sensors.

Extensions of the GLM

Autocorrelation and Generalized Least Squares

One issue with the use of the GLM to analyze imaging data is that it assumes that the observations (images) are all independent of one another. In an fMRI time series, the data are correlated across time and space. This means that the activity in a voxel at one time can be partially determined from the preceding activity, and that there are fewer error degrees of freedom than used in the GLM calculations. Such temporal autocorrelation causes apparent p-values from the GLM to be much smaller than they actually are. As a result, inferences about individual subjects cannot be made without correcting for autocorrelation. Group analyses will not be biased by the existence of autocorrelation within individuals, but they will be less powerful than they would be if an appropriate model were used, because the subject-level parameter estimates will be more variable.

There are two basic ways to handle autocorrelation. One way is to estimate the autocorrelation and reduce the degrees of freedom appropriately, e.g., via the Sattherwaite correction (Neter, 1996). A popular approach using the SPM99 toolbox was to apply temporal smoothing to the data (a low-pass filter, which *introduces* autocorrelation), and then reduce the degrees of freedom based on the applied smoothing. However, a more popular approach recently has been to estimate the autocorrelation and remove it from the data during model fitting. This is

called 'prewhitening' because it is an attempt to create 'white noise' without temporal structure, and it is statistically efficient (high power) if the autocorrelation is modeled correctly, but it can introduce false positives under certain conditions if the autocorrelation estimates are in error (Friston, Josephs et al., 2000).

Prewhitening works by pre-multiplying both sides of the general linear model equation (Eq. 5) by the square root of a filtering matrix W, to create a new design matrix $W^{1/2}X$ and whitened data $W^{1/2}y$. W can stand for 'whitening matrix,' or, as we will see later, more generally for 'weighting matrix.' Just as a properly designed matrix can apply a high-pass or low-pass (smoothing) filter, here the matrix can be used to remove autocorrelation. One way to understand this is to notice that $W^{1/2}X$ is a new linear combination, or weighting, of the predicted values (rows) of X based on the columns of $W^{1/2}$. If the initial predicted responses at each time (in a column of X) are themselves a weighted combination of the current stimulus and past history, then W is designed so that the weights (values) in its rows are the *inverse* of the estimated weights that produced the correlation. Thus, applying $W^{1/2}$ weights the observations so that the autocorrelation is removed. The estimates of the activation magnitudes in this case become:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \quad \textbf{(9)}$$

This equation describes activation parameter estimation in the generalized least squares (GLS) framework. Several other applications of GLS, including to hierarchical modeling, are described below. The crux of the issue for dealing with autocorrelation is designing the matrix W so that the autocorrelation is removed. A popular approach in SPM2 is to use an AR1 model, which assumes each observation is a weighted combination of the current level of activity plus carryover from the previous time point. This model can account for smooth drift, but not sinusoidal or other oscillating noise structures. Other models allow for influences father back in time (AR2 for two time-points, AR3 for three, and so forth). An approach implemented in VoxBo software is to use an empirically determined autocorrelation function from visual and motor cortex during separate 'localizer' scans to prewhiten the data (Aguirre, Zarahn, & D'Esposito, 1997). One observation to note is that SPM2 assumes that the autocorrelation (and other parameters such as the covariance and spatial smoothness of the data) is the same everywhere in the brain. FSLs approach implements local (region-by-region) autocorrelation estimates and prewhitening. In principal, if the covariance of the data is given by V, then one should choose $W = V^{-1}$ as the whitening matrix.

We described the W matrix as a weighting matrix above because the form of Eq. 9 is not only appropriate for whitening time series data, but it is generally useful whenever cases (observations) should be weighted. Consider the case of the group analysis, in which some subjects' estimates may be much more reliable than others. This may be because the 'good' subjects showed consistent task performance, they had strong or regularly-shaped HRFs, the scanner noise was low on the day they were tested, their time series data were more independent of physiological confounds, or for other reasons. In such a case, the $\hat{\beta}$s will be better estimates of the true βs for those subjects, and it is advantageous to weight those cases most highly in the group analysis.

The ability to weight cases based on the intra-subject variance is a primary advantage of using a hierarchical model over the two-level 'summary statistic' approach. Once the intra-subject error terms have been estimated, the individuals with the least certain estimates can be down-weighted in the group analysis. In this case, W is a diagonal matrix whose diagonal elements are $1/(\sigma^2_G+\sigma^2)$, or the inverse of the sum of intra- and inter-subject variance estimates. The off-diagonals are zero because we assume subjects are independent of one another (the prewhitening matrix has high off-diagonal entries to the degree that the time series is

autocorrelated). Group $\hat{\beta}_G$s are estimated using weighted least squares as in Eq. 9.

Outliers, artifacts, and robust regression

Statisticians almost universally agree that when performing statistical tests, the data must be examined for outliers and for violations of the statistical assumptions required for inference. Because outliers are far from the group values and the fitting procedure in the GLM minimizes the sum of squared errors, outliers have a disproportionate influence on the group parameter estimates, and they can create violations of the other assumptions. Ideally, the analyst carefully examines the pattern of data and deals with potential outliers on a case-by-case basis.

A major challenge in neuroimaging is that thousands of statistical tests are typically run in parallel, and assumptions are often not checked. Is this a problem? Outliers can cause both null results and false positives, and they may be much more common in imaging data than in many other kinds of data. Simulations performed in our laboratory show that a small proportion of outliers (10% of the sample) can cause a 50% reduction in power over what could be obtained using one of the improved methods described below (Wager, Keller et al., 2005).

Outliers in imaging experiments come from many sources. As mentioned above, some subjects' activation estimates may be more variable than others for a variety of reasons, and some of this variability may be reflected or not in the intra-subject variance. For example, outliers in the individual time series data are regionally specific, very common, and can dramatically influence activation estimates for an individual. These outliers are likely to produce corresponding increases in the variance for that subject. By contrast, errors in the process of normalization, or warping brains into a standard anatomical space, will produce outliers in the group data for a brain region without influencing the intra-subject variance.

Robust regression techniques are a class of statistical tools designed to provide estimates and inferential statistics that are relatively insensitive to the presence of one or more outliers in the data (Huber, 1981; Hubert, 2004; Neter, 1996). They are most appropriate when a large number of regressions are tested and assumptions cannot be evaluated for each individual regression, such as with neuroimaging data. In our work we have compared a number of simple techniques for eliminating or reducing the impact of outliers, focusing mostly on the group analysis level (Wager, Keller et al., 2005). One technique that works well in a variety of situations is iteratively re-weighted least squares (IRLS). It is based on the principle that outliers can be down-weighted rather than dropped altogether, and that outliers can be identified based on how far away they are from the "center of mass" of the data (after accounting for variability explained by the model). The algorithm uses the same weighting scheme used in Eq. 9. As with the hierarchical model, the matrix W is a diagonal matrix containing weights for each subject. The regression line and computation of weights based on residual values is iterated until convergence.

Robust regression is a practical tool for group imaging analysis. Firstly, it does not increase the false positive rate under the variety of conditions and types of outliers encountered in imaging studies; and false positive rates are lower than those of the ordinary GLM under some particularly problematic conditions (Wager, Keller et al., 2005). If there are no outliers in a particular brain region, then there is a small cost in power, because the weighting scheme effectively reduces the error degrees of freedom; but the cost is generally relatively minimal compared to the benefits in regions containing outliers. And while running such an analysis on the computers of the 1990's would have been prohibitive in terms of computation time, a consumer-model computer in 2005 can run IRLS on every voxel in the brain in a typical group imaging study in a matter of several hours.

Group analysis

The analysis described so far has been, for fMRI datasets, an analysis of data from a single subject. However, researchers are often interested in making inferences about a population, not just about a single subject or even a set of individual subjects, which requires a group

analysis.

There are two main approaches to group analysis which differ in the underlying assumptions that they make and the conclusions that may be drawn from the results of the analysis. The *fixed-effects model* does not model variability across subjects, and the *mixed-effects* (often erroneously referred to as 'random-effects') *model* provides for inter-subject variability. Only mixed-effects models allow population inference.

Fixed vs. mixed effects

Fixed-effects models assume the signal strength is identical in all subjects and the only variation present between subjects is due to measurement error. Early approaches in brain imaging collapsed data across multiple observations within a group of subjects into one large GLM analysis. This approach is a fixed-effects model, because the error variance across subjects is assumed to be fixed rather than modeled as a random variable. The hypothesis tests performed are therefore only about the acquired data and cannot be generalized to a wider population. Early PET analyses seems to have taken a wrong turn early on because of the novelty of the data management problem, but fixed effects analyses are rarely used or considered valid now.

The alternative is to treat the subjects as a random effect, meaning that different subjects may actually respond to the experimental task in a different manner. Mixed effects models assume that the signal strength varies across subjects. In these models there are two sources of variation, the first due to measurement error (as in the fixed-effects model) and the second to differences in the individual's response magnitude. Taking this approach, each subject has a random magnitude that is considered to be drawn from a population with a fixed population mean. By testing observed activation values against estimates of the between-subjects error, can provide a traditional test of significance in a population.

Mechanics of mixed effects and hierarchical models

Both PET and fMRI studies nearly always involve collecting more than one image per subject, and testing for the significance of effects in a group of subjects. The full model can be viewed as being hierarchical in nature, with observations (fMRI signal or PET images) nested within subjects, which are in turn nested within a group, and different random variance components are introduced at each level. In fMRI, typically, separate GLM analyses are conducted on the time series data for each subject at each voxel in the brain to estimate the magnitude of activation evoked by the task. This is called a "first level" analysis. These estimates are carried forward and tested for reliability across subjects in a "second level" group analysis.

In the *summary statistics* approach, used in SPM99, VoxBo, FSL, AFNI, and often SPM2, a model is fit for each subject. After the effect of interest is defined, a contrast image is constructed for each subject corresponding to this effect. In the second level, a voxel-wise t-test is performed across all of the contrast images. If we have two populations of interest, then a two-sample t-test can be applied at the second level. This approach is simpler than the full hierarchical model, in that it does not explicitly model error at both levels in the second-level analysis. However, it assumes balanced design matrices (e.g., the same design for each subject) and that the within-subject variance is homogeneous across subjects (i.e., no "noisy" subjects). If these assumptions do not hold, a full hierarchical model, discussed below, is more appropriate.

Mathematically, we can describe the second level analysis in the same way as the first-level GLM discussed earlier. We use X_G to refer to a series of between-subjects predictors. The intercept of this model, which is always included, reflects the mean activity in a population when all other predictors are at zero. If the other predictors are mean-zero, the intercept parameter is the mean activation across subjects. Other predictors in X_G are used to model inter-individual differences (e.g., patients vs. controls, differences in trait anxiety, performance measures, etc.). The vector of the regression slopes (magnitudes) for these effects is β_G, and ε_G is a vector of individual differences not explained by the model. The true activation contrast values for the first-level analysis for each subject are thus a combination of a true population effect plus unexplained

individual differences among subjects. The full two-level model can be written:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \beta &= X_G\beta_G + \varepsilon_G \end{aligned} \quad \textbf{(10)}$$

In particular, note that there are now two sources of noise present in the model, one within subjects and one between subjects.

In the summary statistics approach, we assume that the activation estimates $\hat{\beta}$ are equal to the true values of β. We then run a second GLM, this time with the $\hat{\beta}$s for each subject as the dependent variable and the between-subjects design matrix X_G as predictors. Here the estimate of β_G is dependent on the intermediate estimate $\hat{\beta}$. In the simplest case, a one-sample t-test, $\hat{\beta}$ is an N (subjects) x 1 vector of estimated activation magnitudes for a condition in one voxel (the group data), X_G is an N x 1 column of ones (an intercept), and β_G is a single true group activation parameter.

However, in reality $\hat{\beta}$ is not the same as β, because the estimates $\hat{\beta}$ are influenced by the within-subjects (sampling) error as well. To get the full model, we can replace individual subjects' βs (Eq. 5) by a group model consisting of a population paramter β_G plus individual differences ε_G, so that $\beta = X_G\beta_G + \varepsilon_G$. Thus, the full model that relates the predictors to the data for the group of subjects is:

$$\begin{aligned} y &= XX_G\beta_G + X\varepsilon_G + \varepsilon \\ &= \tilde{X}\beta_G + \gamma \end{aligned} \quad \textbf{(11)}$$

where $\tilde{X} = XX_G$ and $\gamma = X\varepsilon_G + \varepsilon$. The error γ is now composed of two parts—individual difference values ε_G that cannot be explained by the model and error values ε based on the sampling error within subjects. Collecting more data per subject can minimize the within-subjects error variance σ^2, (based on ε) but cannot reduce the between-subjects error variance $\sigma^2{}_G$, based on ε_G. Collecting more subjects can reduce the standard error based on ε_G.

If we solve the equation above for β_G, in the full model we get a solution that depends directly on the data y, and not on the intermediate values $\hat{\beta}$. It can be shown that a single level GLM can be decomposed into an equivalent two-level version if both the $\hat{\beta}$s and their *covariance* are passed down from the first level. Beckmann et al. (Beckmann, Jenkinson, & Smith, 2003) have applied hierarchical modeling to fMRI, and it is implemented in the increasingly popular FILM (fMRIB Improved Linear Model) package as part of the fMRIB Software Library (FSL)(Smith et al., 2004). Hierarchical modeling is also possible in SPM2, though the summary statistics approach is often used in practice (Friston, Stephan, Lund, Morcom, & Kiebel, 2005).

In practice, the relative contribution of the within- and between-subjects error is unknown, and variance components must be estimated. In SPM2 the variance components are estimated using Restricted Maximum Likelihood (ReML) The ReML estimates are calculated using only data from responsive voxels, which are voxels with large F-statistics in a standard pre-analysis of the data, and pooled across those voxels.

Statistical Power and sample size

Statistical power rests on having a large activation response (high contrast values) and a low standard error. In a group study, the standard error comes from two sources: variability across subjects (ε_G) and the variability within each subject (ε). At the group level, power can be increased by increasing the sample size, improving methods for selecting comparable brain

regions across subjects (e.g., normalization, ROI selection), or testing a more homogenous population (at the cost of generalization to other populations).

What sample size is adequate? This depends on the effect size in the group and scanner noise and signal optimization, and it is different for each task and each brain voxel (Zarahn & Slifstein, 2001)(Desmond & Glover, 2002). Using a large group effect size by conventional standards (Cohen's d = 1) and a simple Bonferroni correction for multiple comparisons (described below) over 30,000 voxels, we can get an idea of what sample size is adequate. As

Figure 16

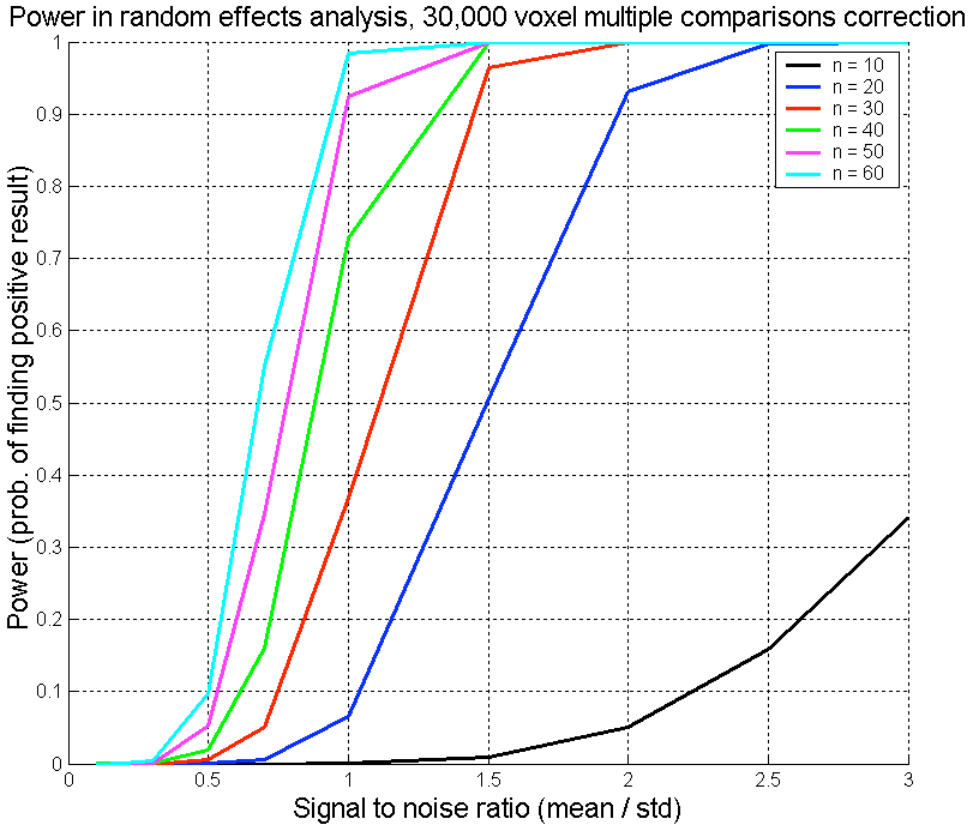
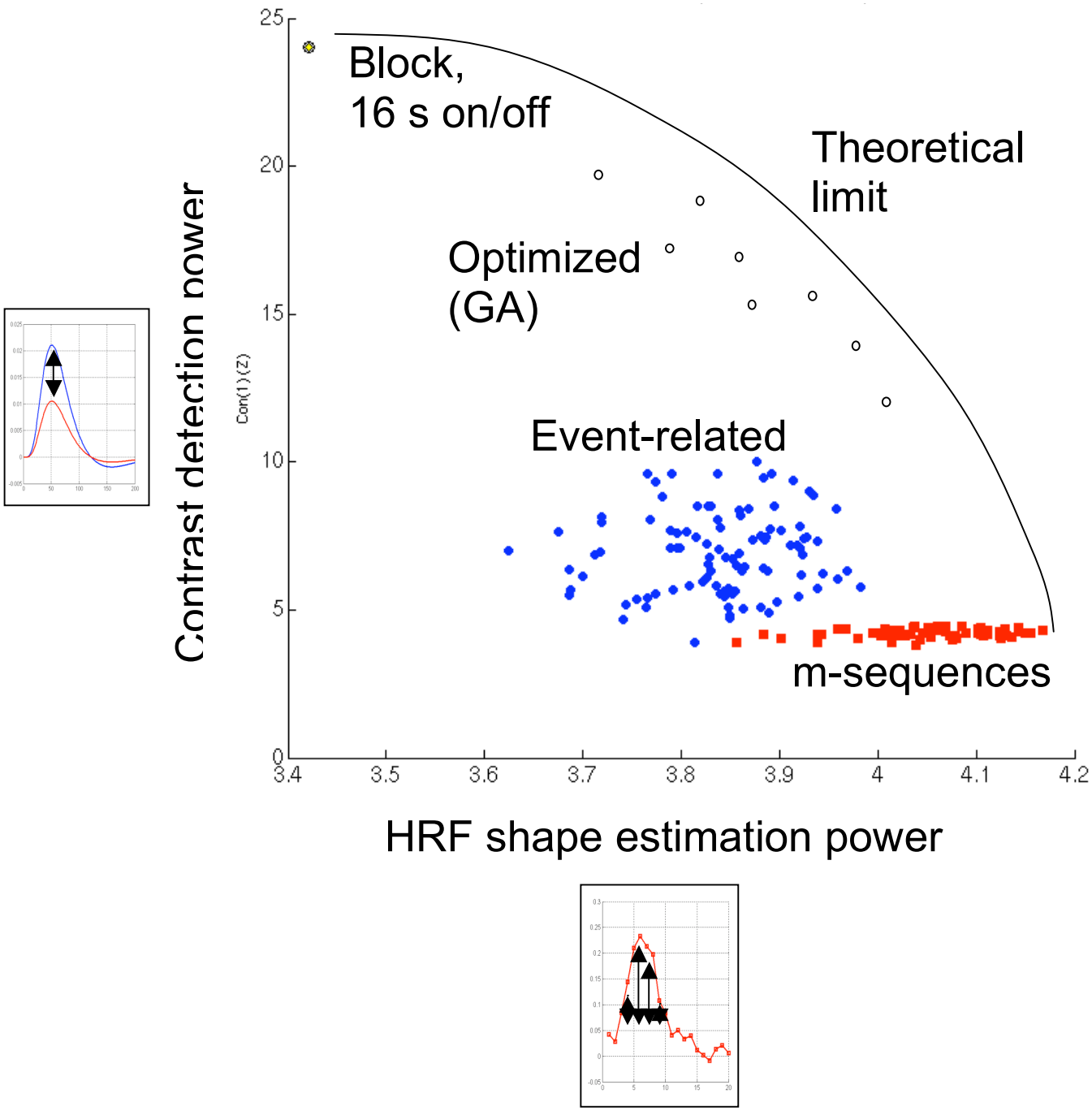


Figure 16 shows, the chances of finding a truly activated voxel with n = 10 are virtually zero. With n = 20, power is still less than 10%, but with 30 and 40 subjects, the power increases to.35 and .7, respectively.

For a given true group effect size and sample size, power depends on the within-subjects standard error, $(se(C^T\hat{\beta}))$. As shown in Eq. 8, the within-subjects standard error depends on the design matrix (X) and the residual variance, σ. σ can be reduced by optimizing data collection (e.g., pulse sequences and hardware) and in the study design by maximizing the engagement of subjects in the tasks.

However, because the standard error depends on the design matrix (X), power can also be substantially increased by carefully choosing the number, sequence, and spacing of events to minimize the design-related component of the standard error, $C^T(X^TX)^{-1}C$ in Eq. 8. Variance in the predicted response and orthogonality in the predicted responses for different events both make a statistically powerful experimental design (Liu, 2004). This is particularly critical in event-related fMRI, where delayed HRFs, overlapping responses, and autocorrelation contribute in complex ways to the overall error. It is possible to build a design in which effects can never

Figure 17



be detected, even if they actually exist! For this reason, many researchers choose to use computer-aided designs to optimize sequences of trials so that the power to estimate contrasts of interest may be maximized (Buracas & Boynton, 2002; Wager & Nichols, 2003).

Both theory and simulations show that there is a substantial tradeoff in power between *detecting* activation differences between conditions using an assumed HRF shape and estimating the *shape* of evoked activations with a more flexible model (Liu, Frank, Wong, & Buxton, 2001). This tradeoff is shown in Figure 17, in which shape-estimation power is shown on the x-axis and contrast-detection power is shown on the y-axis. The points in the model represent designs with

different sequences and timing of events. Blocked designs have the highest detection power, but provide little information about the shape of the response. M-sequences, or sequences which are orthogonal to themselves shifted in time, provide optimal shape estimation power (the nonoptimality in the figure is due to truncation of the m-sequences so they are imperfect), but low detection power (Buracas & Boynton, 2002). Random event-related designs are in between. As the Figure shows, designs optimized with a genetic algorithm (Wager & Nichols, 2003) can produce substantially better results than random designs on both measures.

Bayesian inference

Recently, Bayesian methods have received a great deal of attention in fMRI literature. Bayesian inferential methods are now key components in several major fMRI analysis software packages (e.g. SPM and FSL). Hearkening back to our discussion on inference, we made the point that P(Activation|Task state), or the probability of observing activation given a task state, is not the same as P(Task|Activation). Bayesian statistics differ from classical statistics in that the unknown true population parameters β are considered to be random variables rather than fixed constants, and what is assessed is P(β|y), where y is the observed data. According to Bayes' rule (Eq. 4), the probability of activation P(β|y) is a combination of the likelihood of observing the data and a subjective prior belief about the parameter. This combination is called the posterior distribution, and can be written:

$$P(\beta \mid y) \propto P(y \mid \beta)P(\beta) \quad \textbf{(12)}$$

Here p(β|y) denotes the posterior, p(β) the prior and p(y|β) the likelihood function. Because these are probability distributions, and β takes on a distribution of values, specifying priors involves specifying multiple parameters of the prior distribution, such as the mean and variance, as well as its distributional form. Note that if we were simply to find the values of $\hat{\beta}$ that maximize the likelihood of the observed data, we would obtain the maximum likelihood estimator (MLE) used in classical inference.

The posterior mean or mode is often used as a parameter estimate, i.e we can estimate β using $\hat{\beta} = E(\beta)$, where E denotes expected value. However, it is often difficult to calculate the exact form of the posterior distribution, and Bayesian inference therefore often requires the numerical evaluation of complicated integrals or sums. Techniques such as Markov-chain Monte-Carlo (MCMC) and Variational Bayes (VB) are used to make relevant numerical calculations about the posterior (Penny et al. 2003, Penny et al 2004, Woolrich et al 2005). An excellent overview of Bayesian statistics can be found in Gelman et. al. (2004).

As mentioned above, Bayesian methods require the definition of prior probabilities, and the choice of priors is crucial. Inference is based on a combination of evidence from the observed data and pre-existing beliefs. If strong priors are chosen, the resulting activation maps may reflect prior beliefs more than the story told by the data. If one does not want to impose such beliefs, then it is possible to use *non-informative priors*. This is an approach taken by some neuroimaging applications, e.g., FSL, which implements a fully Bayesian approach towards multi-subject analysis with non-informative priors (Woolrich, Behrens, Beckmann, Jenkinson, & Smith, 2004). For the single-level model this leads to parameter estimates that are equivalent to those obtained using classical inference. Without informative priors, it is unclear whether the Bayesian approach confers an advantage over the classical approach, although the ability to specify priors makes the Bayesian framework more flexible. Another way to choose prior beliefs is by estimating them from data. This is the 'empirical Bayes' approach. It is a hybrid between classical and Bayesian inference which is used in the SPM2 software (Friston, Glaser et al., 2002; Friston, Penny et al., 2002).

Taking a Bayesian approach allows us to calculate the probability that the activation exceeds some specific threshold, given the data (W. Penny, Kiebel, & Friston, 2003). Posterior probability maps (PPMs), e.g. in SPM2, are images depicting these probabilities. In a fully Bayesian framework, there is no hypothesis test and there are no false positives or multiple comparisons problem because one never rejects a null hypothesis. However, if PPM maps are thresholded (e.g., at 0.95) and one wants to infer that some voxels were actually activated by the task, then a hypothesis test is conducted and false positives again become an issue. It is hard to imagine a case where researchers would not want to conclude that some regions were activated by a task, so this theoretical advantage does not translate into a practical advantage. One place Bayesian methods may be advantageous is in specifying spatial priors on the regression coefficients (W. D. Penny, Trujillo-Barreto, & Friston, 2005) that incorporate prior knowledge that the evoked responses are both spatially contiguous and locally homogenous.

Multivariate analysis

The 'massively univariate' GLM approach treats each voxel as a separate dependent measure, but it is perhaps more natural to think of brain activity in terms of time-varying processes that are distributed across the brain. Likewise, it seems natural to think of psychological processes as emerging from the *interactions* among a set of brain regions—a concept which is not captured in the univariate approach. Multivariate methods such as those described below generally model the data by decomposing a large dataset (1000 time points x 100,000 voxels x 20 subjects) into a smaller set of *components* and a series of *weights*. The components may be canonical patterns of activity across time and the weights their distribution across brain space, or the other way around.

Such approaches include Principal Components Analysis (PCA), Independent Components Analysis (ICA), Canonical Variate Analysis (CVA), Partial Least Squares (PLS), Factor Analysis, and the Multivariate Linear Model (MLM). They all share the common core idea of decomposition into simpler components that maximize the amount of variability explained by the model. The approaches differ in the criteria used to select and rotate components, and in whether the experimental design is included as part of the data to be modeled.

In practice, an additional important distinction is whether the techniques are used to model the timeseries data within a participant, trials within a participant, or the individual differences in activation across participants. The interpretation of the results is very different depending on which of these types of data is used. Timeseries data can provide evidence on dynamic functional and effective connectivity across time (Mechelli, Penny, Price, Gitelman, & Friston, 2002). Trial-level data can be used to relate brain activity with performance or emotion within subjects (Rissman et al., 2004). And individual difference data can characterize patterns of individual differences across the brain, i.e., the tendency of a particular type of subject to activate a distributed attention network (Habeck et al., 2005; Lin et al., 2003) (as a note of caution, such data is often mis-interpreted as reflecting dynamic connectivity).

Though the potential for new information offered by dynamic connectivity makes this type of analysis very appealing, special concerns must be taken to prevent observed patterns from being dominated by timeseries artifacts (Lund, 2001). In addition, for most dynamic connectivity applications, it is critical to provide population-level inference by treating subjects as a random effect, as in the GLM above. Freely available toolboxes exist that can perform this type of analysis—two notable ones are the GIFT package (Calhoun, Adali, & Pekar, 2004) and the tensor-ICA routines in FSL (Beckmann & Smith, 2005).

Another important distinction is between *hypothesis-driven* and *data-driven* approaches. The GLM approach is hypothesis driven, because one seeks to find regions that fit a particular predicted model. Most multivariate approaches (e.g., PCA, ICA, FA) are data-driven, in that they simplify or reduce the data into components without reference to a predictive model (i.e., the design matrix, X). Usually, the multivariate analysis is used to generate components, and then those components are tested for relationships with X (Beckmann & Smith, 2004).

Another approach is to include X in the set of data to be reduced, or to decompose the covariance between the data and the model into components. This approach has the potential to pick out specifically those components that are task-related, but has the potential to capitalize on chance variations in the signal, and thus simulations or permutation methods must be used to control the false positive rate. Partial least squares (McIntosh, Chau, & Protzner, 2004) and semi-blind ICA are examples of this approach (Calhoun et al., 2005).

Mechanics of multivariate analysis

Each technique described in this section decomposes a matrix of data, Y, into a set of spatial and temporal components. Y is a [t x v] matrix, with t time points (observations) and v voxels. Each column of Y is the timeseries of one voxel in the brain.

Principal Components Analysis (PCA) decomposes the data by finding linear combinations of timeseries, each a column in matrix U (also of dimension t x v), such that each column of U is uncorrelated with (orthogonal to) every other column of U. The columns of U, called components, are arranged in order of variance explained: the first component explains the most variance possible in Y, the second component explains the maximal amount of remaining variance, and so forth. Together with their spatial maps and variances (described below) these v components perfectly reproduce the data, but most of the total variance is usually captured in just the first few components of U. Thus, the first components are a 'compressed' representation of the data.

Because each component is a weighted sum across timeseries of different voxels, another matrix V (of dimension voxel x component, [v x v]) contains columns of voxel weights used to create each component in U. For example, the first column of V shows how to weight each of the v voxel timeseries in order to capture the most variance in Y, and represents the spatial distribution of the first component. Thus, the columns of U are the temporal components (the 'canonical' timeseries) and those of V are the spatial components (the maps across brain voxels) of these timeseries.

If we think of each voxel as being a variable with its own axis, the weights V describe a rotation of the data Y such that the columns of YV—the new 'component scores'—fall along the axes of greatest covariance across voxels. An algebraic solution for V is given by the eigendecomposition of the covariance matrix of Y, which finds a matrix of eigenvectors V and eigenvalues (weights for V) λ such that cov(Y)V = λV. That is, rotating V by cov(Y) is equivalent to scaling V by scalar weights λ, because V lie along the principal axes of cov(Y). The weights λ are the variances of each component.

In neuroimaging, the components are usually calculated through singular value decomposition (SVD) of the centered (mean-zero) data. SVD is a more general form of PCA that decomposes the data into temporal components U and spatial components V such that:

$$Y = USV^T \quad \textbf{(14)}$$

With centered (mean-zero) data, S is a diagonal matrix (only the diagonal elements are non-zero) whose entries are the 'singular values,' the sums of squared deviations explained by each component. These are related to the eigenvalues such that $\lambda = S^2/(t-1)$. The columns of V are the eigenvectors, as in the eigendecomposition described above, and US are the component scores (components scaled by the amount of variability they explain), equal to YV in the eigendecomposition.

A thorough treatment of eigenvectors, eigenvalues and SVD is provided by Strang (1988).

Once one grasps the central idea of data decomposition into spatial and temporal components, many other techniques can be understood as variations on this theme. Independent Components Analysis (ICA), for example, is a variant of this technique. Rather than maximizing explained variance, the components are chosen to maximize the statistical independence of the

components in a more general sense. The components are not required to be orthogonal; rather, the constraint is that the distribution of one component cannot be predicted from the values of the other, and the joint probability P(A,B) of components A and B is equal to P(A)P(B). In the Infomax variant, mutual information between components (a general measure of relationship that is not necessarily linear or monotonic) is minimized (McKeown 1998a). In broad terms, ICA assumes that Y, is a weighted sum of a number of source signals (timeseries), contained in X. The data Y is a linear mixture of these source components described by the weighting or *mixing matrix* of spatial weights M:

$$Y = MX \quad \textbf{(15)}$$

There is no algebraic solution, so iterative search algorithms are used to estimate both M and X. An alternative decomposition is to transpose the data matrix and treat the spatial components as sources and the temporal components as mixing weights. For more details, we refer the reader to (Bell & Sejnowski, 1995; McKeown & Sejnowski, 1998; McKeown et al., 1998; Petersson, Nichols, Poline, & Holmes, 1999b; Petersson, Nichols, Poline, & Holmes, 1999a).

Both PCA and ICA simply reduce the data to a simpler (lower-dimension than that of the *v* voxels) space by capturing the most prominent variations across the set of voxels. The components may reflect signals of interest or they may be dominated by artifacts, and it is up to the user to determine which are 'of interest' (e.g., task-related). In ICA the order of importance of the independent components cannot be determined. Hence, it is necessary to sift through all of the components to search for ones that are task-related or otherwise of interest. However, both ICA and PCA assume all variability results from signal (noise is not modeled). A popular variant in the social sciences literature is factor analysis, which additionally fits a parameter for the noise (unexplained) variance at each voxel. However, a disadvantage of factor analysis is that the solution is *rotationally indeterminate*, and thus a number of combinations of spatial and temporal components can explain the same variability in the data. While both ICA and PCA are not rotationally indeterminate, there is some question as to what the 'right' rotation is (in PCA it is determined by variance explained). Interpreting thresholded component maps, as is commonly done, depends critically on establishing a rotation that is meaningful and reliable across studies.

These techniques as described so far model only a single subject's data. In a group study there is the additional complexity of making population inference. It is not correct to treat all the data as coming from one 'super-subject' and decomposing the group data matrix, for the same reasons that fixed effects analyses in the GLM are not appropriate. One approach is to decompose the group matrix, and subsequently 'back-reconstruct' or estimate spatial weights for each subject for a component of interest (Calhoun, Adali, Pearlson, & Pekar, 2001). The spatial weights at each voxel across subjects are treated as random variables, and one-sample t-test is conducted to test whether that voxel loaded significantly on that component in the group. Another approach, called *tensor ICA*, is to use a 3-way decomposition, using the group data to estimate temporal components and weights for each subject and each voxel (Beckmann & Smith, 2005). The subject weights at each voxel are then tested for significance.

Thresholding and multiple comparisons

As discussed above, the 'massively univariate' approach requires fitting a model at each of many voxels throughout the brain (often 100,000 or more), and constructing maps of the activation statistics and reliability over the brain (statistical parametric maps, or SPMs). In a typical behavioral experiment, test statistics whose p-values are below 0.05 are considered sufficient evidence to reject the null hypothesis, with an acceptable false positive rate (alpha) of 0.05. However, in a neuroimaging experiment, many such tests (e.g., 100,000) are conducted, and a voxel-wise alpha of 0.05 means that 5% of the voxels on average will show false positive results. In our example, that turns out to be 5,000 false positive results. Thus, even if an experiment produces *no true activation*, there is a good chance that without a more conservative

correction for multiple comparisons, the activation map will show a number of activated regions, leading to erroneous conclusions.

Many researchers use a more stringent, but arbitrary, threshold of 0.005 or 0.001 to help reduce false positives. The problem with this approach is that the chances of finding at least one false positive is very high, and its exact probability is unknown. It is generally considered desirable to limit the chances of finding a false positive somewhere in the brain to 0.05. This is the family-wise error rate (FWER), the rate at which a statistical test would be expected to produce one or more false positives among a family of tests, under the null hypothesis.

Commonly, researchers impose an additional threshold on extent of activation, requiring *k* (for example, 5) or more contiguous voxels to be significant at .005 (for example) before considering a region to be significant. This approach does not necessarily control the FWER either. The assumption that the chances of observing 5 contiguous voxels at $p < .005$ is $.005^5$ (true for independent voxels) is erroneous because imaging data are correlated ("smooth") across space. Without estimating the spatial *smoothness*, one cannot tell what the true probability of observing a contiguous activated cluster of k voxels is under the null hypothesis. Smoothness, number of subjects, and the size of the search volume vary across studies, so the FWER corrected threshold will also be different for different studies.

Another kind of control of the false positive rate is to use the false discovery rate (FDR) control. Imagine that we conduct a study on 100,000 brain voxels at alpha = .001 uncorrected, and we find 300 'significant' voxels. We expect 100 of them or 33% of our significant 'discoveries,' to be false positives. But which ones they are we cannot tell, and 33% is a substantial proportion. We may want to set a threshold such that only 5% of the significant voxels are expected to be false positives. This is FDR control at the 0.05 level. In this case, we might argue that most of the results are likely to be true activations; however, we will still not be able to tell which voxels are truly activated and which are false positives.

A review of the PET and fMRI literature shows that many investigators use uncorrected thresholds; this is likely because these studies do not have the statistical power to correct for multiple comparisons and still detect a reasonable number (or, indeed, any) of the truly activated areas. This is because as the false positive rate is controlled more conservatively, statistical power decreases. However, from the standpoint of making inferences about regional activation within a study, correcting for multiple comparisons is critical. Without correcting, any of the activations in the study might be simply false, and many such false positive regions are expected. Methods for controlling the FWER and FDR are described briefly below, along with a popular alternative, region-of-interest testing.

FWE correction. The simplest way of controlling the FWER is to use the Bonferroni correction, in which the alpha value is divided by the total number of statistical tests performed (i.e., voxels). However, if there is spatial dependence in the data (as there would be because of spatial smoothing) this is an unnecessarily conservative correction that leads to a decrease in the probability of detecting truly active voxels.

Gaussian Random Field Theory (RFT), used in SPM software, is another approach towards controlling the FWER. If the image is smooth and the number of subjects is relatively high (around 20), RFT is less conservative and provides control closer to the true false positive rate than the Bonferroni method. However, with small samples, RFT is often more conservative than the Bonferroni method. It is acceptable to use the more lenient of the two, as they both control the FWER, which is what SPM currently does. In addition, RFT is used to assess the probability of k contiguous voxels exceeding the threshold under the null hypothesis, leading to a "cluster-level" correction. Nichols and Hayasaka (T. Nichols & Hayasaka, 2003) provide an excellent review of FWER correction methods, and they find that while RFT is overly conservative at the voxel level, it is somewhat too liberal at the cluster level with small sample sizes.

RFT correction begins by estimating the spatial smoothness of the data and the consequent number of independent statistical tests, or *resels* ("resolution elements"). The number of clusters that should be found solely by chance at a given threshold is known as the Euler characteristic (EC) of the data. The RFT method assumes that the statistic maps are continuous random 3-dimensional fields of values, and so spatial smoothing is often applied (some estimates are 3 times the voxel dimensions) to satisfy this assumption. In SPM, it also assumes that the smoothness is stationary (does not vary) throughout the brain—an assumption that is often violated.

Both methods described above for controlling the FWER assume that the error values (estimated by the residuals) are normally distributed, and that the variance of the errors is equal across all values of the predictors. Nonparametric methods instead use the data themselves to find the distribution. Using such methods can provide substantial improvements in power and validity, particularly with small sample sizes, and we regard them as the "gold standard" for use in imaging analyses. Thus, these tests can be used to verify the validity of the less computationally expensive parametric approaches.

A popular package for doing non-parametric tests in group analyses, SnPM or "Statistical Non-Parametric Mapping" (T. E. Nichols & Holmes, 2002), is based on the use of permutation tests. Under the null hypothesis for a 1-sample t-test, the signs (+ or -) of the activations are distributed symmetrically around 0. In SnPM, the signs of betas across subjects are permuted, and t-statistics are computed throughout the brain. The maximum t-value from this permuted null hypothesis map is saved, and the simulation is repeated many times (e.g., 10,000) to simulate a distribution of the maximum t-value under the null hypothesis. The 95th percentile of this distribution of maxima is a threshold that provides FWER control at $p < .05$. Similar permutation tests are available for a variety of types of test, including multiple-condition tests (e.g., ANOVA) and brain-behavior correlations.

FDR control. The false discovery rate (FDR) is a recent development in multiple comparison problems developed by Benjamini and Hochberg (1995). While the FWER controls the probability of any false positives, the FDR controls the proportion of false positives among all rejected tests.

The FDR controlling procedure is adaptive in the sense that the larger the signal, the lower the threshold. If all of the null hypotheses are true, the FDR will be equivalent to the FWER. Any procedure that controls the FWER will also control the FDR. Hence, any procedure that controls the FDR only can be less stringent and lead to a gain in power. A major advantage is that since FDR controlling procedures work only on the p-values and not on the actual test statistics, it can be applied to any valid statistical test.

ROI analysis. Because of the difficulty in preserving both false positive control and power without many subjects, researchers often specify regions-of-interest (ROIs) in which activation is expected before the study is conducted. ROI analyses are conducted variously over the average signal within a region, the peak activation voxel within a region, or—a preferred method—on individually defined anatomical or functional ROIs. Another technique involves testing every voxel within an ROI (e.g., the amygdala) and correcting for the number of voxels in the search volume. This is often referred to as a "small volume correction."

Two important cautions must be mentioned. First, conducting many ROI analyses increases the false positive rate. While it may be philosophically sound to independently test a small number of areas in which activation is expected, testing many such regions violates the spirit of *a priori* ROI specification and will lead to an increased false positive rate. Small volume corrections in multiple ROIs also do not preserve the false positive rate across ROIs.

Second, although activated regions can be used as ROIs for subsequent tests, the test used to define the region must be *independent* of the test conducted in that region. Acceptable

examples include defining a region based on a main effect and then testing to see of activity in that region is correlated with performance, or using the main effect of (A+B) to define a region and then testing for a difference (A – B). Problematic examples are defining a region activating in older subjects and then testing to see if its activity is reduced in younger subjects or defining a region based on activity in the first run of an experiment and then testing whether it shows less activity in subsequent runs. Both of these are not valid tests because they do not control for regression to the mean.

A final note about uncorrected thresholds. From the standpoint of accumulation of evidence across studies, when testing large samples is impractical, using low thresholds and then using meta-analysis to see which areas are reliably activated makes some sense. Imagine that there are 20 activated regions in a task, and that 50 studies are conducted of the same task. If each study uses FWE correction, the power in each study might be 5%. Each study might be expected to activate 1 of the 20 regions, which one selected at random, and the result would be a literature of 50 papers that show different results, allowing little meaningful aggregation across those studies. Thus, it is a good idea to report results at a reasonable uncorrected threshold (e.g., $p < .001$ and 10 contiguous voxels) for archival purposes, in addition to reporting corrected results.

References

Aguirre, G. K., Singh, R., & D'Esposito, M. (1999). Stimulus inversion and the responses of face and object-sensitive cortical areas. *Neuroreport, 10*(1), 189-194.

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1997). Empirical analyses of BOLD fMRI statistics. II. Spatially smoothed data collected under null-hypothesis and experimental conditions. *Neuroimage, 5*(3), 199-212.

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage, 8*(4), 360-369.

Alsop, D. C., & Detre, J. A. (1996). Reduced transit-time sensitivity in noninvasive magnetic resonance imaging of human cerebral blood flow. *J Cereb Blood Flow Metab, 16*(6), 1236-1249.

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage, 13*(5), 903-919.

Andreasen, N. C., Arndt, S., Swayze, V., 2nd, Cizadlo, T., Flaum, M., O'Leary, D., et al. (1994). Thalamic abnormalities in schizophrenia visualized through magnetic resonance image averaging. *Science, 266*, 294-298.

Aron, A., Fisher, H., Mashek, D. J., Strong, G., Li, H., & Brown, L. L. (2005). Reward, motivation, and emotion systems associated with early-stage intense romantic love. *J Neurophysiol, 94*(1), 327-337.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., & Hyde, J. S. (1992). Time course EPI of human brain function during task activation. *Magn Reson Med, 25*(2), 390-397.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging, 23*(2), 137-152.

Beckmann, C. F., & Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *Neuroimage, 25*(1), 294-311.

Bendriem, B., Townsend, D.W. (1998). *The theory and practice of 3D PET*. (Vol. 32). Boston: Dordrecht; Boston: Kluwer Academic, 1998.

Birn, R. M., Saad, Z. S., & Bandettini, P. A. (2001). Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. *Neuroimage, 14*(4), 817-826.

Buckner, R. L., Koutstaal, W., Schacter, D. L., Dale, A. M., Rotte, M., & Rosen, B. R. (1998). Functional-anatomic study of episodic retrieval. II. Selective averaging of event-related fMRI trials to test the retrieval success hypothesis. *Neuroimage, 7*(3), 163-175.

Buracas, G. T., & Boynton, G. M. (2002). Efficient design of event-related fMRI experiments using M-sequences. *Neuroimage, 16*(3 Pt 1), 801-813.

Burock, M. A., Buckner, R. L., Woldorff, M. G., Rosen, B. R., & Dale, A. M. (1998). Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. *Neuroreport, 9*(16), 3735-3739.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*(6), 215-222. [Record as supplied by publisher].

Buxton, R. B., & Frank, L. R. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J Cereb Blood Flow Metab, 17*(1), 64-72.

Buxton, R. B., Uludag, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage, 23 Suppl 1*, S220-233.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp, 14*(3), 140-151.

Calhoun, V. D., Adali, T., & Pekar, J. J. (2004). A method for comparing group fMRI data using independent component analysis: application to visual, motor and visuomotor tasks. *Magn Reson Imaging, 22*(9), 1181-1191.

Calhoun, V. D., Adali, T., Stevens, M. C., Kiehl, K. A., & Pekar, J. J. (2005). Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis. *Neuroimage, 25*(2), 527-538.

Cheng, K., Waggoner, R. A., & Tanaka, K. (2001). Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging. *Neuron, 32*(2), 359-374.

Constable, R. T., & Spencer, D. D. (1999). Composite image formation in z-shimmed functional MR imaging. *Magn Reson Med, 42*(1), 110-117.

Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory. In (pp. 18-26). New York: Wiley.

Critchley, H. D., Elliott, R., Mathias, C. J., & Dolan, R. J. (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study. *J Neurosci, 20*(8), 3033-3040.

Critchley, H. D., Mathias, C. J., Josephs, O., O'Doherty, J., Zanini, S., Dewar, B. K., et al. (2003). Human cingulate cortex and autonomic control: converging neuroimaging and clinical evidence. *Brain, 126*(Pt 10), 2139-2152.

D'Esposito, M., Zarahn, E., Aguirre, G. K., & Rypma, B. (1999). The effect of normal aging on the coupling of neural activity to the bold hemodynamic response. *Neuroimage, 10*(1), 6-14.

Dagher, A., Owen, A. M., Boecker, H., & Brooks, D. J. (1999). Mapping the network for planning: a correlational PET activation study with the Tower of London task. *Brain, 122*(Pt 10), 1973-1987.

Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human*

Brain Mapping, 5, 329-340.
de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254-1258.
Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., & Pruessner, J. C. (2005). The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *J Psychiatry Neurosci, 30*(5), 319-325.
Disbrow, E. A., Slutsky, D. A., Roberts, T. P., & Krubitzer, L. A. (2000). Functional MRI at 1.5 tesla: a comparison of the blood oxygenation level-dependent signal and electrophysiology. *Proc Natl Acad Sci U S A, 97*(17), 9718-9723.
Duann, J. R., Jung, T. P., Kuo, W. J., Yeh, T. C., Makeig, S., Hsieh, J. C., et al. (2002). Single-trial variability in event-related BOLD signals. *Neuroimage, 15*(4), 823-835.
Duong, T. Q., Yacoub, E., Adriany, G., Hu, X., Ugurbil, K., Vaughan, J. T., et al. (2002). High-resolution, spin-echo BOLD, and CBF fMRI at 4 and 7 T. *Magn Reson Med, 48*(4), 589-593.
Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science, 302*(5643), 290-292.
Elster, A. D. (1994). *Questions and answers in magnetic resonance imaging*. St. Louis, Mo.: Mosby.
Elster, A. D., Link, K. M., & Carr, J. J. (1994). Patient screening prior to MR imaging: a practical approach synthesized from protocols at 15 U. S. medical centers. *AJR Am J Roentgenol, 162*(1), 195-199.
Fan, J., Flombaum, J. I., McCandliss, B. D., Thomas, K. M., & Posner, M. I. (2003). Cognitive and brain consequences of conflict. *Neuroimage, 18*(1), 42-57.
Frackowiak, R. S. (1997). *Human brain function*. San Diego, CA: Academic Press.
Friston, K. J., Frith, C. D., Turner, R., & Frackowiak, R. S. (1995). Characterizing evoked hemodynamics with fMRI. *Neuroimage, 2*(2), 157-165.
Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *Neuroimage, 16*(2), 484-512.
Friston, K. J., Josephs, O., Rees, G., & Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magn Reson Med, 39*(1), 41-52.
Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., & Poline, J. (2000). To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *Neuroimage, 12*(2), 196-208.
Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage, 12*(4), 466-477.
Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. *Neuroimage, 16*(2), 465-483.
Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *Neuroimage, 24*(1), 244-252.
Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage, 9*(4), 416-429.
Goutte, C., Nielsen, F. A., & Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Trans Med Imaging, 19*(12), 1188-1201.
Habeck, C., Krakauer, J. W., Ghez, C., Sackeim, H. A., Eidelberg, D., Stern, Y., et al. (2005). A new approach to spatial covariance modeling of functional brain imaging data: ordinal trend analysis. *Neural Comput, 17*(7), 1602-1645.
Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nat Rev Neurosci, 3*(2), 142-151.
Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., & Pike, G. B. (1999). Investigation of BOLD signal dependence on cerebral blood flow and oxygen consumption: the deoxyhemoglobin dilution model. *Magn Reson Med, 42*(5), 849-863.
Huber, P. J. (1981). *Robust Statistics*. New York: Wiley-Interscience.
Hubert, M., Rosseeuw, P. J., & Val Aelst, S. (2004). Robustness. In B. Sundt & J. Teugels (Eds.), *Encyclopedia of Actuarial Sciences*. New York: Wiley.
Huettel, S. A., et al. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, Mass: Sinauer Associates.
Johansen-Berg, H., Behrens, T. E., Robson, M. D., Drobnjak, I., Rushworth, M. F., Brady, J. M., et al. (2004). Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. *Proc Natl Acad Sci U S A, 101*(36), 13335-13340.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.
Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci, 23*, 315-341.
Kim, S. G. (1995). Quantification of relative cerebral blood flow change by flow-sensitive alternating inversion recovery (FAIR) technique: application to functional mapping. *Magn Reson Med, 34*(3), 293-301.
Koepp, M. J. (1998). Evidence for striatal dopamine release during a video game. *Nature, 393*(21 May), 266-268.
Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc Natl Acad Sci U S A, 89*(12), 5675-5679.

Liao, C. H., Worsley, K. J., Poline, J. B., Aston, J. A., Duncan, G. H., & Evans, A. C. (2002). Estimating the delay of the fMRI response. *Neuroimage, 16*(3 Pt 1), 593-606.

Lin, F. H., McIntosh, A. R., Agnew, J. A., Eden, G. F., Zeffiro, T. A., & Belliveau, J. W. (2003). Multivariate analysis of neuronal interactions in the generalized partial least squares framework: simulations and empirical studies. *Neuroimage, 20*(2), 625-642.

Liu, T. T. (2004). Efficiency, power, and entropy in event-related fMRI with multiple trial types. Part II: design of experiments. *Neuroimage, 21*(1), 401-413.

Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. *Neuroimage, 13*(4), 759-773.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature, 412*(6843), 150-157.

Lund, T. E. (2001). fcMRI--mapping functional connectivity or correlating cardiac-induced noise? *Magn Reson Med, 46*(3), 628-629.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W. L., & Nichols, T. E. (2005). Non-white noise in fMRI: Does modelling have an impact? *Neuroimage*.

Luo, W. L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *Neuroimage, 19*(3), 1014-1032.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proc Natl Acad Sci U S A, 97*(8), 4398-4403.

McIntosh, A. R., Chau, W. K., & Protzner, A. B. (2004). Spatiotemporal analysis of event-related fMRI data using partial least squares. *Neuroimage, 23*(2), 764-775.

Mechelli, A., Penny, W. D., Price, C. J., Gitelman, D. R., & Friston, K. J. (2002). Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities. *Neuroimage, 17*(3), 1459-1469.

Menon, R. S., Luknowsky, D. C., & Gati, J. S. (1998). Mental chronometry using latency-resolved functional MRI. *Proc Natl Acad Sci U S A, 95*(18), 10902-10907.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage, 11*(6 Pt 1), 735-759.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognit Psychol, 41*(1), 49-100.

Neter, J., Kutner, M. H., Wasserman, W., & Nachtsheim, C. J. (1996). *Applied Linear Statistical Models* (4 ed.): McGraw-Hill/Irwin.

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage, 25*(3), 653-660.

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res, 12*(5), 419-446.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp, 15*(1), 1-25.

Noll, D. C., Fessler, J. A., & Sutton, B. P. (2005). Conjugate phase MRI reconstruction with spatially variant sample density correction. *IEEE Trans Med Imaging, 24*(3), 325-336.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A, 87*(24), 9868-9872.

Ollinger, J. M., Shulman, G. L., & Corbetta, M. (2001). Separating processes within a trial in event-related functional MRI. *Neuroimage, 13*(1), 210-217.

Oswald, L. M., Wong, D. F., McCaul, M., Zhou, Y., Kuwabara, H., Choi, L., et al. (2005). Relationships among ventral striatal dopamine release, cortisol secretion, and subjective responses to amphetamine. *Neuropsychopharmacology, 30*(4), 821-832.

Paus, T. (2001). Primate anterior cingulate cortex: where motor control, drive and cognition interface. *Nat Rev Neurosci, 2*(6), 417-424.

Paus, T., Koski, L., Caramanos, Z., & Westbury, C. (1998). Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: a review of 107 PET activation studies. *Neuroreport, 9*(9), R37-47.

Pawlik, G., Rackl, A., & Bing, R. J. (1981). Quantitative capillary topography and blood flow in the cerebral cortex of cats: an in vivo microscopic study. *Brain Res, 208*(1), 35-58.

Peled, S., Gudbjartsson, H., Westin, C. F., Kikinis, R., & Jolesz, F. A. (1998). Magnetic resonance imaging shows orientation and asymmetry of white matter fiber tracts. *Brain Res, 780*(1), 27-33.

Penny, W., Kiebel, S., & Friston, K. (2003). Variational Bayesian inference for fMRI time series. *Neuroimage, 19*(3), 727-741.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Modelling functional integration: a comparison of

structural equation and dynamic causal models. *Neuroimage, 23 Suppl 1*, S264-274.
Penny, W. D., Trujillo-Barreto, N. J., & Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *Neuroimage, 24*(2), 350-362.
Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single- word processing. *Nature, 331*(6157), 585-589.
Pfeuffer, J., Adriany, G., Shmuel, A., Yacoub, E., Van De Moortele, P. F., Hu, X., et al. (2002). Perfusion-based high-resolution functional imaging in the human brain at 7 Tesla. *Magn Reson Med, 47*(5), 903-911.
Phan, K. L., Taylor, S. F., Welsh, R. C., Ho, S. H., Britton, J. C., & Liberzon, I. (2004). Neural correlates of individual ratings of emotional salience: a trial-related fMRI study. *Neuroimage, 21*(2), 768-780.
Posner, M. I., Petersen, S. E., Fox, P. T., & Raichle, M. E. (1988). Localization of cognitive operations in the human brain. *Science, 240*(4859), 1627-1631.
Price, C. J., Veltman, D. J., Ashburner, J., Josephs, O., & Friston, K. J. (1999). The critical relationship between the timing of stimulus presentation and data acquisition in blocked designs with fMRI. *Neuroimage, 10*(1), 36-44.
Reiman, E. M., Fusselman, M. J., Fox, P. T., & Raichle, M. E. (1989). Neuroanatomical correlates of anticipatory anxiety [published erratum appears in Science 1992 Jun 19;256(5064):1696]. *Science, 243*(4894 Pt 1), 1071-1074.
Riera, J. J., Watanabe, J., Kazuki, I., Naoki, M., Aubert, E., Ozaki, T., et al. (2004). A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals. *Neuroimage, 21*(2), 547-567.
Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage, 23*(2), 752-763.
Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage, 25*(1), 230-242.
Rosen, B. R., Buckner, R. L., & Dale, A. M. (1998). Event-related functional MRI: past, present, and future. *Proc Natl Acad Sci U S A, 95*(3), 773-780.
Sandler, M. P. (2003). *Diagnostic nuclear medicine.* Philadelphia, PA: Lippincott / Williams & Wilkins.
Sarter, M., Berntson, G. G., & Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *Am Psychol, 51*(1), 13-21.
Schacter, D. L., Buckner, R. L., Koutstaal, W., Dale, A. M., & Rosen, B. R. (1997). Late onset of anterior prefrontal activity during true and false recognition: an event-related fMRI study. *Neuroimage, 6*(4), 259-269.
Shulman, R. G., & Rothman, D. L. (1998). Interpreting functional imaging studies in terms of neurotransmitter cycling. *Proc Natl Acad Sci U S A, 95*(20), 11993-11998.
Shulman, R. G., Rothman, D. L., Behar, K. L., & Hyder, F. (2004). Energetic basis of brain activity: implications for neuroimaging. *Trends Neurosci, 27*(8), 489-495.
Sibson, N. R., Dhankhar, A., Mason, G. F., Behar, K. L., Rothman, D. L., & Shulman, R. G. (1997). In vivo 13C NMR measurements of cerebral glutamine synthesis as evidence for glutamate-glutamine cycling. *Proc Natl Acad Sci U S A, 94*(6), 2699-2704.
Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *Neuroimage, 20*(1), 114-124.
Silva, A. C., Zhang, W., Williams, D. S., & Koretsky, A. P. (1995). Multi-slice MRI of rat brain perfusion during amphetamine stimulation using arterial spin labeling. *Magn Reson Med, 33*(2), 209-214.
Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). ROC analysis of statistical methods used in functional MRI: individual subjects. *Neuroimage, 9*(3), 311-329.
Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23 Suppl 1*, S208-219.
Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *Am Sci, 57*(4), 421-457.
Valenstein, E. (1973). *Brain Control*. New York: John Wiley & Sons.
Vazquez, A. L., & Noll, D. C. (1998). Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage, 7*(2), 108-118.
Wager, T. D., Jonides, J., & Reading, S. (2004). Neuroimaging studies of shifting attention: a meta-analysis. *Neuroimage, 22*(4), 1679-1693.
Wager, T. D., Jonides, J., Smith, E. E., & Nichols, T. E. (2005). Toward a taxonomy of attention shifting: individual differences in fMRI during multiple shift types. *Cogn Affect Behav Neurosci, 5*(2), 127-143.
Wager, T. D., Keller, M. C., Lacey, S. C., & Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage, 26*(1), 99-113.
Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage, 18*(2), 293-309.
Wager, T. D., Rilling, J. K., Smith, E. E., Sokolik, A., Casey, K. L., Davidson, R. J., et al. (2004). Placebo-induced changes in FMRI in the anticipation and experience of pain. *Science, 303*(5661), 1162-1167.
Wager, T. D., Sylvester, C. Y., Lacey, S. C., Nee, D. E., Franklin, M., & Jonides, J. (2005). Common and unique

components of response inhibition revealed by fMRI. *Neuroimage, 27*(2), 323-340.
Wager, T. D., Vazquez, A., Hernandez, L., & Noll, D. C. (2005). Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *Neuroimage, 25*(1), 206-218.
Wang, J., Aguirre, G. K., Kimberg, D. Y., & Detre, J. A. (2003). Empirical analyses of null-hypothesis perfusion FMRI data at 1.5 and 4 T. *Neuroimage, 19*(4), 1449-1462.
Williams, D. S., Detre, J. A., Leigh, J. S., & Koretsky, A. P. (1992). Magnetic resonance imaging of perfusion using spin inversion of arterial water. *Proc Natl Acad Sci U S A, 89*(1), 212-216.
Wilson, J. L., & Jezzard, P. (2003). Utilization of an intra-oral diamagnetic passive shim in functional MRI of the inferior frontal cortex. *Magn Reson Med, 50*(5), 1089-1094.
Wong, E. C., Buxton, R. B., & Frank, L. R. (1997). Implementation of quantitative perfusion imaging techniques for functional brain mapping using pulsed arterial spin labeling. *NMR Biomed, 10*(4-5), 237-249.
Wong, E. C., Buxton, R. B., & Frank, L. R. (1998). A theoretical and experimental comparison of continuous and pulsed arterial spin labeling techniques for quantitative perfusion imaging. *Magn Reson Med, 40*(3), 348-355.
Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage, 21*(4), 1732-1747.
Woolrich, M. W., Behrens, T. E., & Smith, S. M. (2004). Constrained linear basis sets for HRF modelling using Variational Bayes. *Neuroimage, 21*(4), 1748-1761.
Zarahn, E. (2002). Using larger dimensional signal subspaces to increase sensitivity in fMRI time series analyses. *Hum Brain Mapp, 17*(1), 13-16.
Zarahn, E., Aguirre, G., & D'Esposito, M. (1997). A trial-based experimental design for fMRI. *Neuroimage, 6*(2), 122-138.
Zarahn, E., & Slifstein, M. (2001). A reference effect approach for power analysis in fMRI. *Neuroimage, 14*(3), 768-779.

Figure Captions

Figure 1. Publications by year indexed in Medline for PET and fMRI studies. Neuroimaging studies were identified by inclusively combining the search terms fMRI, PET (but not animal or companion), positron emission tomography, functional magnetic resonance imaging , or brain imaging, limited to human studies. Memory publications were identified using a title search for the word "memory," and ERP/EEG publications were identified by using an inclusive search for these terms.

Figure 2. A schematic diagram of the main components of a PET scanner.

Figure 3. PET image reconstruction. The raw data are a set of projections (sums) at different angles as shown in A. "Backprojecting" the raw data onto the image means adding the numbers of counts in the projection to the pixels that are aligned with each point in the projection, as shown in B. An image can be obtained after the data from all the projections has been added as shown in C.

Figure 4. The magnetic spin ensemble in MRI

Figure 5. Tipping the magnetization vector from the z-axis onto the xy-plane. The duration and strength of the B1 field determine how far the vector is tipped (ie – the "flip angle")

Figure 6. The dephasing process occurs because all the spins in the ensemble do not precess at the exact same rate. Some of them get a ahead, and some of them lag behind. The net effect is that they start canceling each other out, shortening the length of the magnetization vector.

Figure 7. The same slice of brain tissue can appear very different, depending on which relaxation mechanism is emphasized as the source of the contrast in the pulse sequence. Using long echo times emphasizes T2 differences between tissues, and shortening the repetition time emphasizes T1 differences in tissue. Left: one slice of a T1 image. Right: the same slice acquired as a T2 image.

Figure 8. Refocusing of spins in MR imaging by Gradient Echoes and Spin Echoes.

Figure 9. Influences on T2*-weighted signal in BOLD fMRI imaging. Courtesy Dr. Doug Noll.

Figure 10. Axial slices showing brain regions responsive to different types of switching and their overlap, from Wager et al. (2005). All voxels identified show significant switch costs in at least two switch-no switch contrasts ($p < .05$ corrected in each). Thus, many regions not shown here may also show brain switch costs at less stringent thresholds. Regions colored in red are *common activations* that show no significant differences among costs for different types of switch (at $p < .05$ uncorrected). Other regions show evidence for greater activation in some switch types than others, as indicated in the legend. I, internal; E, external; O, object; A, attribute switch types.
Figure 11. Left: the Montreal Neurological Institute average of 152 brains, used here as a spatial normalization template. Center: a subject that turned out OK. Right: a problematic normalization.

Figure 12. Examples of hemodynamic response functions (HRFs) derived from different brain regions in different studies. The brain region for each HRF is shown in the inset panels. The gray boxes at lower left in each panel show the duration of stimulus presentation. Solid lines show group-averaged high-resolution HRF estimates derived using an finite impulse response

model. Dashed lines show the fitted responses using the canonical SPM HRF, composed of two gamma functions. The impulse HRF was convolved with a boxcar function equal to the stimulus duration in A and D, providing linear predictions for the response to the epoch. The impulse HRF was fitted to the briefly presented events in B and C.

Figure 13. Construction of an event-related fMRI design matrix with four event types, using the canonical SPM HRF.

Figure 14. Basis functions (left panels), their intensity-mapped images (center), and fitted responses to data for three popular basis sets. Solid lines show group-averaged HRF estimates using a 1-s resolution FIR model. Dashed lines show the fitted response.

Figure 15. A) Principal components of fMRI noise extracted from the ventricles (fluid spaces) of a single subject. B) The power spectrum of each component, with frequency on the x-axis and energy on the y-axis. C) A true signal (blue) sampled at a lower temporal resolution (red circles), illustrating aliasing. D) The power spectra of original and estimated signals.

Figure 16. Power, or the probability of detecting activation in a truly activated voxel, as a function of effect size (Cohen's d, x-axis) and sample size (lines), using Bonferroni correction for 30,000 voxels. The search area corresponds roughly to a whole-brain search over gray matter only with 3.5 x 3.5 x 5 mm voxel sizes.

Figure 17. The tradeoff between contrast detection and HRF shape estimation power, and the performance of different types of designs on each. Power is expressed here in terms of z-scores in a simulated group analysis (n = 10, effect sizes estimated from visual cortex data in Wager et al., 2005). Yellow circle: an optimal-periodicity block design is shown in yellow. Blue circles: Randomized event-related designs. Red squares: m-sequences. Open circles: genetic algorithm (GA) optimized designs.