THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2005, 58A (2), 193–233

Psychology Press
Taylor & Francis Group

What can functional neuroimaging tell the experimental psychologist?

Richard Henson

Institute of Cognitive Neuroscience, Department of Psychology, and Wellcome Department of Imaging Neuroscience, University College London

I argue here that functional neuroimaging data—which I restrict to the haemodynamic techniques of fMRI and PET—can inform psychological theorizing, provided one assumes a "systematic" function–structure mapping in the brain. In this case, imaging data simply comprise another dependent variable, along with behavioural data, that can be used to test competing theories. In particular, I distinguish two types of inference: function-to-structure deduction and structure-to-function induction. With the former inference, a qualitatively different pattern of activity over the brain under two experimental conditions implies at least one different function associated with changes in the independent variable. With the second type of inference, activity of the same brain region(s) under two conditions implies a common function, possibly not predicted a priori. I illustrate these inferences with imaging studies of recognition memory, short-term memory, and repetition priming. I then consider in greater detail what is meant by a "systematic" function–structure mapping and argue that, particularly for structure-to-function induction, this entails a one-to-one mapping between functional and structural units, although the structural unit may be a network of interacting regions and care must be taken over the appropriate level of functional/structural abstraction. Nonetheless, the assumption of a systematic function–structure mapping is a "working hypothesis" that, in common with other scientific fields, cannot be proved on independent grounds and is probably best evaluated by the success of the enterprise as a whole. I also consider statistical issues such as the definition of a qualitative difference and methodological issues such as the relationship between imaging and behavioural data. I finish by reviewing various objections to neuroimaging, including neophrenology, functionalism, and equipotentiality, and by observing some criticisms of current practice in the imaging literature.

Correspondence should be addressed to Richard Henson, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 2EF, UK. Email: rik.henson@mrc-cbu.cam.ac.uk

This paper was written at the kind invitation of the Experimental Psychology Society, following the 11th prize lecture. The author would like to thank Tim Shallice, Mike Page, Mick Rugg, Cathy Price, Chris Frith, Ken Paller, John Duncan, Andrew Mayes, Alan Baddeley, the Methods group at the Functional Imaging Laboratory, and, in particular, Karl Friston. The author was supported by a Wellcome Trust Fellowship.

http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980443000502

Few can have escaped the recent excitement surrounding functional neuroimaging. Pictures of brains are becoming more frequent at psychology conferences; papers describing functional imaging experiments are appearing in psychological journals; many psychology departments are contributing to the purchase of brain scanners. But what do functional neuroimaging data really tell us? And do they justify the large costs associated with brain scanning, or will the brain scanners be "white elephants"? In other words, is the excitement justified?

I shall try to argue that functional neuroimaging data can be informative to the experimental psychologist. I believe that the excitement and costs are justified to some extent, although in practice there may also be unrealistic hopes and reasons to reconsider the costs if they detract from funding other aspects of psychological research. My main argument is that, provided one makes the assumption that there is some "systematic" mapping from psychological function to brain structure, then functional neuroimaging data simply comprise another dependent variable, along with behavioural data, that can be used to distinguish between competing psychological theories. I elaborate this argument below and give some examples of the use of functional neuroimaging data to inform psychological theories. I then return to discuss the validity of the "function–structure mapping" assumption and define more clearly different meanings of "systematic". Finally, I address some common criticisms levelled at functional neuroimaging.

In the present article, the term "functional neuroimaging"—or simply "imaging"—is used to refer to the so-called haemodynamic techniques of Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI). It is the data from these techniques that would appear most controversial in their relevance to psychologists. One might also broaden the definition of functional neuroimaging to include the electrophysiological techniques of electroencephalography (EEG) and magnetoencephalography (MEG), although I shall consider data from these techniques only in passing—namely, in relation to "multimodal" integration of electrophysiological and haemodynamic data.

What do the haemodynamic techniques measure? In simple terms, they measure changes in metabolic demands owing to variations in the mean activity of a large number of neurons, which result in changes in the local blood supply. By "large", we are probably referring to millions of neurons within a volume of several cubic millimetres. In fact, the PET signal, and even more so the Blood Oxygenation Level Dependent (BOLD) signal measured by fMRI, is a complex function of blood flow, oxygenation level, and volume (for more information, see, for example, Attwell & Iadecola, 2002; Heeger & Ress, 2002). This signal is typically sampled every few minutes (with PET) or seconds (with fMRI) from multiple positions within the brain—typically of the order of 100,000 such positions, or "voxels"—which gives rise to a series of three-dimensional images of brain "activity". Because of the relatively sluggish nature of blood-flow changes, the haemodynamic signal in fact reflects the integration of several seconds of neural activity (even with "event-related fMRI", in which the signal is time-locked to stimulus onset and often sampled at a sub-second rate).

It is probably mainly because of the above properties that I have heard "the use of functional imaging to understand the brain" compared with "trying to understand how a car engine works, using only a thermal sensor on a geostationary satellite" (original source unknown; apologies for plagiarism). This metaphor illustrates the relative paucity of the haemodynamic signal, the thermal aspect emphasizing its indirect and temporally impoverished nature, and the geostationary satellite emphasizing its relatively poor resolution

(relative to individual neurons). The haemodynamic signal is far from the ideal measurement of brain activity, assuming that ideal involves measuring simultaneously the activity of many individual neurons. However, no other technique in neuroscience comes close to this ideal either. Recording from single neurons is possible in animals but is rarely ethical in humans, and it is typically limited to a handful of neurons at any one time. Nonetheless, to the extent that one can still detect reliable differences in the pattern of haemodynamic signal across the brain despite the poverty of that signal, I will argue that one has learned something. In other words, its usefulness becomes an empirical question, and I will describe findings that suggest that the spatial scale of the functional organization of the brain is such that imaging data can show reliable regional dissociations as a function of psychological manipulations.

I should note that there are many cognitive neuroscientists who maintain that imaging data are far more valuable than the above "car-engine" analogy might suggest. These scientists believe, like me, that "brain data" are essential to understand the mind, as is elaborated in the later section on algorithmic versus implementational levels of description. They believe that imaging data can constrain cognitive theories in ways that cannot be addressed by behavioural experiments alone. For example, determination of the circumstances under which a tactile stimulus produces activity in visual cortex gives clues about the mechanisms of sensory integration that are not provided by behavioural data. Furthermore, imaging data allow contact with other sources of information, such as single-cell recording data from non-human primates and anatomical data about how different brain regions are connected, which can further constrain cognitive theories. Nonetheless, it is important to note that the question addressed by the present article is not whether brain data are essential to understand the mind (though this is, of course, an encompassing issue), but the more specific question of whether imaging data in particular can "feed back" to inform psychological-level theories.

Imaging data comprise another dependent variable

In what sense do imaging data comprise "another dependent variable" for the psychologist? An example might help. Imagine one is testing a psychological theory by measuring the effect of some experimental manipulation on a simple bimanual response task. Rather than recording the reaction time (RT) of key presses, a typical behavioural measure, one could also measure electromyographic (EMG) data from the relevant extensor muscle in the arms, which would provide different, but related and possibly richer, information (Figure 1a). Or one could measure the lateralized readiness potential (LRP) over the scalp above the motor cortex using EEG. Or one could measure the lateralized haemodynamic response within the motor cortex using fMRI (the "lateralized BOLD response", LBR). Although in the latter case one would lose the temporal information available in the RT or EMG/EEG data, one could still test for differences in the mean activity of left versus right motor cortex across different experimental conditions (e.g., to test for covert priming of the contralateral response: Dehaene et al., 1998). I do not see any privileged status to the behavioural data: all types of data are, in principle, observations about the system we are trying to understand, viz., the mind (with the implicit assumption of materialism—i.e., that the mind is a product of the brain). Put flippantly, accuracy and reaction times have been the "meat and two veg" of experimental psychology; I am simply highlighting a larger range of dietary options.

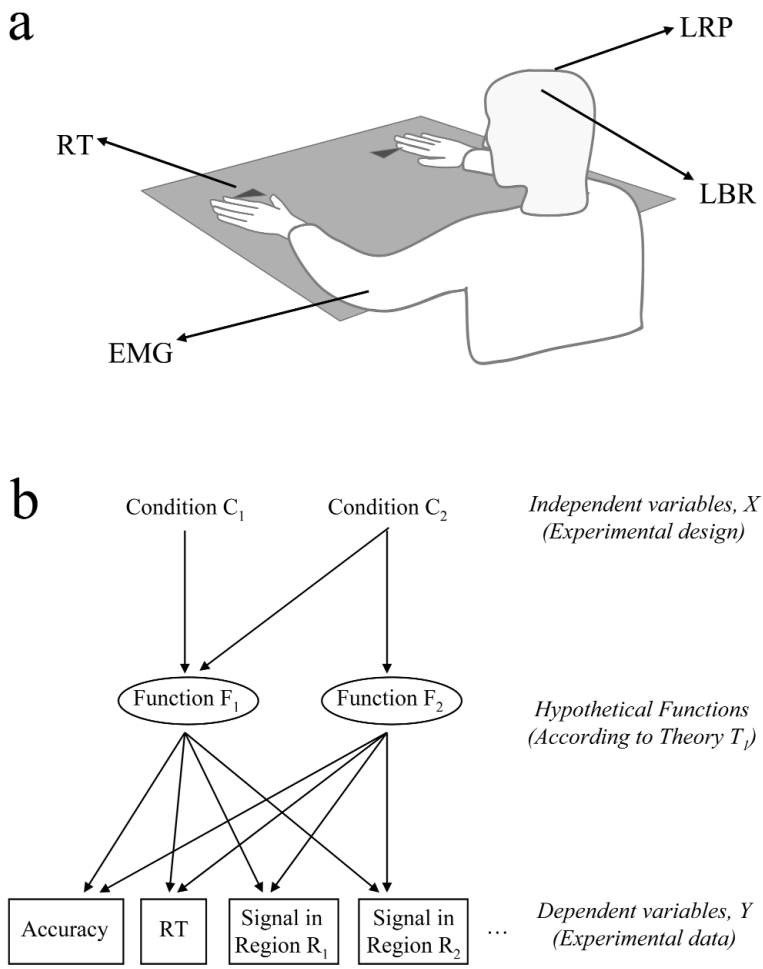

Figure 1. Dependent variables. **(a)** Different types of dependent measures in a bimanual response task: reaction time (RT), electromyographic (EMG) data, lateralized readiness potential (LRP), and lateralized BOLD response (LBR). **(b)** The hypothetico-deductive method commonly used by experimental psychologists. Note that this example assumes that the experimental psychologists are cognitive psychologists rather than behaviourists, and hence are allowed to postulate functions F_1 and F_2 as unobservable theoretical constructs.

The obvious response to this claim is that one could measure an infinite number of things in an experiment (e.g., the participant's hair length), but not all of them are relevant to the operation of the mind, or, more specifically, to the psychological-level theory under test. What makes haemodynamic changes in different parts of the brain relevant? Here we come to the crux of the argument. Functional imaging data are only relevant if there is some systematic mapping between "which" psychological process is currently engaged and "where" activity is changing in the brain. This assumption of a function–structure mapping is, in fact, a "linking" or "bridging" hypothesis (possibly unprovable in a strict sense: see later)

that needs to be made before one can use neuroimaging data to address psychological theories. It relates to the "localizationist" position that has been much debated over the years (e.g., Lashley, 1929/1963; Marshall, 1984; Uttal, 2001), particularly within neuropsychology (e.g., Caplan, 1981; Mehler, Morton, & Jusczyk, 1984; Shallice, 1988). I discuss this assumption in greater detail later, but in order to keep the argument flowing, I shall presume for the moment that it holds and illustrate how it is used.

Two types of inference from imaging data

The hypothetico-deductive reasoning employed by most experimental psychologists can be illustrated as follows. According to theory T_1, experimental conditions C_1 and C_2 differ, owing to good experimental control, only in the engagement of hypothetical function F_2 (Figure 1b). According to theory T_0, the two conditions do not differ, leading to the null hypothesis. If function F_2 affects dependent variable Y, then a reliable difference in Y under conditions C_1 and C_2 favours theory T_1 over theory T_0 and, in turn, supports the existence of function F_2. Variable Y might be a univariate behavioural variable like RT or a multivariate measure of activity over the brain. (In the present context, I will assume that the dimensions of the haemodynamic data are spatial only, ignoring their temporal dimension, and that the data are averaged over trials/replications and over subjects, which does not matter for the following inferences but is revisited later in relation to different function–structure mappings across time or across individuals.)

One can consider two types of inference made from imaging data within this framework. The first I term "function-to-structure deduction". This is a context-specific—and theory-specific—inference made from a single experiment. The second I term "structure-to-function induction". This is a context-independent inference, but one that requires stronger assumptions than does function-to-structure deduction.

Function-to-structure deduction

A function-to-structure deduction is of the following form: if conditions C_1 and C_2 produce qualitatively different patterns of activity over the brain, then conditions C_1 and C_2 differ in at least one function, F_2. The definition of "qualitatively" is considered in greater detail later but entails a reliable statistical interaction between conditions C_1 and C_2 and at least two brain regions, R_1 and R_2 (i.e., factorizing the dimension of space).[1] Note that there is an implicit inference that function F_2 is distinct from at least one other function also engaged in the experiment (e.g., Function F_1 in Figure 1b), though this other function does not need to be specific to either condition or region. The absence of any reliable difference

[1] The question of what defines a region anatomically is difficult and unresolved. A common means of classification are the Brodmann areas (Brodmann, 1909), which are defined on the basis of their cytoarchitecture (differences in cell types and their distribution). More recent classifications use multiple criteria, including data from receptor profiles and connectivity patterns, which, though they do not always agree, do suggest sharp boundaries between distinct areas (Kaas, 1987). Given that the precise areas are still debated, however, and that they are unlikely to be the smallest units of functional specialization, I do not use the term "area" (or Brodmann) in the present article. Rather, I stick with broad descriptions of functionally defined "regions" in terms of gross gyral/sulcal anatomy (more precise information can be obtained from the papers cited).

in brain activity, or the presence of only a quantitative difference (as discussed later), does not favour either T_1 or T_0—that is, it is a "null result" in classical statistics, though some might argue that this result favours T_0 if T_0 is more parsimonious.

With this type of inference, it does not actually matter "where" in the brain the qualitative differences arise, only that they exist (in the same way as absolute RTs may not matter for a given experiment, only that there is a reliable difference).[2] The nature of the function–structure mapping assumption needed to link the imaging data to psychological-level theories is also simple: the only assumption is that the same psychological function will not give rise to different patterns of brain activity within that experiment. This would seem a reasonable assumption. Note that, if one constructs "activation maps" of significant differences in C_1 and C_2 with respect to some third, baseline condition, C_0, this assumption does not require that only one region is activated in each map or that no regions are activated in common (Figure 2a). It only requires that the sets of regions activated for C_1 and C_2, relative to C_0, are not coextensive. (Note that the maps in Figure 2a are not, in fact, sufficient for a qualitative difference: additional criteria are described in the later section on qualitative versus quantitative differences.)

Structure-to-function induction

For the second type of inference, structure-to-function induction, the precise location of haemodynamic changes does matter. This argument runs: if condition C_2 elicits responses in brain region R_1 relative to some baseline condition C_0, and region R_1 has been associated with function F_1 in a different context (e.g., in comparison of condition C_1 versus C_0 in a previous experiment), then function F_1 is also implicated in condition C_2. This is illustrated in Figure 2b. The precise location matters in order to identify that it is the same brain region R_1 in both experiments, in the same way as absolute RTs might matter in order to compare across experiments. Again, there may in fact be a set of multiple brain regions in common. Note that this inference does not require the "same" regions by anatomical criteria, only that the activity differences are coextensive at the spatial scale of the data. When the engagement of function F_1 was not predicted a priori for condition C_2, this type of inference can be inductive (in a loose sense), leading to new psychological theories that relate the different experimental conditions.

Importantly, structure-to-function induction requires stronger assumptions. First, there is the recurrent problem of "association" rather than "dissociation": to accept that it is the

[2] Though location may not matter within the strict confines of the present inference, it is likely to be important within a larger theoretical framework (e.g., in relation to other data about the activated regions, such as their anatomical connectivity). Furthermore, because no two experimental conditions are identical (by definition), one could argue that it is not surprising to find a difference in corresponding activity somewhere in the brain. A similar argument has been used against the task dissociation methodology in neuropsychology (though to view a double dissociation between tasks and patients, or between conditions and activated regions, as unsurprising would suggest that an implicit assumption has already been made of the existence of two qualitatively rather than quantitatively distinct processes). This is where location may also become important—for example, that it is particularly parietal and temporal patients who show a double dissociation. Likewise, in the case of imaging, where two classes of stimuli produce different activity is generally important: a difference in visual cortex between two classes of emotionally laden visual stimuli is less interesting than a difference in amygdala, given other brain data on the importance of amygdala for some emotional responses.

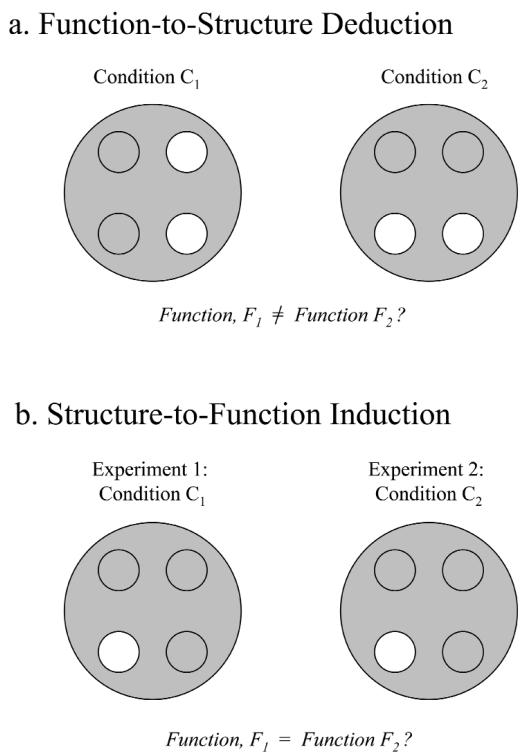

Figure 2. Two types of inference from imaging data: **(a)** function-to-structure deduction, **(b)** structure-to-function induction. Schematic statistical parametric maps (SPMs) over four brain regions (white circles = regions reliably activated versus a common baseline condition, C_0). Note that the SPMs depicted in Figure 2a are only meant to give a rough idea of the type of inference; they are not sufficient to claim qualitatively different patterns of activity (see Figure 5).

same region R_1 activated across the two experiments is to implicitly accept the (often untested) null hypothesis that there is no statistically reliable difference in the activation across experiments (though one can calculate the probability that a voxel would be activated across two orthogonal contrasts, a so-called conjunction analysis: Price & Friston, 1997). This problem is not unique to neuroimaging, however; it applies whenever one tries to generalize across experiments. Nonetheless, there is an additional assumption required for imaging data—namely, that the same regions are involved in the same functions in all contexts. This is a stronger sense of "systematicity" in the function–structure mapping.

Some examples

The above types of inference might become clearer with some examples. They all come from fMRI studies of memory (and clearly represent a biased selection!). The first two examples illustrate function-to-structure deduction within the context of recognition and encoding in long-term memory.

Recognition memory

Psychological investigation of recognition memory has been an active area for several decades, producing a rich set of behavioural findings. However, theorists can still not agree on whether a single memory process is sufficient to explain those findings, or whether one needs to appeal to at least two different types of memory process. Though one might be under the impression that "dual-process" models, which distinguish two types of memory process—such as familiarity and recollection—represent the consensus view (e.g., Yonelinas, 2002), some have argued that the evidence from recognition memory is not sufficient to reject the more parsimonious "single-process" model (e.g., Heathcote, 2003).

One type of data that has been influential in this debate derives from the "Remember/Know" paradigm introduced by Tulving (1985). In this paradigm, subjects are simply asked to indicate, when they think a stimulus has been encountered previously, whether they think that because they recollect a particular aspect of the prior encounter ("Remember"), or whether the stimulus just seems familiar ("Know"). This paradigm is clearly meant to operationalize different types of memory, and in support of this dissociations have been found between Remember/Know judgements and various experimental manipulations (see, e.g., Gardiner & Java, 1990; Rajaram, 1993; Yonelinas & Jacoby, 1995). Although the mapping between these phenomenological judgements and the underlying theoretical memory processes may not be direct (Knowlton & Squire, 1995), according to most dual-process models Remember judgements entail recollection (and possibly familiarity too), whereas Know judgements do not (for which familiarity is sufficient).

However, most of the dissociations between Remember/Know judgements are single dissociations, and others have argued that they can be explained by a single-process model and signal-detection theory (e.g., Donaldson, 1996; Hirshman & Master, 1997; Inoue & Bellezza, 1998), in which the judgements simply reflect different response criteria along a single continuum of "memory strength". In other words, it is argued that the Remember/Know dissociations can be explained by a quantitative rather than a qualitative difference in the underlying memory process.

We conducted an event-related fMRI study of recognition memory for words using the Remember/Know paradigm (Henson, Rugg, Shallice, Josephs, & Dolan, 1999). When we contrasted the haemodynamic response associated with each type of judgement (conditionalized on those judgements being correct), we found some cortical regions, particularly in left hemisphere, that were more responsive for Remember (R) than for Know (K) judgements, and other regions, particularly in the right prefrontal cortex, that were more responsive for Know than for Remember judgements. Indeed, comparison of the haemodynamic response in, say, left inferior parietal cortex with that in right dorsolateral prefrontal cortex (Figure 3a) would reveal a double dissociation—indeed, a cross-over interaction relative to New judgements. In other words, qualitatively different patterns of brain activity were associated with different subjective experiences to stimuli that were objectively all "old".

In the framework outlined above, one could view dual-process models as examples of theory T_1 and single-process models as examples of theory T_0. In this case, making only the weak assumption that nonidentical patterns of brain activity imply at least one different function, the imaging data support dual-process models over single-process models. Whether

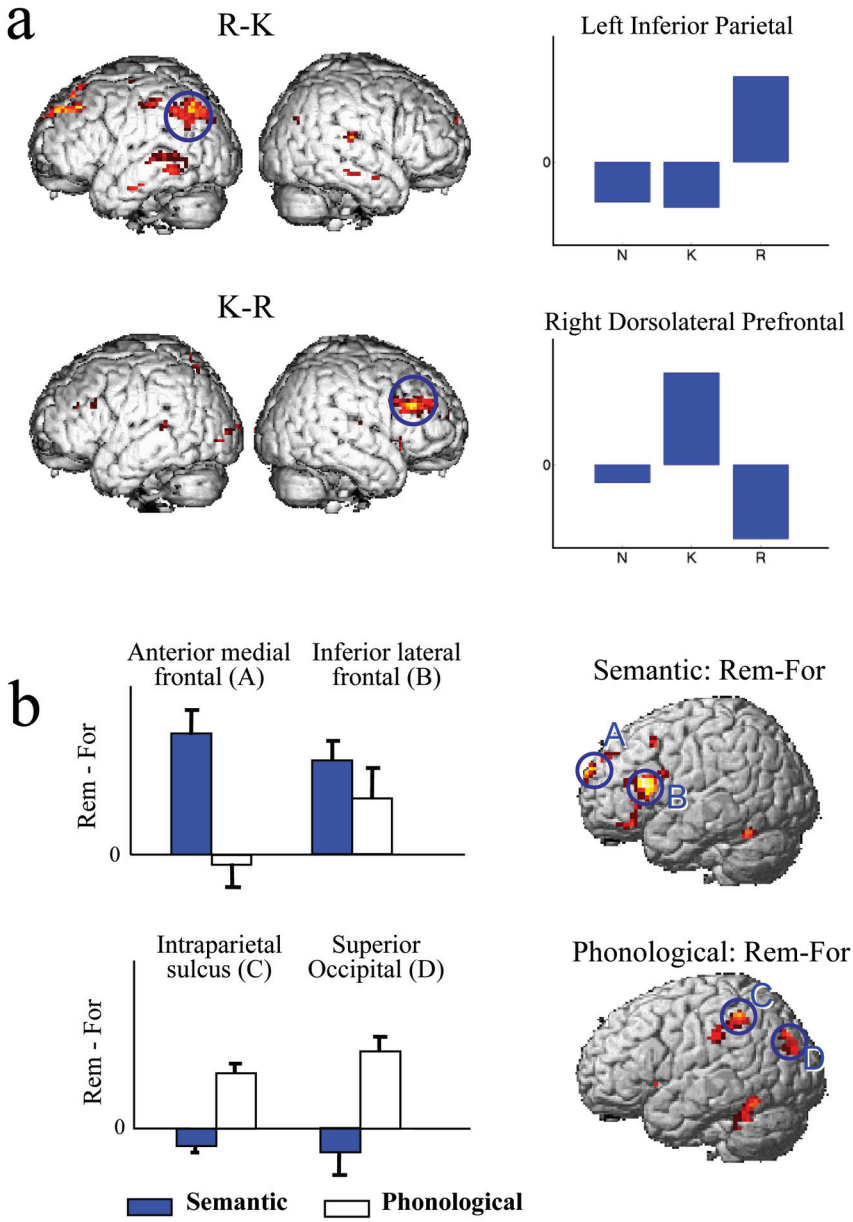

Figure 3. Function-to-structure deduction: examples from long-term memory. **(a)** Red/yellow "blobs" consist of voxels reliably more active for correct Remember (R) than for correct Know (K) judgements (top image), or more active for correct K than correct R judgements (bottom image), during recognition of studied words (Henson et al., 1999), rendered onto the left and right lateral surfaces of a canonical brain. Bar plots show estimates of the magnitude of event-related BOLD responses (units and zero-value arbitrary) for each condition, including correct rejections of unstudied ("new") words (N), from the left inferior parietal and right dorsolateral prefrontal regions circled in blue on the brains. **(b)** Subsequent memory effects during semantic or phonological study tasks (Otten & Rugg, 2001). Bar plots show differences in magnitude of event-related responses to (confidently) Remembered (Rem) and Forgotten (For) words within circled regions in anterior medial and left inferior prefrontal cortex (top image, only former showed interaction that reached significance) and left intraparietal and superior occipital cortex, respectively (bottom image).

that function is recollection or familiarity depends on the precise model of how these processes map to Remember/Know judgements (Knowlton & Squire, 1995). In this case, further imaging experiments (using source memory tasks, for example, where recollection can be objectified) have indeed associated left lateral (and medial) parietal responses with recollection (Rugg & Henson, 2002). I think it would be difficult to explain these imaging data in terms of purely quantitative differences in memory strengths.

Now, one might argue that the differential brain responses reflect other, nonmnemonic differences between Remember and Know judgements—that is, other psychological differences that confound this comparison: a violation of the "pure insertion" assumption (discussed later). Indeed, in subsequent experiments, we interpreted the right dorsolateral prefrontal response in terms of post-retrieval decision processes rather than familiarity or memory strength (Henson, Rugg, Shallice, & Dolan, 2000c) and have since associated familiarity with response reductions in anterior medial temporal cortex (Henson, Cansino, Herron, Robb, & Rugg, 2003c). However, this possibility of confounding variables/alternative explanations is not unique to neuroimaging: it applies equally well to any experimental science. The solution is simply to conduct further experiments to test the alternative explanations.

Memory encoding

Another example of a function-to-structure deduction comes from work on encoding into long-term memory. In the so-called subsequent-memory paradigm, participants perform a simple task on a series of stimuli (the "study" phase), though they are not told of any requirement to remember the stimuli. In a later test phase, they are presented with a surprise memory test, data from which are used to "back-sort" the stimuli at study into those subsequently remembered and those subsequently forgotten. Imaging data acquired during the study phase can then be used to see which brain regions "predict" subsequent memory.

This paradigm has been used to investigate different theories of memory encoding. According to what I shall call "structural" theories, there exists a cognitive system specialized for episodic memory (e.g., Cohen & Squire, 1980; Schacter & Tulving, 1994; the distinctions between "episodic" and "declarative" do not matter here). According to "proceduralist" theories on the other hand (e.g., Kolers & Roediger, 1984; Morris, Bransford, & Franks, 1977), memory is better viewed as a by-product of the processes performed when a stimulus is encountered (again, this view is simplified somewhat for the present argument). An important difference between these two types of theories is whether successful remembering always involves a specific psychological process, supported by a specialized memory system, or whether remembering can be associated with different processes on different occasions. This can be tested by comparing subsequent memory effects under different study tasks: according to the structural theories (T_0), the brain regions correlating with subsequent memory should not differ across tasks, whereas according to procedural theories (T_1), they should.

Otten, Henson, and Rugg (2001) compared brain activations for subsequently remembered versus subsequently forgotten words as a function of whether the study task encouraged semantic processing (whether the word referred to an animate or inanimate entity) or

orthographic processing (whether the first and last letters in the word were in alphabetical order). In both cases, activity in the left medial temporal lobe correlated with subsequent memory.[3] Though only an association rather than dissociation (see above), this finding is at least consistent with the structural view, and with the neuropsychological evidence that structures within the medial temporal lobe are critical for memory. However, a subsequent experiment by Otten and Rugg (2001) found clear evidence for the procedural view. This experiment compared a semantic study task with a phonological—rather than an orthographic—study task and found a region within anterior medial prefrontal cortex that showed greater subsequent memory activations under the semantic task, and regions within left intraparietal sulcus and superior occipital cortex that showed greater subsequent memory activations under the phonological task (Figure 3b). The same prefrontal and parietal/occipital regions were also more responsive (to all study items) under one task than under the other, suggesting a general role in semantic or phonological processing of words, respectively, and further supporting the procedural view. (An interesting follow-up to these experiments will be to test whether the brain regions predicting subsequent memory depend on the type of memory test, in addition to the type of study task, as would be expected by procedural theories such as "transfer-appropriate processing": Morris et al., 1977.)

Note that the activation patterns in the two experiments by Otten and colleagues would not simply appear to reflect unreliability in the imaging methods, since both patterns have been replicated (e.g., medial temporal encoding-related activations across different study tasks by Fletcher, Stephenson, Carpenter, Donovan, & Bullmore, 2003, and dissociable encoding-related cortical activations as a function of study task by Otten, Henson, & Rugg, 2002). Rather, these imaging studies suggest that elements of both the structural and procedural theories are correct and that it may be fruitful to consider how those elements can be combined.

Short-term memory

An example of the second type of inference, structure-to-function induction, comes from an imaging experiment on short-term memory (Henson, Burgess, & Frith, 2000b). The background to this experiment concerns computational models of verbal short-term memory. A plethora of such models have emerged recently (e.g., Anderson & Matessa, 1997; Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1999; Henson, 1998; Lee & Estes, 1981; Page & Norris, 1998), which all do a reasonable job in capturing the main aspects of the behavioural data. Indeed, distinguishing between the models requires consideration of the minutiae of somewhat contrived tasks (in the sense that they are far removed from everyday tasks). This reflects a trend in experimental psychology for very detailed investigation of a limited number of specialized domains. One could argue that the danger of such focused "cottage industries" is that we forget the larger question of how the psychological constructs

[3] Note that remembered and forgotten words did not differ in their reaction times in the study tasks, ruling out simple accounts in terms of "time on task". Note also that there was an apparent difference in the latency by which the subsequent memory effect emerged in the haemodynamic response, and an alternative, "proceduralist" explanation (to be tested experimentally) is that the particular items remembered in the orthographic task underwent incidental semantic processing, despite the task instructions.

within these domains are interrelated. One hope, perhaps overly optimistic, is that the brain can provide a common framework within which these constructs might be related, by comparing functional neuroimaging results across such domains.

In any case, one of the key problems addressed by computational models of short-term memory is the problem of serial order: that is, how we maintain the order of a novel sequence of items, such as the digits in a telephone number. We designed two probe tasks that differed in their requirement to maintain temporal order, while controlling for memory for the items themselves. In the item probe task, participants saw a temporal sequence of six letters, followed by a single probe letter, and decided whether or not the probe was one of the letters in the previous sequence (a task originally popularized by Sternberg, 1969b). In the list probe task, participants saw the same type of sequence, but the probe now consisted of the same six letters presented simultaneously, and participants decided whether the order of the letters—from left to right—matched the temporal order of their prior presentation. Performance of this task, but not of the item probe task, involves maintenance of serial order (see Henson, Hartley, Flude, Burgess, & Hitch, 2003b, for behavioural evidence for this claim).

Furthermore, temporal grouping of a sequence of items (e.g., into threes or fours, as is common with telephone numbers) is known to improve memory for their order. According to some models, such grouping reflects a change in a "timing" signal that is used to maintain the order of items (e.g., Burgess & Hitch, 1999; Brown et al., 2000). We therefore added a grouped version of the list probe task, splitting the to-be-remembered sequence into two groups of three with an extra pause, holding the total presentation time constant. We were interested in whether the same brain regions implicated in maintaining serial order by comparison of list probe and item probe tasks would be further modulated in comparing grouped and ungrouped versions of the list probe task.

The fMRI data revealed that part of dorsal left premotor cortex (dorsal to Broca's area) was more active in the list probe than the item probe task, but less active in the grouped than the ungrouped list probe task (Figure 4a). This is consistent with utilization and modulation, respectively, of a timing signal like that proposed by Burgess and Hitch (1999).[4] It is also consistent with independent behavioural work by Saito (2001). But more important for the present argument is the observation that dorsal left premotor cortex has also been implicated in timing by other studies. An imaging study by Catalan, Honda, Weeks, Cohen, and Hallett (1998) found that this region was more active for sequential than for repetitive finger movements, while a neuropsychological study by Halsband, Tanji, and Freund (1993) found that patients with damage to this region had difficulties reproducing rhythmic motor sequences. One could argue that a common activation site in the brain not only links these traditionally separate domains (verbal short-term memory and motor control) but mutually reinforces the corresponding domain-specific theories that relate a timing function to certain tasks. Furthermore, imagine that there were no prior theorizing about timing signals in

[4] The question of whether grouping increases or decreases the metabolic demand in a brain region generating a timing signal is not directly addressed by this model (though it could be elaborated to do so). This does not detract from the point here that the presence of reliable differences in the same cortical area (at the spatial scale of the data) when comparing the List Probe against Item Probe tasks and grouped against ungrouped List Probe tasks, whether activations or deactivations, is consistent with a timing-signal account of memory for serial order.

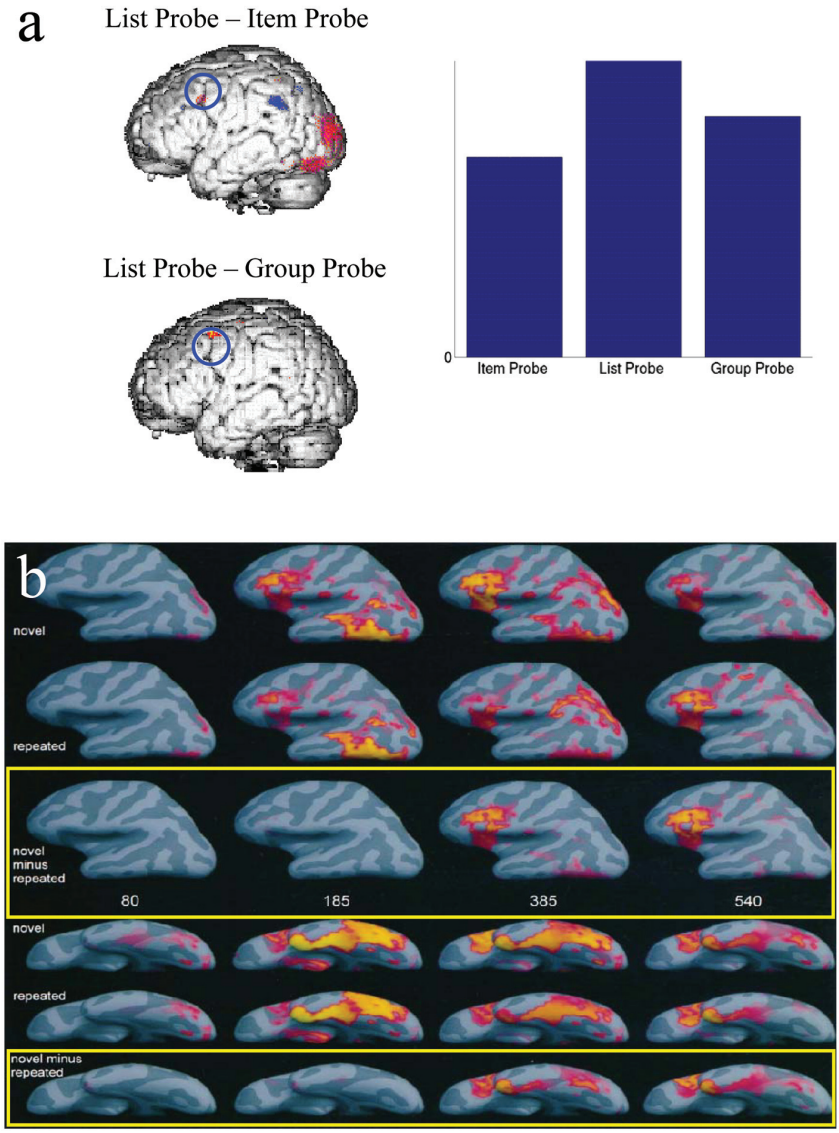

Figure 4. Structure-to-function induction. **(a)** Comparison of list probe and item probe tasks (top image) and grouped (group probe) and ungrouped versions of the list probe task (bottom image) in verbal short-term memory (Henson et al., 2000b). Bar plots show mean haemodynamic response during blocks of each task (scale and zero-value arbitrary) from the left dorsal premotor region circled in blue on both images (activations are coextensive at a lower threshold). See Figure 3 legend for more details. **(b)** Snapshots of estimated neural activity based on MEG data constrained by fMRI data, shown on lateral (upper 3 rows) and inferior (lower 3 rows) views of an inflated left hemisphere for first presentations of words ("novel"), repeated presentations, and their difference (highlighted by yellow rectangles), at four post-stimulus times (80, 185, 385, and 540 ms, columns left to right) in a semantic decision task. [Adapted from *Neuron, 26,* 55–67, Dale et al., "Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity", Copyright (2000), with permission from Elsevier Science.]

short-term memory. Then one could use this common activation site to induce a new theory of how serial order is maintained in short-term memory. In the earlier terminology, the prior attribution of function F_1 to this region can be used to infer that function F_1 was also engaged in the present experimental context (again, making a strong systematicity assumption in the function–structure mapping).

In short, different experiments can be linked through a function–structure mapping if they activate the same regions. This enables the interpretation of one experiment to inform another and implicitly enforces an internal consistency over these interpretations.

Repetition priming

The final example is a more accurate historical example of theoretical induction from neuroimaging data, at least from a personal perspective, though it derives from combining both haemodynamic and electrophysiological data. As such, it also illustrates the exciting new technique of "multi-modal" imaging, in which these different types of data are formally combined to estimate spatiotemporal "movies" of activity over the brain (marrying the temporal information in EEG/MEG with the spatial information in PET/fMRI).

The relevant psychological phenomenon is repetition priming. Neuroimaging has shown that repetition of visual stimuli often results in a reduction in the haemodynamic responses in extrastriate occipitotemporal cortex (cortex that is reasonably "early" in the visual processing stream, though not as "early" as V1 and V2 in striate cortex). This reduction can survive multiple intervening stimuli and often parallels the amount of perceptual priming. The reduction also tends to be restricted to parts of occipitotemporal cortex that are particularly responsive to the specific category of stimuli used. For example, we find decreased haemodynamic responses for repeated versus initial presentations of familiar faces in parts of fusiform cortex (Henson, Shallice, & Dolan, 2000a) that are likely to correspond to the "Fusiform Face Area" (FFA), an area that some have argued is specialized for faces (Kanwisher et al., 1997; though see Gauthier, Skudlarski, Gore, & Anderson, 2000; Haxby et al., 2001).

One of the earliest components of the event-related potential (ERP) recorded with EEG that is believed to be specific for faces is the "N170", a prominent peak occurring at around 170 ms poststimulus (e.g., Bentin, Allison, Puce, Perez, & McCarthy, 1996). One might suppose that the FFA and N170 are the haemodynamic and electrophysiological correlates, respectively, of the same process of face perception (i.e., the constituents of this form of structure-to-function induction might be viewed as a brain region and an ERP component, rather than as two brain regions). Unlike the FFA, however, there is little evidence that the N170 shows any effect of repeating faces (Henson et al., 2003a). Rather, ERP repetition effects associated with priming (discounting the potentially special case of immediate repetition) tend to emerge somewhat later, typically onsetting 300–400 ms poststimulus (Henson & Rugg, 2002). Together with the above fMRI findings, these data suggest that haemodynamic repetition effects in occipitotemporal regions do not necessarily reflect facilitation of initial stages of visual object recognition (though they may reflect processes giving rise to later ERP repetition effects). This is contrary to what one might expect from an automatic, "bottom-up" theory of perceptual priming and reinforces the fact that, because the haemodynamic response integrates over several seconds of neural activity, haemodynamic differences in a

brain region that is viewed as "early" in the visual processing pathway do not necessarily occur early in time.

However, such informal comparison of haemodynamic and electrophysiological data is difficult. The N170 might not originate solely from the FFA, for example (Henson et al., 2003a). In fact, it is difficult to localize in the brain the sources of EEG or MEG data recorded at the scalp—indeed, it is impossible if the number of sources is unknown: the so-called "inverse" problem. This is where formal techniques for combining the two types of data become useful. Though the inverse problem is ill-posed, the addition of further constraints on the possible locations of the EEG/MEG sources (such as from structural and functional MRI data) allows at least a best estimation of their location subject to those constraints.[5]

Such methods were used by Dale et al. (2000) to examine the spatiotemporal evolution of MEG repetition effects (using words). Snapshots of the resulting "movies" are shown in Figure 4b. The important insight I gained from these multimodal data is that the repetition effects emerge not only relatively late in time (peaking at about 385 ms), but also after "activity" in response to the onset of the words has reached more anterior brain regions, including temporal pole and prefrontal cortex (by 185 ms). This prompted me to consider models of repetition priming that emphasize feedback to "early visual" regions from more anterior, "higher-order perceptual" regions. In particular, following the ideas of Friston (2002b), the data encouraged me to pursue ideas behind predictive-coding models (Rao & Ballard, 1999). In psychological terms, priming may have more to do with alterations in top-down expectancy than with bottom-up evidence. This theoretical induction from imaging data, though possibly trivial to some, certainly influenced my thinking.

Some further clarifications

I hope that the above examples have illustrated the type of psychological inferences one might make from treating imaging data as another dependent variable. We now need to return to the assumption of a systematic function–structure mapping that underlies these inferences. Before doing so, however, I would like to revisit a few subsidiary issues that have emerged during the above examples.

Qualitative versus quantitative differences

In order to make a function-to-structure deduction, I argued that one must find a qualitative, rather than simply quantitative, difference in the pattern of activity over the brain. The most common analysis of haemodynamic data is a "mass, univariate" approach, as exemplified by the popular Statistical Parametric Mapping (SPM) software package (Friston et al., 1995). This means that statistical tests across conditions (such as *t*-tests) are performed separately for each voxel. The SPM package, when used for classical inference, includes

[5] These constraints can be "soft" rather than absolute constraints, as implemented, for example, by priors within a Bayesian framework, which is important given that one can imagine hypothetical situations where fMRI and EEG/MEG data would diverge (e.g., brain regions with electrically closed fields, or changes in the synchrony of neural activity in the absence of metabolic changes).

sophisticated methods for correcting the resulting *p*-values for the fact that multiple tests are being performed across voxels—taking into account, in particular, the fact that the tests are not independent, given that data from nearby voxels tend to be correlated. This leaves one with a subset of voxels, normally clustered into regions, that survive this statistical threshold. The researcher typically infers that these regions are "specialized" for the function of interest (that is hypothesized to differ across the conditions). This inference is valid as long as one does not try to claim that they are the only regions, since other regions may not survive the threshold owing to type II error, or that they are "more specialized" than are other regions.[6]

Yet it is the latter type of claim—that responses are greater in some regions than in others—that underlies what I have called a qualitative difference. To test this claim, one needs to perform statistical tests across two or more regions. In other words, one needs to include "region" as a factor in the analysis and show a reliable interaction between condition and region. Failure to do so is like testing simple effects without testing the interaction (with some simple effects not reaching significance simply because of greater noise; see Figure 5a). In other words, a minimal requirement for deducing the presence of a different function (F_2) is an interaction in which one region shows a reliably greater change in activity across conditions than at least one other region.

One could argue that comparisons across regions are potentially confounded by differences in the mapping from neural to haemodynamic responses in different regions (owing, for example, to variability in the vasculature across the brain). Differences in the multiplicative gain of this neural-to-haemodynamic mapping, in particular, could lead to an interaction in the haemodynamic data using an additive model like ANOVA, even if no interaction were present in the underlying neural activity. Nonetheless, I would argue that testing for an additive interaction is still better than not testing for regional interactions at all. But, most importantly for the present context, similar issues arise in possible mappings from psychological processes to behavioural measures (e.g., whether or not an RT measure of priming, RT_1–RT_2, scales with absolute RT, RT_1). Generally, we need to remind ourselves of the various assumptions made in interpreting behavioural data, such as RTs (e.g., Meyer, Osman, Irwin, & Yantis, 1988).

Though a Region × Condition interaction is necessary, it is not sufficient. The particular regions should also be differentially active in C_1 and/or C_2 relative to a third baseline condition, C_0. This partly reflects the fact that imaging signals are usually relative rather than absolute, and so a common reference is needed to compare across regions—for example, to determine whether an interaction is a cross-over interaction. It would not be sufficient to show that the difference between C_1 and C_2 in region R_1 (e.g., in visual cortex) is reliably greater than the difference between C_1 and C_2 in another region R_2 (e.g., in auditory cortex) that was not active in either C_1 or C_2 relative to C_0. Indeed, it is probably always possible to find a voxel that is not differentially active across any of the experimental conditions and therefore liable to furnish an interaction with another region that is. This pattern could still reflect the same function, F_1, subserved by region R_1 but not R_2, that is engaged by both C_1

[6] Note that it is not sufficient to report two statistical maps, one for each condition versus a common baseline, and observe that they look different (as in Figures 2a and 5a). This is a common mistake ("imager's fallacy"), since it involves an implicit acceptance of the null hypothesis (where the maps differ). One should not eyeball differences in statistics, but explicitly test statistics of differences.

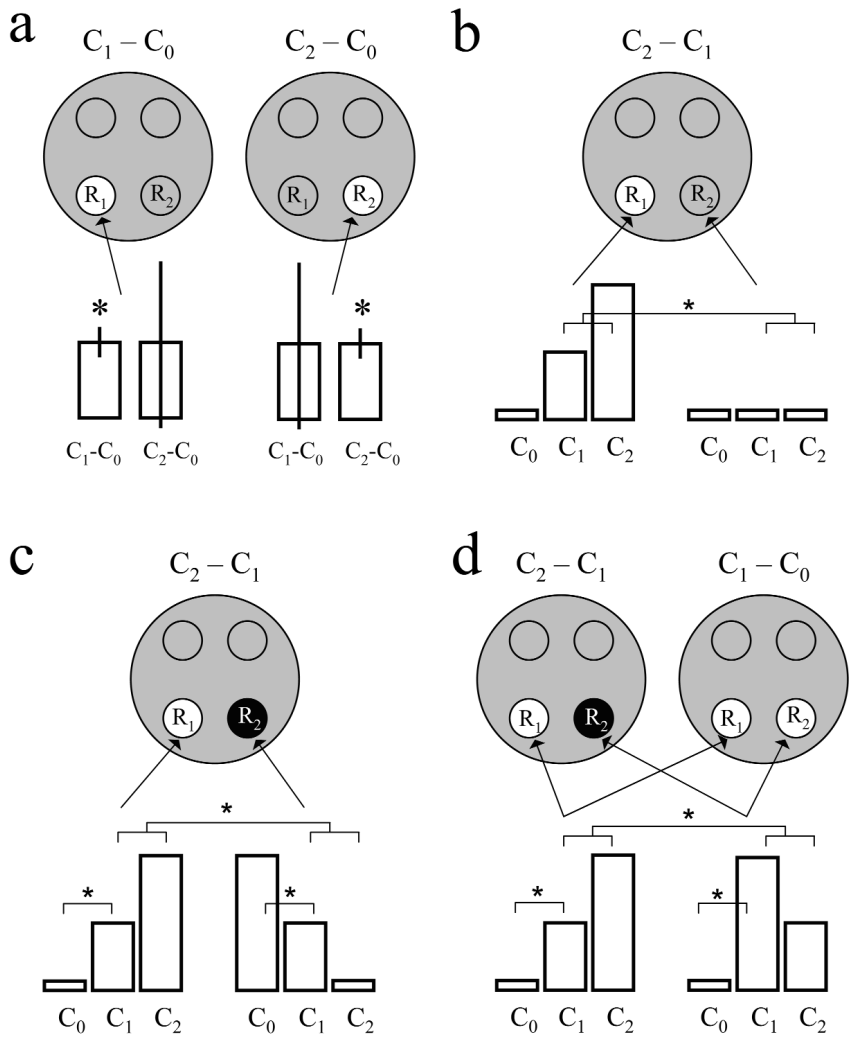

Figure 5. Qualitative versus quantitative differences. Schematic SPMs over four brain regions together with activity profiles for selected regions (white circles = regions reliably activated, black circles = regions reliably deactivated, for stated subtraction of conditions, asterisks indicate reliable differences between indicated conditions). **(a)** Region R_1 but not R_2 shows a significant difference between conditions C_1 and C_0, and region R_2 but not R_1 shows a significant difference between conditions C_2 and C_0. Such a finding is not sufficient for either a qualitative or quantitative difference since the null results for C_1–C_0 in R_2 and for C_2–C_0 in R_1 could reflect differences in means that are comparable to those in R_1 and R_2, respectively, but simply in the presence of greater variability ("imager's fallacy"). **(b)** A significant interaction is found between regions R_1 and R_2 and conditions C_1 and C_2 (a single dissociation), but R_2 is not reliably differentially active across conditions C_0, C_1, or C_2. This is a quantitative but not qualitative difference. **(c)** A significant interaction is found between regions R_1 and R_2 and conditions C_1 and C_2 (a double dissociation), but R_1 is activated in C_1 relative to C_0, whereas R_2 is deactivated for C_1 relative to C_0. This is not sufficient for a qualitative difference (see text). **(d)** A significant interaction is found between regions R_1 and R_2 and conditions C_1 and C_2 (a dissociation), and both regions are activated in C_1 relative to C_0 (an association). This is sufficient for a qualitative difference.

and C_2, but simply twice as much—or for twice as long—in C_2 than in C_1 (Figure 5b). In other words, this single dissociation could reflect only a quantitative difference. One needs to show that R_2 is also differentially active in either C_1 or C_2 relative to C_0. In the earlier examples, condition C_0 would correspond to the New judgements in Figure 3a, or the Forgotten items in Figure 3b, in which the bars already reflect differences between remembered and forgotten items.

Even differential activity with respect to a baseline condition is not sufficient, however: a final criterion is that the difference between C_1 and C_2 is not in the same direction as that between C_0 and C_1 in both R_1 and R_2. This reflects the fact that, unless one wants to make a further assumption that increases in the haemodynamic response necessarily reflect increases in the degree to which a function is engaged (via the concept of a "resource"—Shallice, 2003; see also Sternberg, 2001), activity in some regions may be positively correlated with a function, whereas activity in other regions may be negatively correlated. This might arise, for example, if regions R_1 and R_2 have reciprocal connectivity in their implementation of function F_1. Then a pattern like that in Figure 5c could still reflect a quantitative difference in the engagement of the same function F_1, even though that function "activates" R_1 and "deactivates" in R_2. Note that this means that even a double dissociation in the activity of R_1 and R_2 across conditions C_1 and C_2 is not, on its own, sufficient. Conversely, a pattern like that in Figure 5d would meet the above criteria.

The cross-over pattern in Figure 5d resembles what Dunn and Kirsner (1988) call a "reversed association", in whose terms the different brain regions would correspond to different tasks—that is, the activity of the two brain regions is positively related in one contrast, C_1–C_0, but negatively related in another, C_2–C_1. The present criteria are slightly less constrained, however, in that only one of the two regions needs to be differentially active in C_1 or C_2 versus C_0 (e.g., $C_0 < C_1 < C_2$ in R_1, and $C_0 = C_1 > C_2$ in R_2) and in that a single dissociation across C_1 and C_2 (e.g., $C_0 < C_1 < C_2$ in R_1, and $C_0 < C_1 = C_2$ in R_2) is sufficient (though the latter would further rely on the additivity assumption implicit in a significant non-cross-over interaction, as described earlier in relation to the neural-to-haemodynamic mapping).

The Remember/Know experiment described earlier and in Figure 3a meets these criteria (as does the encoding example in Figure 3b, though with four conditions). First, there is a significant interaction between left parietal versus right prefrontal regions and R versus K judgements. Second, both regions include at least one condition that is reliably different from a baseline condition C_0, which corresponds here to N (correct New) judgements, viz., for R judgements in the left parietal region and for K judgements in the right prefrontal region. Finally, the R versus K difference is not in the same direction relative to N judgements in the two regions: in the right prefrontal region, for example, R judgements produce significantly less activity than do K judgements, but K judgements produce significantly more activity than do N judgements.

The above criteria are not meant to be prescriptive. While experimental findings that meet all of the criteria are an attractive goal, evidence for or against a theory clearly constitutes a continuum, in which the presence of some form of dissociation is better than no evidence at all. I have also considered only two regions, perhaps selected by independent contrasts or based on a priori knowledge. There are practical problems if one has a large number of regions (in the limit, all voxels in an image) and wishes to factorize these regions for

analysis (along spatial dimensions, e.g., left versus right). While multivariate statistical techniques have been applied to neuroimaging (see, e.g., Kherif et al., 2002), they are not sufficient to test all the above criteria, and new techniques may need to be developed in such cases.

Now one might imagine imaging data that satisfy the above criteria, but which one might still not want to claim imply functional differences. For example, imagine experimental conditions of silence (C_0), low-frequency tones (C_1), and high-frequency tones (C_2), and two regions (R_1/R_2) within a tonotopic map in auditory cortex (M. Page, personal communication, January 2004). Indeed, evidence for such tonotopic maps in human primary auditory cortex has been reported using high-field fMRI (Formisano et al., 2003). The imaging data might show a significant interaction between tone frequency and location within the map (e.g., $C_1 > C_2$ in R_1; $C_1 < C_2$ in R_2), together with activation by tones versus silence (i.e., $C_1 > C_0$; $C_2 > C_0$) in both regions—that is, a cross-over interaction like that in Figure 5d. (This is basically an example of a quantitative dimension being mapped onto a spatial dimension in the brain, analogous to what Shallice, 1988, calls a "continuum of processing space".) Would one really want to claim functional dissociations between the two conditions? (See Rugg & Coles, 1995, for a related argument.)

This is where closer consideration of the psychological theories under test is required: what are the alternate and null hypotheses? One might argue that the same psychological function of "pitch determination" is operating under both conditions (i.e., there are not different processes for determining the pitch of high and low tones), and the null hypothesis has been falsely rejected. But one might also argue that this null hypothesis is framed at the wrong functional level, just as it would be inappropriate to state that both high and low tones engage "hearing", so one should see no difference in the brain. At a more detailed level of functional description, the null hypothesis would be that pitch is a monotonic function of the output of a single "pitch detector"—that is, one could imagine neurons whose firing got stronger and stronger as tone frequency increased, until the upper limit of human hearing was reached. This null hypothesis would appear to be rejected correctly: one could deduce that pitch is not a monotonic function of mean activity in a single set of neurons (which probably have nonmonotonic tuning curves instead) and favour the alternative hypothesis that pitch derives from a distributed pattern of activity over neurons (which, fortunately for the experimenter, happen to be spatially organized at a scale detectable by haemodynamic techniques). This issue of levels of functional abstraction is picked up later, when the function–structure mapping assumption is considered.

The "pure insertion" assumption

An assumption often made in imaging experiments is that of "pure insertion". This is the assumption that C_1 and C_2 differ only in the one function of interest, such that the subtraction of imaging data in one condition from those in the other reveals the neural correlates of that function. When C_1 and C_2 involve different tasks, this assumption is the same as that made by the subtractive method of Donders and addressed by Sternberg's (1969a, 2001) "additive factors". The pure insertion assumption that has since been levelled at imaging experiments has been used more generally to refer to any experimental conditions (e.g., two different stimulus types within the same task). In this sense, however, the same assumption

is made in just about every design in experimental psychology. While it may be true that the brain is a complex, nonlinear system in which pure insertion is unlikely (Friston et al., 1996), this has the same consequences for both imaging and behavioural data, since behaviour arises from the same complex system. Thus I do not think that pure insertion is a criticism specific to neuroimaging.

It is easy to imagine differences in brain activity that appear incidental to the tasks performed by the subjects. For example, if subjects find C_1 more "difficult" than C_2, one may observe differences in brain activity, which relate to increased anxiety of the subjects, that are incidental to the functional difference under investigation. This is a valid concern, and the answer, as mentioned earlier, is to design experiments in which anxiety is controlled, or manipulated factorially—ideally with enough sensitivity to reveal a main effect of anxiety in regions other than those showing effects of the factors of interest.[7] Importantly, the same differences in anxiety might, in principle, also confound behavioural measures of task performance. One reason this possibility may seem less obvious for behavioural than for imaging data may be an implicit bias that behavioural data somehow provide a more direct index of underlying functions than do haemodynamic data. On the contrary, my personal experience has been that the deliberation of possible confounds within imaging experiments forced me to think more carefully about the precise cognitive processes involved in tasks, even those as "vanilla" as yes–no recognition memory, than did deliberation of existing behavioural data. For example, fMRI data (Herron, Henson, & Rugg, 2004) prompted consideration of the relationship between the nature of the stimuli in recognition tests (old/new) and what subjects regard as the "targets", as is likely to be influenced by the ratio of old:new items and the precise task instructions (a consideration that has received much attention in the ERP literature, but not, to my knowledge, in the behavioural literature).

Note that imaging data can nonetheless include contributions from psychological processes that are known, on the basis of other experiments, not to influence behavioural data that are being acquired simultaneously. Examples include processes that operate subsequent to the behavioural response, but which still influence the haemodynamic response owing to its poor temporal resolution (e.g., explicit memory for a prime that arises only after unconscious facilitation of a speeded decision to the target; Henson & Rugg, 2002). I suspect that this is where the concern about pure insertion in neuroimaging really resides: in practice, one often has independent evidence that a particular behavioural measure in a particular task is not confounded by various other factors, such as anxiety. Yet this independent evidence will not necessarily apply to the imaging data. However, the same answer applies: one should then collect imaging data from the same experiments that addressed the confounds behaviourally. This relationship between imaging and behavioural data is considered further below.

The use of concurrent behavioural data

One often acquires behavioural data simultaneously with imaging data. Imagine that, in doing so, one finds significant RT differences, as well as haemodynamic differences, between

[7] Of course, no experiment can control for every possible confound; the goal is to control for those factors that are thought to be relevant on a priori theoretical grounds (Boring, 1963).

two conditions. A common worry in this situation is that the haemodynamic differences are "confounded" by the behavioural differences. This leads some neuroimagers to covary out the behavioural differences or to select subsets of trials in which behaviour is better balanced—or, in the worst case, to specifically design experiments in which behavioural performance is at ceiling, which is, of course, inadequate.

I think such worries are often unfounded. Within the framework adopted here, imaging data and behavioural data are both dependent variables. Both types of data are therefore *effects* of a hypothetical underlying functional difference, and so behavioural data cannot *cause* imaging data (Figure 6a). Thus there is no need, in principle, to covary out behavioural differences (though see below). Indeed, in doing so, one can remove any reliable

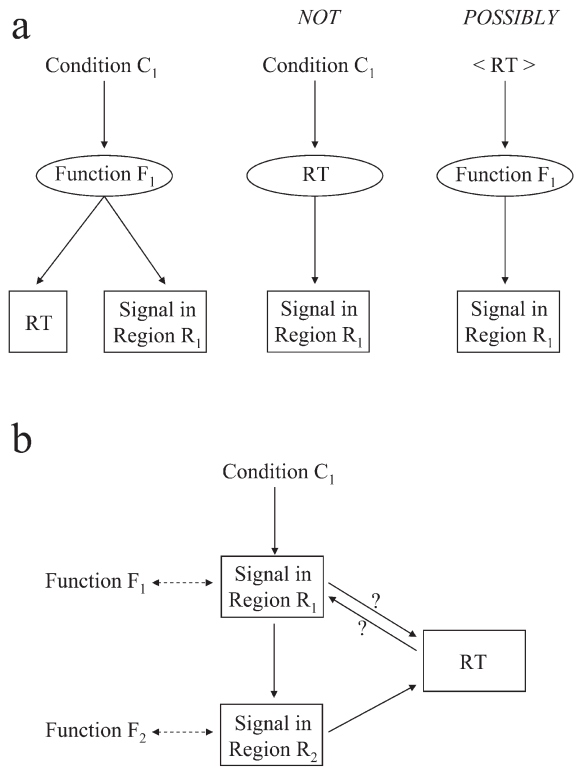

Figure 6. Relationship between imaging data and behavioural data. **(a)** Valid (left) and invalid (middle) conceptions within a causal interpretation of present framework (direction of causation is top-down); ellipses represent causes, boxes represent dependent measures. On the right is a valid alternative in which behavioural data are used only as a surrogate to index a hypothetical functional cause. **(b)** An alternative, more sophisticated model in which activity in some regions (region R_2, e.g., motor cortex) depends on activity in other regions (region R_1, e.g., visual cortex) and therefore possibly only indirectly on the experimental manipulation. In such models, the relationship between behavioural and imaging data might be tested more directly (e.g., by structural equation modelling: whether the connections indicated with a "?" are necessary). Note that, in these models, activity in a region might be equated directly with the imaging data, in which case one dependent variable, such as signal in R_2, can be treated as the "cause" of another, such as RTs (though regional activity is probably better treated as a hidden variable that is directly "connected" to the dependent imaging variable, allowing for example the possibility of independent measurement noise).

differences in the haemodynamic data too. For example, behavioural priming, when indexed by faster RTs for primed than for unprimed stimuli, and haemodynamic response decreases, when comparing primed and unprimed stimuli, may be different expressions of the same cause (e.g., facilitation of a perceptual process). Though the RTs may not covary perfectly with the haemodynamic changes in a "perceptual" region (owing to independent motoric components to the RTs, for example), introducing RTs as a confounding covariate in analysis of the imaging data is still likely to remove reliable haemodynamic differences by virtue of the variance that is shared. But this does not negate the role of that region in the perceptual process that has been facilitated, because the facilitation itself has been effectively discounted!

Nonetheless, the situation is a little more complicated. There are cases where behavioural data can be used sensibly to define covariates in the analysis of imaging data (confounding or otherwise); these are situations where behavioural data are used as surrogates for the true functional "causes". This is illustrated by the Remember/Know experiment by Henson et al. (1999) described earlier. Because the experimenters had no control over which stimuli would be recollected and which would just seem familiar, the experimental conditions had to be defined by behaviour (i.e., which key was pressed). In other words, the behavioural data, though still not strictly the cause of the imaging data, were used to index underlying psychological processes (Figure 6a). In this role, one of the dependent variables is treated instead as an independent variable in the statistical model of the imaging data. In the study of Henson et al. (1999), it was these differences that were of interest. In other situations, they may be potential confounds, such as error trials. The difference between this last case, when behavioural data are used to define a confounding variable, and the above example of priming is that now an explicit assumption is made that the behavioural data index a distinct functional cause.

I am not aware of a general principle that determines whether the inclusion of behaviourally defined confounds in the analysis of imaging data is correct or incorrect. (It is really a question of "model selection".) Note, though, that I have been implicitly talking about conventional, mass univariate analyses in which the same statistical model is tested at each voxel. In these models, an experimental manipulation (the independent variable) can potentially "influence" every region directly. In more sophisticated models—such as structural equation models—one might constrain the "causality" by modelling explicit connections between regions, such that some regions are only influenced by experimental manipulations indirectly via the activity of other regions (Figure 6b). Thus the visual stimulation of a visuomotor task might be modelled as influencing one region (e.g., in visual cortex) but might only influence another region (e.g., in motor cortex) via its effect on the visual region. In such models, the relationship between imaging data and behavioural data can be tested more explicitly by treating the behavioural data as another "region" (node) in the model (e.g., connected from the motor cortex region; see Figure 6b legend). Such models are, however, outside the present framework. My point here is simply to warn against the automatic assumption, often owing to an implicit precedence given to behavioural data, that concurrent behavioural differences confound interpretations of imaging data.

Note, finally, that simultaneous acquisition of behavioural data is generally advisable to confirm that the task is being performed in the scanner in the manner that was hypothesized

(i.e., to ensure the psychological validity of the experimental manipulations). This does not mean that concurrent behavioural data are always required. One might be interested in functional differences between trial types for which concurrent behavioural measurement would interfere with those functions (e.g., requiring a differential response to two trial types may confer on them a different status, such as target versus nontarget; Rugg, 1995). Indeed, imaging data may reveal a greater range of functional differences than is possible with behavioural data, since behavioural measures capture only the final outcome of task-related processes. (For excellent further discussion of the relationship between behavioural and imaging data, see Wilkinson & Halligan, 2004.)

The systematic function–structure mapping assumption

Here we come to the crux of the argument. At some level, different parts of the normal, healthy brain subserve different functions. I expect that few people would dispute this statement—for example, that different parts of the brain are involved in the initial processing of visual versus auditory input (in the occipital and superior temporal lobes, respectively). Within the occipital lobes, there is also good evidence that different parts are specialized for processing colour (V4) and motion (V5). So is the question of whether there is a mapping between function and structure simply a question of scale? And if the appropriate scale is identified, do we really believe that all we need to do is to find the *N* distinct functions associated with the *N* distinct structures?

To answer these questions, it is useful to consider the function–structure mapping more formally: the mapping could be one-to-one, one-to-many, or many-to-one (see ahead to Figure 8). For the purpose of function-to-structure deduction, the type of mapping would not appear to matter, because it is an assumption of the experimental design, according to theory T_1, that the experimental conditions being contrasted differ only in the single function of interest (see earlier comments on "pure insertion"). Nor does it matter if there are "many" regions differentially activated or if there are "many" other functions subserved by those regions in other contexts. The only assumption is that, within the current experimental context, it cannot be the case that some regions are associated with a function in one condition, but other regions are associated with the same function in the other condition. This is the weak sense of "systematicity".

For the purpose of structure-to-function induction on the other hand, one might argue that a one-to-one mapping—a stronger sense of "systematicity"—is necessary. Otherwise, if a region can be associated with different functions, one cannot infer from its activation in one experiment that the same function occurred in another experiment, since there is no explicit control of potential confounds across experiments. But what if the unit of mapping is not a single region, but a set of regions (or "network", in the loose sense)? Thus, a particular set of regions may be associated with a single function, even if each region within that set may be associated with many different functions when part of different sets (Figures 7a, 7b). In other words, the function of a given region may depend on its co-activation with one or more other regions. This claim, at least, has the power of combinatorics in the number of possible functions that might arise from *N* regions.

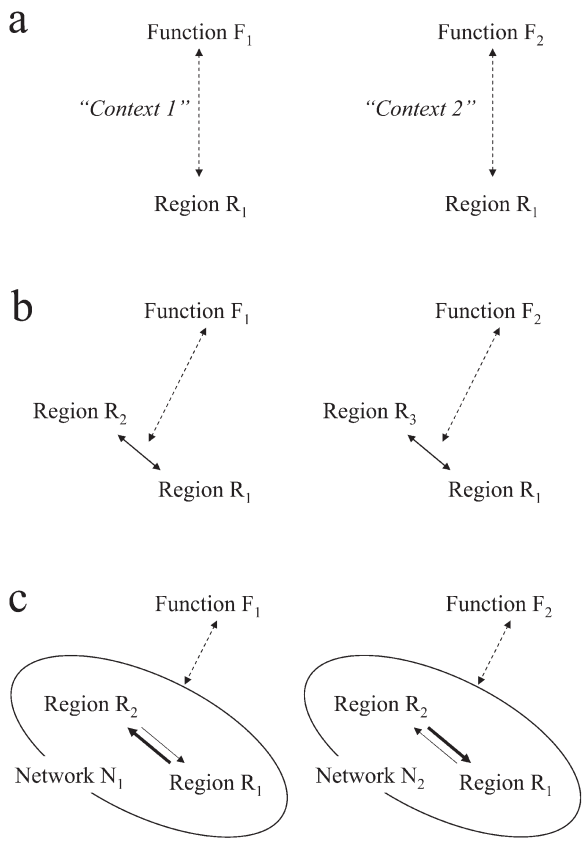

Figure 7. Networks. **(a)** The function of a region may depend on the experimental "context". But what are these "contexts" any more than differences in activity elsewhere in the brain? **(b)** Thus, functions may map to coactivation of sets of regions, such that a given region (R_1 here) may participate in different functions when part of different sets (i.e., regions R_2 and R_3 can be thought of the "neural context" that affects the function of R_1). **(c)** In fact, functions may map to sets of regions that are interacting in specific ways, or "networks". (Thickness of arrows indicates strength of effective connectivity; see text.)

There is also a stronger sense of a "network", in which the activities of a number of regions correlate over time (or replications) within a condition. This is called "functional connectivity". When the degree of covariation between regions—or, more precisely, the slope of their regression—changes reliably between conditions, it becomes an example of "effective connectivity" (Friston, 2002a). For example, attending to motion in a visual stimulus has been shown to increase the effective connectivity between V2 and V5 (Friston & Buchel, 2000). In principle, such changes in effective connectivity can occur in the absence of changes in the mean activity of the regions—that is, the same regions would not be seen in a conventional subtraction analysis like those described above. Several methods have been introduced to test effective connectivity in such networks, such as structural equation modelling (Buchel & Friston, 1997) and nonlinear dynamical systems (Friston, Harrison, & Penny, 2003). Such

methods allow formal tests of the idea of "neural context" (McIntosh, 1999): that the function of a region depends on its interaction ("hot-wiring") with other regions.[8] In the following discussion, the structural units of the function–structure mapping are therefore generalized to sets of brain regions interacting in specific ways (i.e., distinct "networks"; Figure 7c). However, there is a sense in which this redefinition of the "structural units"—and its associated power of combinatorics—simply dodges the question of whether there is a one-to-one function–structure mapping: the question now becomes whether there exists a one-to-one mapping between a psychological function and a specific network of interacting regions.

Note that there is a more subtle sense of a one-to-many function–structure mapping; the sense of "degeneracy" (Noppeney, Friston, & Price, in press). An example of degeneracy would be two networks that can perform exactly the same function, but which need not perform that function at the same time (other examples in biology include different neucleotide triplets coding for the same protein). At any given time, one network might be latent—suppressed by the active network, for example; this notion is thus distinct from "redundancy", in which two or more degenerate networks happen to be active at the same time. Noppeney et al. (in press) distinguish degenerate functional anatomies within a brain from degenerate functional anatomies across brains (i.e., across individuals). The latter is an important possibility that is discussed in the later section on equipotentiality and plasticity. The former possibility deserves consideration, since it potentially violates the one-to-one mapping needed for structure-to-function induction. While degeneracy is always a possibility, and might be advantageous in both ontogenetic and phylogenetic terms, it is unclear what determines whether a particular degenerate network will be active on a given occasion since redundant activation of all of them would make little sense from a metabolic perspective. If the selection of one of a set of degenerate networks were essentially random from trial to trial or from experiment to experiment, one would not observe reliable or reproducible activations. If it differed from context to context, then the problem simply becomes specifying the nature of the differences in the context (i.e., refinement of the precise function). Thus, though the concept of degeneracy cannot be ruled out as a reason for failing to find reproducible activations, and though it may be vital when relating imaging data to neuropsychological data (i.e., from patients in whom one or more degenerate networks are damaged; Noppeney et al., in press), it is not clear that it affects the type of inferences based on imaging data that are under consideration here.

A more fundamental issue concerns the level of "structure" and the level of "function". Most psychologists would appear happy with the proposal that there is a hierarchy of abstraction of functions, with any one function (such as "visual perception") being divisible into sub-functions (such as "colour perception", "form perception", and "motion perception").

[8] This possibility raises problems when interpreting data from only a small number of regions (e.g., in many single-cell recording studies). The ability of imaging methods like fMRI to measure haemodynamic changes over the whole brain effectively simultaneously (given their slow dynamics) is an advantage in this respect. Note, however, that there may be changes in the interactions between neurons in different regions at a millisecond timescale that would be invisible to haemodynamic techniques.

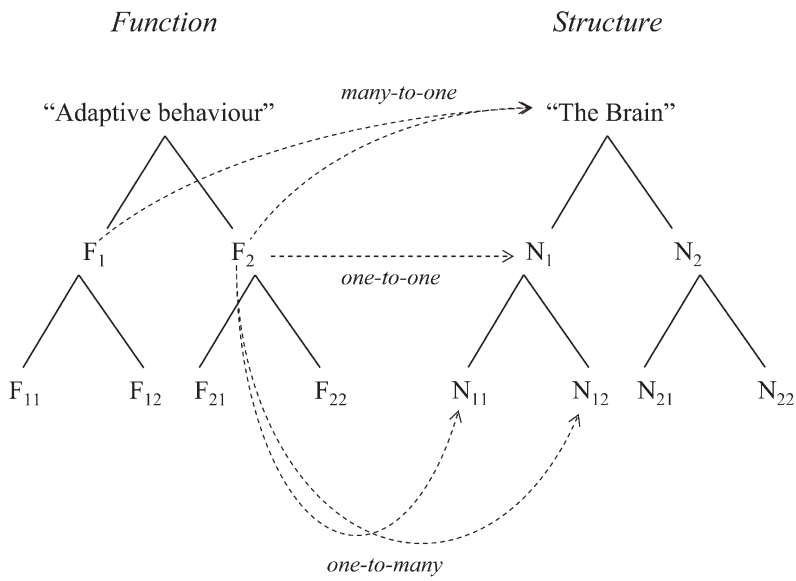

Figure 8. Functional and structural hierarchies. The nature of the function–structure mapping—one-to-one, many-to-one (pluripotential), or one-to-many (degenerate)—may depend on the relative levels within the two hierarchies. F refers to a function; N refers to a network (a set of brain regions interacting in a certain way; see text), which may reduce to a single brain region.

Most anatomists classify brain structures hierarchically too. At the top of these two hierarchies might sit "adaptive behaviour" and "the brain", respectively; at a much lower level might sit "detect incident photons" and "retinal cone" (though the latter may not be viewed as a "psychological" level of function, and clearly extant neuroimaging techniques can only resolve down to a certain level of the anatomical hierarchy). Such hierarchies have been called "ontologies" (Price & Friston, in press), in analogy with functional genomics (Lan, Montelione, & Gerstein, 2003). The question of whether there is a one-to-one, one-to-many, or many-to-one mapping then depends on which levels of the two hierarchies—functional and structural—are being compared. One-to-one mappings may be restricted to comparisons at the appropriate level (or, more accurately, the "appropriate" level can be defined as that in which there is a one-to-one correspondence). The mapping of a function at a "higher" level of abstraction is likely to be one-to-many; the mapping of a function at a lower level of abstraction is likely to be many-to-one (Figure 8).

But rather than dodging the question of a one-to-one mapping, this possibility would appear to beg the question. If one does not know the functional hierarchy of the mind (even if one does know the anatomical hierarchy of the brain), how can one use activations within the anatomical hierarchy to determine the functional hierarchy? According to the argument above, there needs to be one-to-one mapping before one can infer from structure to function—that is, for structure-to-function induction. Yet in order to know whether there is a one-to-one mapping, one needs to know the appropriate level of the functional hierarchy—in other words, to already know the relevant functions. This is, of course,

a chicken-and-egg situation. This is why I said earlier that a systematic function–structure mapping, in the strong sense of systematicity—that is, one-to-one mapping—may be unprovable in a logical sense.

For example, take the earlier question of tonotopic maps in auditory cortex. Imagine that, in Figure 8, function F_2 corresponds to "pitch determination"; functions F_{21} and F_{22} to "detect high-frequency" and "detect low-frequency" sounds, respectively; network N_1—here a single region—to "caudal Heschl's gyrus"; and networks N_{11} and N_{12} to "caudomedial" and "rostrolateral" portions of caudal Heschl's gyrus, respectively (Formisano et al., 2003). To infer from activation of the caudomedial portion (N_{11}) that high-frequency sounds are being perceived requires a one-to-one function–structure mapping. To know that there is such a one-to-one mapping, one needs to know that the functions and structures equated are at the appropriate levels of the functional and structural hierarchies. But to know whether they are at the appropriate level (and here the circularity is completed), one needs to already know the relevant parts of the hierarchies: if one observed activation of the whole of caudal Heschl's gyrus (i.e., was at the structural level of N_1), one could not infer perception of high-frequency sounds, because there is a many-to-one function–structure mapping. Conversely, if one's psychological theory did not distinguish detection of high- and low-frequency sounds (i.e., was at the functional level of F_2), one would not observe consistent activation within caudal Heschl's gyrus when stimulating with tones that happened to be of different frequencies, because there is a one-to-many function–structure mapping.

Nonetheless, such "circularity" of logic is not, in fact, uncommon in science. One could argue that the "transparency" assumption in cognitive neuropsychology (Caramazza, 1986), "renormalization" in theoretical physics, or even the materialist view of the mind—that is, rejection of dualism—are comparable types of assumption, which are necessary to proceed but cannot be proved on independent grounds. They might be viewed as working hypotheses under which a particular scientific paradigm operates (Kuhn, 1970). The practical response is, of course, to "bootstrap": to assume that such working hypotheses are true and then see how much progress is made. The hope is that, if they are true, interpretations of the resulting empirical findings will converge on a "consistent" body of theorizing (in the present case, that a consistent functional hierarchy will emerge). Conversely, if the working hypotheses are false, no consistency will emerge, and no progress will become apparent. Put simply, the assumption—and paradigm—stands or falls by its success; put flippantly, the proof of the pudding is in the eating.

One might not regard this as a very satisfactory answer. In particular, it obviously raises the question of whether functional neuroimaging has produced a consistent body of theorizing. Answering this question is just as difficult, and one might claim that the field is still too young to tell. But I wonder whether closer inspection of other fields, such as experimental psychology, would reveal similar working hypotheses and also raise difficulties in answering claims that the field is converging on a consistent body of theory?

Note that both types of inference—function-to-structure deduction and structure-to-function induction—would appear important for scientific progress. Though a single imaging experiment may be sufficient to favour one psychological theory over another, using only function-to-structure deduction, in practice one almost invariably wants to extend the imaging findings to slightly different conditions in further experiments, which soon requires

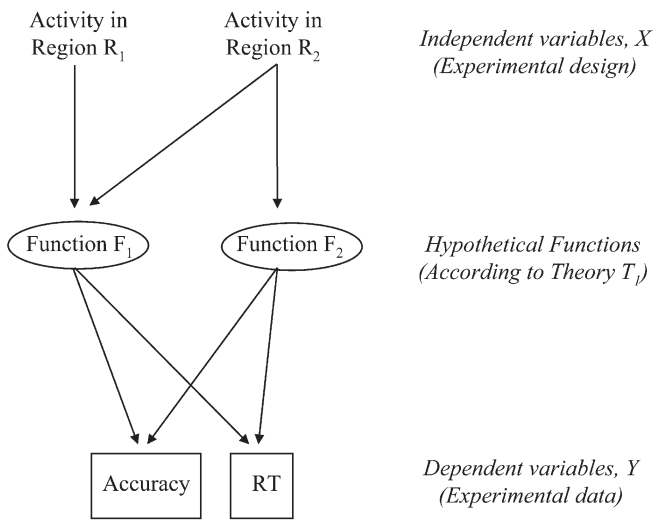

Figure 9. Characterization of methods that group or manipulate regional activity and measure corresponding changes in behaviour—such as neuropsychology, electrical stimulation, and transcranial magnetic stimulation—which offer an important complementary approach to neuroimaging (cf. Figure 1b).

structure-to-function induction, with the extent to which the "same-region/same-function" assumption is "stretched" depending on the extent of differences in the experimental details.[9] In this case, one could argue that, for functional neuroimaging as a whole to make progress, it needs to assume a one-to-one function–structure mapping.

In summary, I have argued that inferences of the "function–structure deduction" type do not require strong assumptions of systematicity: that is, one can learn from imaging data without making the "localizationist" assumption of a simple, invariant mapping between function and structure. For inferences of the "structure–function induction" type, however, I believe that one must assume a one-to-one mapping between function and structure, in order to discount the possibility that the same structure implements different functions across the experimental contexts over which the inference is made. This assumption may not be provable (at least without other types of data, Sarter, Bernstein, & Cacioppo, 1996; see Figure 9), but the same would seem to be true of working hypotheses in other scientific

[9] The role of these two types of inference might be illuminated by considering the history of cognitive ERP research. Early studies identified obvious visible components of the ERP (e.g., "P1", "N400", "P3") and tried to attach a specific function to each. This proved a difficult task, and the functions soon became quite abstract (e.g., "context updating", Donchin & Coles, 1988). One reaction to this was to restrict interpretation of ERP effects to certain psychological domains (e.g., recognition memory). In this case, it mattered less which particular ERP components differed, but simply that a difference was found between two conditions (that was either predicted or not predicted by a theory within that domain). This, of course, mirrors the more restricted type of function-to-structure deduction inference discussed here. However, in order to say whether the same ERP effect replicated in another experiment, a further inference analogous to structure-to-function induction was required. One then either moves back towards labelling specific effects/components with specific functions or sticks with qualifying each function as restricted to a certain experimental domain.

fields, and this is not necessarily an impediment to progress. Finally, when appealing to a one-to-one function–structure mapping, it is important to consider different levels of functional and structural abstraction in relation to alternative many-to-one or one-to-many mappings, and it may be important to allow the structural unit of the mapping to be a network of brain regions interacting in certain ways, rather than a unique brain region.

Common criticisms

I have encountered a number of common criticisms of neuroimaging, organized by the "soundbites" below. Many of these relate to issues discussed above; some are similar criticisms but from a different starting point. Nonetheless, it may be useful to consider each of them in greater detail.

"Imaging is just glorified phrenology"

This criticism compares neuroimaging to the once popular, but now largely ignored, field of phrenology. There may be several definitions of this term, but I refer here to the specific approach of using the shapes of people's brains to infer something about their mental faculties, as exemplified by the work of Gall [1822–1825] (see Marshall, 1984). This is a type of structure-to-function induction, and the field faced the same criticisms of "localizationism" as function neuroimaging. However, there are (at least) three important differences between this historical definition of phrenology and modern-day neuroimaging.

The first concerns the "structures" under investigation. The structure for phrenologists concerned parts of the cranium, and the dependent variable was their shape or size, with the shape of the cranium being intended as an indirect measure of the shape of the underlying cortex without the skull having to be opened. The relevance of cortical shape to psychological function is unclear, however (indeed, the two may be unrelated). The relevance of the haemodynamic dependent variable used by neuroimagers, though not fully established, is certainly much better established. The blood flow and BOLD signals correlate with other neurophysiological measures of neural/synaptic activity (when recorded simultaneously, for example, in the nonhuman primate; Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001), and much work has correlated the same neural/synaptic activity with behaviour, particularly in the nonhuman primate.

The second difference between neuroimaging and classical phrenology concerns the functions under investigation. The "faculties" identified with cortical regions by the phrenologists (based on information content in Gall's view—e.g., "number", "colour", "music"—or pretty much on speculation in Fowler's popularization of phrenology—e.g., "hope", "mirthfulness", "love of family") were not particularly well founded (on behavioural experiments, for instance). To the extent that one believes that experimental psychology has made progress during the intervening centuries, the functions considered by neuroimagers are based on a much larger body of theory and data. One of the main arguments made by Uttal (2001), who associates neuroimaging with neophrenology, is that we have not, in fact, made much progress in the fractionation of cognitive processes. This is, of course, an open question (related to the earlier question of how one defines the "success" of a field) and one on which, I expect, most experimental psychologists would disagree with Uttal (see, e.g., Hubbard, 2003).

The third and fundamental difference from phrenology is that functional imaging can be interventional rather than simply correlational: that is, experiments can be designed in which a factor is manipulated, perturbing the system under investigation, and the consequences, such as changes in brain activity, observed. This is not true of phrenology, in which the dependent variable is fixed (for an individual) and one can only correlate differences in function with differences in skull shape across groups of individuals. This opportunity for causal intervention is, in fact, the main reason that I think functional imaging is more likely to succeed where phrenology failed.

"Imaging data are only correlational"

This comment refers to a different sense of correlational from that mentioned above. It is often made when comparing imaging data with neuropsychological data, relating, for example, to claims that imaging data indicate brain regions *sufficient* for a given task, but not *necessary* for that task (Price, Warburton, Moore, Frackowiak, & Friston, 1999). Rather, it is the disruption of a brain region—for example, following a lesion—that is claimed to demonstrate that the region is necessary for a task (Figure 9). For example, imaging data might show activity in medial temporal regions that correlates with delay conditioning, even though lesions to those regions do not impair typical measures of such conditioning (unlike lesions to the cerebellum, which do cause impairments; Daum, Schuygens, Ackermann, Lutzenberger, & Birbaumer, 1993). Thus one can conclude that medial temporal activity is not necessary for this specific task (and measure) of delay conditioning. This type of conclusion is important when comparing imaging and neuropsychological data from the same tasks (Price et al., 1999). However, this example does not indicate an unreliable or spurious nature of imaging data, as might be suggested by "correlational", because this task is unlikely to engage a single type of learning in normal subjects—that is, it is unlikely to be "process-pure". Rather, it is likely to engage parallel learning of both the declarative and non-declarative type, in medial temporal and cerebellar regions, respectively (Cohen & Squire, 1980). In other words, subjects are likely to become aware of the contingency between a conditioned stimulus (e.g., tone) and an unconditioned stimulus (e.g., airpuff), independently of the (unconscious) learning of an unconditioned response (e.g., eyeblink). Medial temporal activity may still be necessary for such declarative learning in this task; it is simply that the typical measure of task performance (eyeblinks) does not access this type of learning (though other measures might). What are needed, of course, are other types of manipulation (e.g., performance of a concurrent distracter task) that dissociate the two types of learning within the task (see earlier section on pure insertion). Thus it is important to distinguish the use of "correlational" to refer to performance of a specific task from its use above to refer to the nature of the independent variable (experimental intervention in imaging versus post-hoc classification in phrenology). In its latter use, imaging data from an experimental manipulation of delay conditioning are no more "correlational" than are behavioural data from that manipulation.

"Algorithm is independent of implementation"

This claim is based on the views of Marr (1982), and the strong interpretation of these views within the fields of Artificial Intelligence (AI) and Cognitive Science. Marr (1982)

claimed that the brain—or any complex system—can be studied at three levels, and the questions posed at one level are separable from those posed at another. Psychological theories are normally phrased at the "algorithmic level", in terms of the representations and processes used to achieve a particular goal, whereas neuroimaging might be regarded as a method of studying the "implementational level", viz., how these algorithms happen to be realized in the brain (following years of evolution). The related AI view is that the same algorithm (software) can be realized by any number of implementations (hardware, such as transistors, hydraulics, neurons, etc.). Conversely, if the neural architecture of the brain is sufficient to constitute a Universal Turing Machine, then it can, in principle, implement any algorithm. Another related view in Cognitive Science is the "functionalist" perspective: that psychological functions are defined purely by their inputs and outputs, within the context of the system as a whole, regardless of the mechanism that transforms the inputs into outputs (i.e., are "black boxes"). These views lead many psychologists to think that study of the neural architecture of the brain cannot be informative about psychological function. Hence location in the brain is irrelevant (a view called the "ultra-cognitive" stance in neuropsychology; Shallice, 1988).

For example, both Harley (2004b) and Coltheart (2004) question the value of imaging data because they believe knowledge of hardware tells one nothing about the software that runs upon that hardware. First, I think this is misleading. Harley and Coltheart appear to equate imaging data with knowledge about hardware, yet functional imaging provides more than information about hardware: it provides "performance" information about the spatial distribution of processes occurring while "software" is running. (In the example of a chocolate factory used by Coltheart, 2002, the analogy would be to measure simultaneously the temperature of different components within the factory while certain inputs are changed and outputs observed, rather than simply scrutinizing the structure of each component separately.)[10]

Second, despite the apparent truth of the AI view—the "in principle" independence of software and hardware—once one also has performance (e.g., behavioural) data, I think knowledge of hardware can tell you something about software. For example, one might postulate a psychological theory that face recognition involves serial comparison of (some abstraction of) the visual input with a template representation of every face one knows. Yet

[10] Indeed, it is ironic that Harley (2004a) would appear to favour connectionist models in cognitive psychology—models that are at least loosely inspired by the architecture of the brain. Nonetheless, while I agree with Harley that psychological theories benefit from being formalized in computational models, I do not think such models need to be connectionist models, in the sense that the components of the models do not need to be related to neural concepts like synapses (any more than they need to be related to specific parts of the brain). If intended as purely psychological models, this would correspond to a conflation of explanatory concepts at different levels (as Marr would caution). However, even if the psychological-level computational models make no predictions about specific brain regions, they can still be tested by neural-level data—namely, by finding dissociations between brain activity under conditions that would, or would not, be expected to differ in the model (as elaborated above). For example, while Coltheart (2002) is right to question whether a failure to find a difference in brain activity between reading exception words and reading nonwords would falsify a dual-route model (not least because it is based on a null result, to which imagers should be just as sensitive as psychologists), it is puzzling that he does not consider the opposite, positive finding of dissociable regional activations for exception words and nonwords. I would view this as evidence against single-route connectionist models of word reading.

consideration of the minimal time for information processing by neurons (e.g., tens of milliseconds for a rate code) would suggest that such a serial algorithm is incompatible with the typical number of known faces (of order hundreds or thousands) and the typical time taken to recognize a face (e.g., hundreds of milliseconds).

Though a trivial example, which raises obvious solutions in terms of parallel neural processing, this illustrates how one of Marr's levels can influence another. Indeed, Marr believed that his levels were, in fact, "logically and causally related"; his point was that "explication of each level involves issues that are rather independent" (p. 25). In other words, though psychological theories can be *phrased* purely in algorithmic terms—representations, processes, mappings, transforms—without reference to brain regions or neurons, this does not mean that they are not ultimately constrained by such implementational details. Theories at each level are not independent in the face of empirical evidence, such as that from behavioural or imaging data. A similar point can be made in relation to the field of genetics. Though one might study Mendelian genetics (at the algorithmic level) without knowledge of DNA (the implementational level), there is no doubt that the discovery of DNA and mapping of the human genome have proved vital to our current understanding of genetics.

"Functional imaging only tells you 'where', not 'how'"

I think this criticism is misguided. Data do not tell us directly "how" a task is performed (imaging data any more or less than behavioural data). It is theories that describe "how" a task is performed. Data are used to distinguish between different theories. Therefore, as argued above, imaging data can tell us "how" a task is performed by virtue of supporting or contesting a theory.

"The cortex is equipotential; the cortex is plastic"

This comment does raise important issues. One of the strong reactions to localization of function was the view that the cortex is equipotential in some functional capacities (Lashley, 1929/1963). According to this view, it is more likely to be the size of a brain lesion, rather than the particular location of that lesion, that correlates with a behavioural deficit. Even if this view is too extreme, and there is a default organization of function, the brain is certainly highly plastic, in that the recovery of function that sometimes occurs following brain damage suggests that some brain regions have the potential to "take up" the function of other damaged regions—or there was prior degeneracy of function in such cases.

I do not wish to review the evidence concerning the degree of plasticity/degeneracy in the brain, but I do want to consider its implications. Foremost, it raises the problem of a change in the function–structure mapping over time. For example, the particular structure responsible for a particular function may change during our development. This would prevent generalization of neuroimaging findings across different developmental stages (though the functions themselves are also likely to change during development). Moreover, different developmental trajectories (based, for example, on the order of exposure to different stimuli, such as language) may result in different "final" function–structure mappings. This does invalidate some of the possible psychological inferences from imaging data described above. I think the only solution is the nomothetic solution: to define "normal"

psychological functions—in the healthy adult, for example—and to assume that there is single "normal" mapping of these functions to the brain. This then becomes an empirical question: if the function–structure mapping changed so rapidly over time, or so dramatically as a function of different developmental trajectories, one would not observe reliable differences in imaging data when averaging over random samples of normal individuals. The fact that neuroimaging can produce reliable and reproducible differences suggests that there is a normal (default) function–structure mapping. Nonetheless, one should bear in mind that partial variability in this mapping might be one cause for a failure to find reliable differences.[11]

A related concern is a change in the function–structure mapping on a faster timescale, such as learning within an experiment. In discussing an "adaptive coding model" of prefrontal cortex, for example, Duncan (2001) describes "flexible" neurons, in parts of prefrontal cortex of the nonhuman primate, that, following training, come to represent particular abstract rules (e.g., "match/nonmatch") or novel categorical distinctions (e.g., "dogs/cats"). Moreover, the responses of these neurons can be remapped when the animals are retrained on a different categorization (Freedman, Riesenhuber, Poggio, & Miller, 2001). This flexibility makes it difficult to define the functional role of (these parts of) prefrontal cortex in general, context-independent terms. An extrapolation of this flexibility is the AI concept of a "virtual machine": if the prefrontal cortex is powerful enough to dynamically "rewire" itself in order to perform a range of different tasks, then one is unlikely to observe dissociations across those tasks. That is, the particular neurons that implement a function in one task may be somewhat arbitrary (i.e., different from those in the prefrontal cortex of another individual, or different from those if the task is forgotten and then relearned). However, this possibility also becomes an empirical question. If the same regions of prefrontal cortex are consistently activated across a range of different tasks (e.g., Duncan & Owen, 2000), then the function of those regions needs to be phrased at a reasonably abstract level in a functional hierarchy (see earlier). Moreover, if one does find dissociations between two regions within prefrontal cortex across two conditions of a task, even though the absolute location of these regions may be arbitrary, depending on the particular virtual machine running, the presence of the dissociation can still say something about how that particular task was solved after training—for example, that cats and dogs were distinguished by a single feature or by a value along some more abstract configural dimension.

"Name three psychological insights from neuroimaging"

This is a devious request. Before attempting to meet it, I note that almost any example will be controversial, just as there is unlikely to be unanimous agreement with any example of what we have learned from experimental psychology. For example, Coltheart (2002) says

[11] There are approaches in functional imaging that allow for variability in the precise localization of functions across individuals. For example, Kanwisher, McDermott, and Chun (1997) define the "fusiform face area" (FFA) separately for each individual (by the maximal activation when contrasting faces against various nonface stimuli), which may be in differing locations in anatomical terms. This contrasts with the nomothetic approach adopted here, in which it is assumed that there is single functional anatomy for all (normal) individuals and, furthermore, that one can match or "normalize" anatomy across brains (an important issue that I do not consider here). The results then stand or fall by this assumption. Relaxing this assumption, as Kanwisher does, may be appropriate to some extent, but it can lead to difficulties (e.g., if an individual's maximal face activation is in prefrontal cortex, is it still the FFA?).

that he cannot think of any cognitive model that has been "refuted on the basis of imaging data" (p. 22). Likewise, I am not aware of any behavioural data, based on healthy or brain-damaged individuals, that has refuted single- (or dual-) process models of word reading: the debate between such models would still appear very much alive! Another example might be the debate between "early" and "late" theories of attentional selection (Allport, 1993; though see Lavie, 1995, for a unified theory, and Rees, Frith, & Lavie, 1997, for an imaging study that used structure-to-function induction to test this theory). The reason is that Popperian "falsification bullets" probably do not represent a good description of science. I believe a better description is competition between theories, in which "natural selection" occurs on the basis of the size and convergence of different types of evidence. In this sense, imaging data increase the scope of the evidential base.

The answer to the above request also depends on what is meant by "insights". If this refers to tests of existing psychological theories, then I would argue that there are many examples in the imaging literature of such function-to-structure deduction inferences. I discussed some examples above in the context of our own work (e.g., dual- versus single-process models of recognition memory). Other examples include tests of inattention blindness versus inattentional amnesia (Rees, Russell, Frith, & Driver, 1999), dependent versus independent routes for processing facial identity and facial expression (Winston, Henson, Fine-Goulden, & Dolan, 2004), verbal versus visuospatial slave systems in working memory (Smith & Jonides, 1997), and "simulation theory" versus "theory theory" of representing others' intentions (Ramnani & Miall, 2004). Some examples of structure-to-function induction include the use of activity in right dorsolateral prefrontal cortex as a signature of "prediction error" during learning, which can be used to distinguish various learning theories (Corlett et al., 2004), and the use of the level of activity in V5 to indicate the degree of processing of unattended moving stimuli (Rees, Frith, & Lavie, 1997).

I expect what is probably meant by "insights", however, is new psychological theories that have emerged from imaging data. Importantly, as discussed above, these must be theories that make no reference to anatomical terms (Page, 2004). There are probably fewer of these. A weak answer is that it is still "early days" for the functional imaging field, which only became established in the 1990s and needs to "catch up" with current levels of psychological theorizing. However, there are certainly psychological domains in which neuroimaging experiments have spurred the development of a number of psychological constructs, if not fully fledged theories, even if these are still open to debate. The most striking of these concerns distinctions between different types of executive function, which may or may not (Duncan & Owen, 2000) be subserved by different parts of prefrontal cortex. These include maintenance, manipulation, monitoring, refreshing, selection, integration, and branching (see, for example, Fletcher & Henson, 2001). This has also included the development of new tasks that might be usefully extended for further behavioural and neuropsychological testing, such as the "N-back" task (Smith & Jonides, 1997). Other neuroimagers may have clearer examples of new theories in other domains, with which I am less familiar.

There are also examples of findings from neuroimaging that, though they have not yet led to new theories, have certainly highlighted an absence of existing psychological theorizing. In the earlier example of memory encoding, I discussed differences in the haemodynamic response to individual items according to whether or not they are later remembered. Another fMRI study (Otten et al., 2002) identified more sustained differences, independent

of responses to individual items, that also correlate with subsequent memory. While Tulving has theorized about the importance of such "state effects" during retrieval (viz., the concept of a "retrieval mode"; Tulving, 1983), I am not aware of any existing psychological theory that addresses state effects at encoding (indeed, such state effects have been difficult to test behaviourally, Rugg & Wilding, 2000).

"Psychologists are only interested in behavioural data, by definition"

This is a tricky one, since I cannot argue with how one chooses to define a scientific discipline. If this definition holds, confining the experimental psychologist to, for example, accuracy and RTs, then I accept that imaging data are irrelevant. Indeed, one might use this question of relevant data to distinguish "(cognitive) psychology" from "cognitive neuroscience". My response, however, is to note that this is a very narrow definition of psychology, which poses both a practical and a theoretical problem. The practical problem is that I know many researchers who regard themselves as experimental psychologists, and who are happy testing their theories using some form of biological data, such as galvanic skin response or pupil diameter. Are we really to retract their professional label? The theoretical problem is that I would imagine that the main criterion that binds experimental psychologists together is not, in fact, the choice of relevant data, but the common goal of understanding how the mind works, and that most share the materialist view that the mind derives, somehow, from the brain. If so, I would think it unwise to discount any type of data other than behavioural data, which, after all, simply reflect the final output of a series of brain processes—a subset of variables that happen to be easily measurable via key presses or tape recorder.

Neuroimaging in practice

Despite the above arguments for the conditions under which imaging data are informative to the experimental psychologist, these conditions would rarely be met from a cursory glance at the functional imaging literature. It is unfortunate that most discussion sections in functional imaging papers simply list a number of post-hoc interpretations of the findings. These are usually types of structure-to-function induction (e.g., "Region R_1 in the present data was also activated in a previous study in a slightly different context, so may reflect function F_1; region R_2 was also activated in a previous study, so may reflect function F_2 . . ."; etc.). There are fewer examples of explicit tests of a priori hypotheses (e.g., function-to-structure deduction). This may partly reflect the history of the field, in which many of the early imaging studies tended to neglect the psychological literature and were often poorly designed. Some were simply "suck-it-and-see" experiments (e.g., "Can we find a difference in brain activation between novice and expert jugglers?"). I think that a new wave of papers has since emerged, as more and more imaging experiments are conceived and designed by psychologists.

There are also many sources of statistical errors in functional imaging papers. Examples include inferences based on differences in statistical maps (e.g., comparing a map of C_1 versus C_0 with a map of C_2 versus C_0), rather than on statistical maps of the difference

(i.e., C_1 versus C_2; see Note 6), or inferences following numerous post-hoc contrasts of different conditions, without correcting for the number of such contrasts. This partly reflects the fact that the datasets are large and the methods of analysis complex and constantly evolving, in tandem with improvements in the rapidly changing technology. Statistical packages like SPM are a double-edged sword: they greatly facilitate access to functional imaging data for the novice, but they can also be misused if their principles are not fully understood. Furthermore, few imaging studies—including my own—would meet the strict criteria put forward earlier regarding statistical tests of interactions across brain regions. I hope that this will change once imagers think more carefully about the nature of the inferences they are trying to make.

This potential for statistical error may be one reason that the functional imaging field also appears peppered with contradictory results. Another reason is that there are few attempts at replication. The imaging literature would certainly benefit from a convention (like that in the *Journal of Experimental Psychology*) that several experiments on a related topic are typically required for a paper (though this lack of replication may reflect the costs of imaging experiments, I would not regard it is as an excuse for relaxing scientific rigour). I do not think that the apparent contradictions in the literature reflect any intrinsic unreliability of imaging data, however. My colleagues and I are fortunate to have performed numerous experiments using variants of the basic yes/no recognition memory task, for example, and there is no doubt that the same set of brain regions keeps emerging (e.g., Rugg & Henson, 2002).

Final points

Finally, I wish to clarify a few points that I am *not* making. I am not arguing that neuroimaging is the panacea to theoretical problems in psychology, or that imaging data are "superior" to behavioural data. I think there is a real danger that pictures of blobs on brains seduce one into thinking that we can now directly observe psychological processes ("reifying" theoretical constructs). To the neuroscience undergraduate in particular, imaging data may seem to offer a more direct window on the mind than do behavioural data. Yet such high hopes are unlikely to be met. Indeed, from a pragmatic point of view, one could argue that, given the expense of acquiring imaging data and a finite amount of research funding, imaging experiments are of less value than behavioural experiments when the experimental hypothesis can be addressed by both (though I think it would be unwise never to seek converging evidence from imaging data because of its greater expense). Having said this, there are of course many other reasons for purchasing a brain scanner. MRI in particular is an extremely flexible tool (Buxton, 2002) and offers much hope for anatomical, physiological, and clinical research.

I am not claiming that there is a one-to-one mapping between (psychological) function and (brain) structure. I am only arguing that such a mapping is necessary if imaging data are to be used to inform psychological theories. I am also not claiming that a one-to-one function–structure mapping is necessary for imaging data to have any value. For example, one might fruitfully study the effects of attention using imaging methods without caring about the localization of those effects—for instance, whether attention raises baseline excitability prior to stimulation and/or augments the response to stimulation (Chawla, Rees, & Friston, 1997)—an inference that appeals to the temporal rather than spatial properties of imaging data and does not easily

conform to function-to-structure deduction or structure-to-function induction. Or one might be interested solely in the brain's implementation of a "known" function without intending to use the data to modify psychological characterization of that function. In other words, the value of imaging is not solely related to new knowledge about psychological theory; it also provides new knowledge about how the brain mediates psychological processes.

Functional imaging is also only one of several other methods of potential interest to experimental psychologists. In particular, the methods of neuropsychology, electrical stimulation, and transcranial magnetic stimulation (TMS) face similar issues regarding function–structure mapping. Within the present framework, these methods can be viewed as experimental groupings or manipulations of activity within different brain regions, with the dependent variable now being behaviour (Figure 9). With neuropsychology and TMS, activity of a certain region might be assumed to be absent or disrupted; with electrical stimulation, it might be assumed to be boosted. As such, these methods are complementary to functional imaging, and both types of methods may, in fact, be necessary to establish a function–structure mapping (Noppeney et al., in press; Price et al., 1999; Sarter et al., 1996). Moreover, functional imaging is useful to test the homology between human and nonhuman primates that is needed to transfer what we have learned from neurophysiological studies in the nonhuman primate to the human (Tootell, Tsao, & Vanduffel, 2003).

In summary, I hope that experimental psychologists will begin to share some of the excitement surrounding functional neuroimaging, while retaining a certain amount of caution regarding the nature of the inferences made, the statistical bases of those inferences, and, most importantly, the implicit assumptions made.

REFERENCES

Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In D. E. Meyers & S. Kornblum (Eds.), *Attention and performance XIV* (pp. 183–218). Cambridge, MA: MIT Press.

Anderson, J., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Attwell, D., & Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends in Neuroscience*, *25*, 621–625.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*, 551–565.

Boring, E. G. (1963). *History, psychology & science: Selected papers* (R. I. Watson & D. T. Campbell, Eds.). New York: John Wiley and Sons.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargelstellt auf Grund des Zellesbaues*. Leipzig: Barth.

Brown, G. D., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127–181.

Buchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, *7*, 768–778.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Buxton, R. B. (2002). *Introduction to functional magnetic resonance imaging*. Cambridge: Cambridge University Press.

Caplan, D. (1981). On the cerebral localization of linguistic functions: Logical and empirical issues surrounding deficit analysis and functional localization. *Brain and Language*, *14*, 120–137.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*, 41–46.

Catalan, M. J., Honda, M., Weeks, R. A., Cohen, L. G., & Hallett, M. (1998). The functional neuroanatomy of simple and complex sequential finger movements: A PET study. *Brain*, *121*, 253–264.

Chawla, D., Rees, G., & Friston, K. (1997). The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, *2*, 671–676.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, *210*, 207–209.

Coltheart, M. (2002). Cognitive neuropsychology. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology* (3rd ed.) (pp. 139–174). New York: John Wiley.

Coltheart, M. (2004). Brain imaging, connectionism and cognitive neuropsychology. *Cognitive Neuropsychology*, *21*, 21–25.

Corlett, P. R., Aitken, M., Dickinson, A., Shanks, D. R., Honey, G., Honey, R. A. E., Robbins, T. W., Bullmore, E. T., & Fletcher, P. C. (2004). *Using lateral prefrontal error signal to explore mechanisms of associative learning*. Manuscript submitted for publication.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*, 55–67.

Daum, I., Schugens, M. M., Ackermann, H., Lutzenberger, W., & Birbaumer, D. J. (1993). Classical conditioning after cerebellar lesions in humans. *Behavioral Neuroscience*, *107*, 748–756.

Dehaene, S., Naccache, L., Le Clec, H. G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–533.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*, 355–372.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*, 820–829.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neuroscience*, *23*, 475–83.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

Fletcher, P. C., & Henson, R. N. (2001). Frontal lobes and human memory: Insights from functional neuroimaging. *Brain*, *124*, 849–881.

Fletcher, P. C., Stephenson, M. E., Carpenter, T. A., Donovan, T., & Bullmore, E. (2003). Regional brain activations predicting subsequent memory success: An event-related fMRI study of the influence of encoding tasks. *Cortex*, *39*, 1009–1026.

Formisano, E., Kim, D.-S., Di Salle, F., van de Moortele, P.-F., Ugurbill, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *13*, 859–869.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312–316.

Friston, K. (2002a). Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience*, *25*, 221–250.

Friston, K. J. (2002b). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*, 113–143.

Friston, K. J., & Buchel, C. (2000). Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proceedings of the National Academy of Sciences USA*, *97*, 7591–7596.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*, 1273–1302.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*, 189–210.

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S., & Dolan, R. J. (1996). The trouble with cognitive subtraction. *Neuroimage*, *4*, 97–104.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*, 23–30.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*, 191–7.

Halsband, U., Ito, N., Tanji, J., & Freund, H. J. (1993). The role of premotor cortex and the supplementary motor area in the temporal control of movement in man. *Brain*, *116*, 243–266.

Harley, T. A. (2004a). Does cognitive neuropsychology have a future? *Cognitive Neuropsychology*, *21*, 3–16.

Harley, T. A. (2004b). Promises, promises. *Cognitive Neuropsychology*, *21*, 51–56.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2405–2407.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1210–1230.

Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, *3*, 142–151.

Henson, R. N. A. (1998). Short-term memory for serial order: The Start–End Model. *Cognitive Psychology*, *36*, 73–137.

Henson, R. N., Burgess, N., & Frith, C. D. (2000b). Recoding, storage, rehearsal and grouping in verbal short-term memory: An fMRI study. *Neuropsychologia*, *38*, 426–440.

Henson, R. N. A., Cansino, S., Herron, J. E., Robb, W. G. K., & Rugg, M. D. (2003c). A familiarity signal in human anterior medial temporal cortex. *Hippocampus*, *13*, 259–262.

Henson, R. N., Goshen-Gottstein, Y., Ganel, T., Otten, L. J., Quayle, A., & Rugg, M. D. (2003a). Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cerebral Cortex*, *13*, 793–805.

Henson, R. N. A., Hartley, T., Flude, B., Burgess, N., & Hitch, G. J. (2003b). Selective interference with verbal short-term memory for item and order information: Tests of a timing signal hypothesis. *Quarterly Journal of Experimental Psychology, 56A*, 1307–1334.

Henson, R. N. A., & Rugg, M. D. (2002). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, *41*, 263–270.

Henson, R. N. A., Rugg, M. D., Shallice, T., & Dolan, R. J. (2000c). Confidence in recognition memory for words: Dissociating right prefrontal roles in episodic retrieval. *Journal of Cognitive Neuroscience*, *12*, 913–923.

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. (1999). Recollection and familiarity in recognition memory: An event-related fMRI study. *Journal of Neuroscience*, *19*, 3962–3972.

Henson, R., Shallice, T., & Dolan, R. (2000a). Neuroimaging evidence for dissociable forms of repetition priming. *Science*, *287*, 1269–1272.

Herron, J. E., Henson, R. N., & Rugg, M. D. (2004). Probability effects on the neural correlates of retrieval success: An fMRI study. *Neuroimage*, *21*, 302–310.

Hirshman, F., & Master, S. (1997). Modeling the conscious correlates of recognition memory. *Memory & Cognition*, *25*, 345–351.

Hubbard, E. M. (2003). A discussion and review of Uttal (2001), "The New Phrenology". *Cognitive Science Online*, *1*, 22–33.

Inoue, C., & Bellezza, F. (1998). The detection model of recognition memory using know and remember judgments. *Memory & Cognition*, *26*, 299–308.

Kaas, J. H. (1987). The organization of neocortex in mammals: Implications for theories of brain function. *Annual Reviews in Psychology*, *38*, 129–151.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialised for face perception. *Journal of Neuroscience*, *17*, 4302–4311.

Kherif, F., Poline, J. B., Flandin, G., Benali, H., Simon, O., Dehaene, S., & Worsley, K. J. (2002). Multivariate model specification for fMRI data. *Neuroimage*, *16*, 1068–1083.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 699–710.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, *23*, 425–449.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lan, N., Montelione, G. T., & Gerstein, M. (2003). Ontologies for proteomics: Towards a systematic definition of structure and function that scales to the genome level. *Current Opinion in Chemical Biology, 7,* 44–54.

Lashley, K. S. (1963). *Brain mechanisms and intelligence: A quantitative study of injuries to the brain*. New York: Dover. (Original work published 1929.)

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468.

Lee, C. L., & Estes, W. K. (1981). Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 149–169.

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

Marr, D. (1982). *Vision*. San Francisco: WH Freeman.

Marshall, J. C. (1984). Multiple perspectives on modularity. *Cognition*, *17*, 209–242.

McIntosh, A. R. (1999). Mapping cognition to the brain through neural interactions. *Memory*, *7*, 523–548.

Mehler, J., Morton, J., & Jusczyk, P. W. (1984). On reducing language to biology. *Cognitive Neuropsychology*, *1*, 83–116.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology*, *26*, 3–67.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Noppeney, U., Friston, K. J. & Price, C. J. (in press). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*.

Otten, L. J., Henson, R. N., & Rugg, M. D. (2001). Depth of processing effects on neural correlates of memory encoding: Relationship between findings from across- and within-task comparisons. *Brain*, *124*, 399–412.

Otten, L. J., Henson, R. N., & Rugg, M. D. (2002). State-related and item-related neural correlates of successful memory encoding. *Nature Neuroscience*, *28*, 28.

Otten, L. J., & Rugg, M. D. (2001). Task-dependency of the neural correlates of episodic encoding as measured by fMRI. *Cerebral Cortex*, *11*, 1150–1160.

Page, M. P. A. (2004). What can't functional imaging tell the experimental psychologist? Abstract from EPS January conference, London.

Page, M. P. A., & Norris, D. G. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761–781.

Price, C. J., & Friston, K. J. (1997). Cognitive conjunction: A new approach to brain activation experiments. *Neuroimage*, *5*, 261–270.

Price, C. J., & Friston, K. J. (in press). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*.

Price, C. J., Warburton, E. A., Moore, C. J., Frackowiak, R. J. S., & Friston, K. J. (1999). Delineating necessary and sufficient neural systems with functional imaging studies of neuropsychological patients. *Journal of Cognitive Neuroscience*, *11*, 4371–4382.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, *21*, 89–102.

Ramnani, N., & Miall, R. C. (2004). A system in the human brain for predicting the actions of others. *Nature Neuroscience*, *7*, 85–90.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rees, G., Frith, C., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science*, *278*, 1616–1619.

Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science*, *286*, 2504–2506.

Rugg, M. D. (1995). ERP studies of memory. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind* (pp. 132–170). Oxford: Oxford University Press.

Rugg, M. D., & Coles, M. G. H. (Eds.). (1995). *Electrophysiology of mind*. Oxford: Oxford University Press.

Rugg, M. D., & Henson, R. N. A. (2002). Episodic memory retrieval: An (event-related) functional neuroimaging perspective. In A. E. Parker, E. L. Wilding, & T. Bussey (Eds.), *The cognitive neuroscience of memory encoding and retrieval* (pp. 3–37). Hove, UK: Psychology Press.

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Science*, *4*, 108–115.

Saito, S. (2001). The phonological loop and memory for rhythms: An individual differences approach. *Memory*, *9*, 313–322.

Sarter, M., Bernstein, G. G., & Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience. *American Psychologist*, *51*, 13–21.

Schacter, D. L., & Tulving, E. (Eds.). (1994). *Memory systems 1994*. Cambridge, MA: MIT Press.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shallice, T. (2003). Functional imaging and neuropsychology findings: How can they be linked? *Neuroimage*, *20*, S146–154.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*, 5–42.

Sternberg, S. (1969a). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.

Sternberg, S. (1969b). Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist*, *57*, 421–457.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologia*, *106*, 147–246.

Tootell, R. B. H., Tsao, D., & Vanduffel, W. (2003). Neuroimaging weighs in: Humans meet macaques in "primate" visual cortex. *Journal of Neuroscience*, *23*, 3981–3989.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist*, *26*, 1–12.

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

Wilkinson, D., & Halligan, P. (2004). The relevance of behavioural measures for functional-imaging studies of cognition. *Nature Reviews Neuroscience*, *5*, 67–73.

Winston, J., Henson, R., Fine-Goulden, M., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology, 92,* 1830–1839.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language*, *34*, 622–642.

Original manuscript received 4 August 2004
Accepted revision received 10 September 2004
PrEview proof published online 17 November 2004