In press @ A.Duarte, M.Barense, and D.R. Addis (Eds) *The Wiley-Blackwell Handbook on the Cogn. Neurosci. of Memory*

Activation and information in working memory research

Bradley R. Postle
Dept. of Psychology, University of Wisconsin-Madison

Recent studies applying multivariate pattern analysis (MVPA) to functional magnetic resonance imaging (fMRI) data sets have called into question long-held assumptions about the importance of sustained, elevated activity for the short-term retention of information (i.e., short-term and working memory). Findings from these studies include the fact that delay-period activity in frontal and parietal cortex may not correspond to information storage, per se, and that information that is in short-term memory, but outside the focus of attention, may be retained in a transient structural code.

Keywords: multivariate pattern analysis; MVPA; short-term memory, PFC, fMRI

Word count: 11,178

1202 West Johnson St.
Madison, WI   53726
USA
tel: +1 608-262-4330
fax: +1 608-262-4029
postle@wisc.edu

In a 2006 review I wrote that "working memory functions arise through the coordinated recruitment, via attention, of brain systems that have evolved to accomplish sensory-, representation-, and action-related functions" (Postle, 2006). By-and-large, ensuing developments in cognitive, computational, and systems neuroscience have been consistent with this perspective. One salient example is a 2011 special issue of *Neuropsychologia* that is devoted to the interrelatedness of the constructs of attention and working memory (Nobre and Stokes, 2011). Interestingly, however, the past several years have witnessed developments in the analysis of high-dimensional datasets (including those generated by functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and multi-unit extracellular electrophysiology) that, in some cases, call for a reconsideration of the interpretation of many of the studies that feature in the Postle (2006) review and, indeed, in some of those in the more recent, above-mentioned special issue of *Neuropsychologia*. This chapter will consider some of the implications of these recent developments for working memory and attentional research. (A second theme that has recently been gaining in prominence in working memory research, although by no means a "new development", is the critical role of network-level oscillatory dynamics in supporting working memory and attentional functions. This second theme has been covered extensively elsewhere, including in a recent chapter by this author (Postle, 2011), and so won't be addressed in detail here.)

Activation and information in the interpretation of physiological signals

The signal-intensity assumption

The idea of sustained activity as a neural basis of the short-term retention (STR) of information (i.e., the "storage" or "maintenance" of functions of short-term memory (STM) and working memory) has been a potent one that can be traced back at least as far as the reverberatory trace in Hebb's dual-trace model of long-term memory (LTM) formation, the active reverberation within a circuit being the initial trace that served the function of the STR of the memory until synapses making up the circuit could be strengthened to create the (second) long-lasting trace (Hebb, 1949). Since the 1970s, neurons in the monkey (and other species) that demonstrate sustained activity throughout the delay period of delay tasks have been seen as a neural embodiment of this trace. First observed in prefrontal cortex (PFC) and MD thalamus by Fuster and Alexander (1971), sustained delay-period activity has since been observed in many brain areas, including not only "high level" regions of parietal and temporal cortex (e.g., Gnadt and Andersen, 1988; Nakamura and Kubota, 1995; Suzuki et al., 1997), but also, in a modality-dependent manner, in primary sensory cortex (e.g., Zhou and Fuster, 1996; Super et al., 2001). In the human, delay-period neuroimaging signal with intensity that is elevated above baseline has long been considered a correlate of the STR of information (e.g., Jonides et al., 1993; Courtney et al., 1997; Zarahn et al., 1997), and the strength of this elevated signal, in comparison with other conditions, used to support models of the neural organization of working memory function. Thus, for example, statistically greater delay-period activity for, say, object information vs. spatial information, has been taken as evidence for the neural segregation of the STR of these two types of information (e.g., Courtney et al., 1996; Owen et al., 1998). The gold standard of evidence that a signal represents the STR of information has been evidence of monotonic increases in signal intensity with increasing memory load ("load sensitivity", e.g., Postle et al., 1999; Jha and McCarthy, 2000; Leung et al., 2004; Todd and Marois, 2004; Xu and Chun, 2006).

Last to be reviewed here is the use of functional localizers to identify putatively category-selective regions of the brain. The classic example is that of the "fusiform face area" (FFA), a region in mid-fusiform gyrus that is typically found to respond with stronger signal intensity to the visual presentation of faces than of objects from other categories, such as houses (Kanwisher et al., 1997). In working memory and attention research, a commonly used strategy has been to identify "category specific" regions of cortex with functional localizer scans (e.g., alternating blocks of faces vs. houses), then to see how activity in these regions of interest (ROI) varies during cognitive tasks that feature stimuli from these same categories. Thus, for example, a neural correlate of object-based attention is inferred when signal intensity in the FFA and in an analogous region of "house-selective" cortex is positively correlated with endogenous attentional cues, despite the fact that face and house stimuli are always present in a superimposed, translucent display (O'Craven et al., 1999). Similarly, neural correlates of the STR of face vs. scene information are inferred from the fact that delay-period activity in a FFA ROI is greater for face memory than for house memory, and the converse is true for a Parahippocampal Place Area (PPA) ROI (Ranganath et al., 2004).

Each of the types of experimental strategy reviewed in the preceding paragraph draws on a common underlying assumption, which is that one can infer the active representation of a particular kind of information from the signal intensity in a local area of the brain. (For expository expediency, I will refer to this as the *signal-intensity assumption*.) In recent years, however, it has become increasingly clear that the signal-intensity assumption is subject to important limitations. Empirically, this has been seen in an increasing number of studies in which it fails to account for working memory performance. And, as we shall see, an increasing appreciation for the multivariate nature of neuroimaging datasets (and, indeed, of brain function) provides a perspective from which the limitations of the signal-intensity assumption become clearer.

First, a brief review of empirical demonstrations that the seemingly straightforward interpretation of elevated delay-period activity as serving a mnemonic function can be problematic. These can be organized into two categories: failures of specificity; and failures of sensitivity. The former refers to instances in which elevated delay-period activity can be shown to serve a function other than the STR of information; the latter, to instances in which behavioral performance makes clear that the subject is successfully remembering information, yet no evidence of elevated delay-period activity can be found. Examples of failures of specificity include:

- Neurons with elevated delay-period activity in a memory task exhibit similarly sustained activity during the "delay" period of a visually guided saccade task, when no memory is required (Tsujimoto and Sawaguchi, 2004).
- Neurons that in a "standard" paradigm seem to encode a sensory representation of the to-be-remembered sample stimulus can be shown in a rotation condition to dynamically change during a single delay period from retrospectively representing the location of the sample to prospectively representing the target of the impending saccade (Takeda and Funahashi, 2002; Takeda and Funahashi, 2004; Takeda and Funahashi, 2007).
- A study designed to dissociate the focus of spatial attention from the focus of spatial memory finds the majority of delay-active neurons to track the former (Lebedev et al., 2004).

(Limits of the specificity assumption will also factor importantly in the consideration of "reverse inference" in neuroimaging, which appears further along, in the section on *Implications of MVPA for ROI-based analyses*.)

Examples of failures of sensitivity include:

- In the monkey, STM for the direction of moving dots in a sample display can be excellent, despite the failure to find directionally tuned neurons, in either area MT or the PFC, that sustain elevated activity across the delay-period (Bisley et al., 2004; Zaksas and Pasternak, 2006).
- In a human fMRI study in which subjects maintained one of two different memory loads across a 24-sec delay period, although sustained, although elevated delay-period activity was observed in several frontal and parietal sites during the delay, none showed load sensitivity, leaving uncertain whether these regions were actually involved in storage (Jha and McCarthy, 2000).

Against this backdrop, there has been an increased appreciation for limitations of the univariate analytic framework within which hypotheses about differences in signal intensity are most commonly tested. With neuroimaging data, the most familiar approach is to solve the general linear model (GLM) in a mass univariate (e.g., Friston et al., 1995) manner. (That is, the GLM is solved effectively independently at each of the (typically) thousands of data elements in a data set.) Typically, this approach leads to the identification of elevated (or decreased) signal intensity in voxels occupying a several-cubic-millimeter (or larger) volume of tissue, and the pooling across these voxels to extract a spatially averaged time course. Implementing this univariate approach to implement the signal-intensity assumption often engages a second assumption that can also be problematic, that of homogeneity of function. That is, by pooling across "activated" (or "deactivated") voxels, one is assuming that they are all "doing the same thing". Finally, the interpretation of the activity from this cluster of voxels often entails a third, often implicit, assumption, which is that this locally homogenous activity can be construed as supporting a mental function independent of other parts of the brain (i.e., modularity)[i]. Each of these assumptions is difficult to reconcile with the increasingly common recognition that neural representations are high-dimensional, and supported by anatomically distributed, dynamic computations (e.g., Buzsaki, 2006; Kriegeskorte et al., 2006; Norman et al., 2006; Bullmore and Sporns, 2009; Cohen, 2011).

Information-based analyses

An important conceptual advance in neuroimaging methods occurred with the publication by Haxby and colleagues (2001) of evidence that meaningful information about neural representations can be obtained from the patterns of activity in unthresholded fMRI data. This was soon followed by the application of powerful machine learning algorithms to fMRI datasets in an approach that has come to be known as multivariate pattern analysis (MVPA) (e.g., Haynes and Rees, 2006; Kriegeskorte et al., 2006; Norman et al., 2006; Pereira et al., 2009). As the name implies, MVPA differs fundamentally from signal intensity-based approaches in that it treats neural datasets as single high-dimensional images, rather than as a collection of independent low-dimensional elements. Therefore, it affords the detection and characterization of information that is represented in patterns of activity distributed within and across multiple regions of the brain. A detailed explication of the details underlying MVPA and its implementation to neuroimaging datasets is beyond the scope of this chapter, but what bears highlighting here is that MVPA is not subject to many of the problematic assumptions associated with signal intensity-based analyses. This includes not only the assumptions of homogeneity of function and of modularity, but also, and most importantly for the topic of this chapter, the very assumption that the STR of information is accomplished via sustained, elevated activity. Indeed, tests of this assumption are the first applications of MVPA to working memory research that I will review here.

The possibility that the STR of information may not depend on sustained activity that is elevated above a baseline (typically, the ITI) was demonstrated by two MVPA studies of visual STM that focused on primary visual cortex (V1). These showed that, although this region did not show elevated activity during the delay period, it nonetheless contained representations of the to-be-remembered stimuli that spanned the delay period (Harrison and Tong, 2009; Serences et al., 2009). In addition to building on what had been reported from V1 in the monkey (Super et al., 2001), these studies clearly demonstrated the increased sensitivity that is typical of MVPA vs. signal intensity-based analyses, in that no studies applying the latter to an fMRI data set had previously implicated V1 in the STR of visual information.

Against this backdrop, a clear next step would be a direct test of the assumption that elevated delay-period activity carries trial-specific stimulus information. To implement it, Riggall and Postle (2012) acquired fMRI data during delayed-recognition of visual motion, and analyzed them with both a GLM and MVPA. The former identified sustained, elevated delay-period activity in superior and lateral frontal cortex and in intraparietal sulcus (IPS), regions that invariably show such activity in studies of STM and working memory. When we applied MVPA, however, the pattern classifiers implementing the analysis were unable to recover trial-specific stimulus information from these delay-active regions (Figure 1). This was not merely a failure of our MVPA methods, because the same classifiers successfully identified trial-specific stimulus information in posterior regions that had not been identified by the GLM: lateral temporooccipital cortex, including the MT+ complex, and calcarine and pericalcarine cortex. Nor was it the case that the frontal and parietal regions were somehow "unclassifiable", because pattern classifiers were able to extract trial-specific task instruction-related information from these regions. Specifically, MVPA showed the frontal and parietal regions to encode whether the instructions on a particular trial were to remember to the speed or the direction of the moving dots that had been presented as the sample stimulus, a finding consistent with previous reports from the monkey (Freedman and Assad, 2006; Swaminathan and Freedman, 2012). Thus, it is unlikely that the failure to recover stimulus-specific information from the frontal and parietal regions (i.e., that the to-be-remembered direction of motion was 42°, 132°, 222°, or 312°) is that they encode information on a finer spatial scale than the posterior regions at which item-level decoding was successful[ii]. Rather, our conclusion is that the elevated delay-period activity that is measured with fMRI may reflect processes other than the storage, per se, of trial-specific stimulus information. Further, and consistent with previous studies (Harrison and Tong, 2009; Serences et al., 2009), it may be that the short-term storage of stimulus information is represented in patterns of (statistically) "subthreshold" activity distributed across regions of low-level sensory cortex that univariate methods cannot detect.

The finding from Riggall and Postle (2012) has potentially profound effects for our understanding of the neural bases of the STR of information, because it calls into question one of the more enduring assumptions of systems and cognitive neuroscience. We have reason to believe that it will hold up, because other groups are reporting compatible findings. For example, Linden and colleagues (2012) have reported a failure with MVPA to recover delay-period stimulus category information from frontal and parietal cortex, and Christophel et al. (2012) a failure to recover delay-period information specific to complex artificial visual stimuli from frontal cortex. (The distinction between classifying at the item level (e.g., Christophel et al., 2012; Riggall and Postle, 2012) vs. at the category level (as by, e.g., Linden et al., 2012) is an important one, in that the former provides the stronger evidence for memory storage, per se.

At this point it is useful to introduce the idea of an *active neural representation*. To illustrate, although Riggall and Postle (2012) contrasted signal intensity, a traditional index of "activation", vs. classifiability, it is important to note that successful classification also depended on evaluation of levels of activity within individual voxels. Thus, there is an important distinction to be made between signal intensity-based *activation*, which can be construed as a first-order physiological property[iii], and a multivariate pattern-based *neural representation*, a second order property (not just activity, but the pattern of activity) that is detectable with MVPA but not with univariate approaches. Nonetheless, a MVPA-detectable neural representation is an active representation, in the sense that neural activity must organize itself to create this pattern, and the neural representation is only present (i.e., only active) for the span of time that we assume it to be psychologically active. (E.g., Riggall and Postle (2012) found that MVPA of stimulus direction was only successful during a trial, when subjects were presumed to be thinking about a stimulus, and not during the ITI, when it is assumed that they were not.) Another way to illustrate this idea of an active neural representation is to consider information held in LTM. For the studies of Polyn et al. (2005) and Lewis-Peacock and Postle (2008), for example, it is assumed that all of the U.S.-citizen subjects for these studies were familiar with the American actor John Wayne prior to volunteering for these studies. However, it is also assumed that none of them were *actively* thinking about John Wayne prior to being shown his image during the course of the study. Thus, there existed in the brains of these subjects an *inactive* neural representation of John Wayne that was not detectable by MVPA during portions of the experiment when subjects were not thinking about John Wayne. This neural representation became *active* when subjects were viewing an image of actor, or retrieving this image from memory, and MVPA was sensitive to this change in the state of the LTM representation. The concept of an active neural representation is of central importance to the next studies to be reviewed here.

One of the questions raised by the Riggall and Postle (2012) findings is *what is the function of the sustained, delay-period activity that has been reported in the hundreds (if not more) of published studies on the neural correlates of STM and working memory since the 1970s?* Several possible answers to this question (and it is almost certainly true that there are several answers) have been reviewed in the first section of this chapter. Our group has also begun addressing this question from the theoretical perspective that working memory performance may be achieved, in part, via the temporary activation of LTM representations. First, in a study employing MVPA that won't be reviewed in detail here, we established the neural plausibility of this idea (Lewis-Peacock and Postle, 2008). In two more recent studies, we have worked from models that posit multiple states of activation, including, variously, a capacity-limited focus of attention, a region of direct access, and a broader pool of temporarily activated representations, all nested within the immense network of latently stored LTM (Cowan, 1988; McElree, 2001; Oberauer, 2002). Importantly, these models distinguish between the STR of information – which can be accomplished in any of the activated states of LTM – from attention to information – which is a capacity-limited resource that can be applied only to a small subset of highly activated representations.

The first of two studies that will be reviewed in this context (Lewis-Peacock et al., 2012) was an fMRI study of a multistep delayed-recognition task (adopted from Oberauer, 2005) presenting two sample stimuli, then retrocues informing the subject which sample was relevant for each of the two successively presented memory probes. More specifically, each trial began with the presentation of two sample stimuli, always selected from two of three categories (lines, words, and pronounceable pseudowords), one in the top half of the screen and one in the bottom half. After offset of the stimulus display and an initial delay period, a retrocue indicating which sample was relevant for the first recognition probe, followed by a second delay, followed by an initial Y/N recognition probe (and response). Critically, during the second delay both items needed to be kept in STM, even though only one was relevant for the first probe. This is because the first probe was followed by a second retrocue that, with equal probability, would indicate that the same item (a "repeat" trial) or the previously uncued item (a "switch" trial) would be tested by the trial-ending second Y/N recognition probe (and response). Thus, the first delay was assumed to require the active retention of two items, whereas the second delay would feature an "attended memory item" (AMI) and an "unattended memory item" (UMI)[iv]. The third delay would only require the retention of an AMI, because it was certain that memory for the item not cued by the second retrocue would never be tested. This design therefore allowed us to assess the prediction that there are different levels of neural activation corresponding to different hypothesized states of activation of LTM representations (Cowan, 1988; McElree, 2001; Oberauer, 2002).

Prior to performing this task, subjects were first scanned while performing a simple one-delay delayed recognition task, and the data from this *Phase 1* scan were used to train the classifier that was then applied to the data from the multistep task described in the previous paragraph. For Phase 1, subjects were trained to indicate whether the probe stimulus matched the sample according to a category-specific criterion -- synonym judgment for words, rhyme judgment for pseudowords, and an orientation judgment for line segments. Our rationale was that by training the classifiers (separately for each subject) on data from the delay period of this task, we'd be training it on patterns of brain activity related to the STR of just a single representational code: phonological (pseudoword trials), semantic (word trials), or visual (line trials). This, in turn, would provide the most unambiguous decoding of delay periods entailing the STR of two AMIs vs. one AMI and one UMI vs. one AMI.

In all trials, classifier evidence for both trial-relevant categories rose precipitously at trial onset and remained at the same elevated level until the onset of the first retrocue. This indicated that both items were encoded and sustained in the focus of attention across the initial memory delay, while it was equiprobable that either would be relevant for the first memory response. Following onset of the first retrocue, however, classifier evidence for the two memory items diverged. Postcue brain activity patterns were classified as highly consistent with the category of the cued item, whereas evidence for the uncued item dropped precipitously, becoming indistinguishable from the classifier's evidence for the stimulus category not presented on that trial (i.e., not different from baseline). If the second cue was a repeat cue, classifier evidence for the already-selected memory item remained elevated and that of the uncued item remained indistinguishable from baseline (Figure 2, Repeat). If, in contrast, the second cue was a switch cue, classifier evidence for the previously uncued item was reinstated, while evidence for the previously cued item dropped to baseline (Figure 2, Switch).

These results suggest that only AMIs are held in an active state. Classifier evidence for an active representation of UMIs returned to baseline levels, despite the fact that they could quickly be reactivated if cued during the second half of the trial. This is an important point because, despite the apparent loss of sustained activity, UMIs were nonetheless easily remembered after a brief delay. Thus, it may be that STM can be preserved across a brief delay despite the apparent loss of sustained representations. Further, it may be that delay-period activity reflects the focus of attention, rather than the STR, per se, of information. These possibilities are provocative, and have potentially profound implications for our understanding of working memory and attention. There are, however, several concerns and possible alternative explanations that need to be considered before these conclusions can be viewed as definitive. Some of these are taken up in the study of LaRocque et al. (2013), to be considered next.

One important caveat about the Lewis-Peacock et al. (2012) findings is that they were derived solely from fMRI data. The possibility exists, however, that UMIs may be retained in an active state via a mechanism to which the BOLD signal measured by fMRI is not sensitive. One candidate for such a mechanism is neuronal oscillations. There is considerable evidence that oscillatory dynamics in large populations of neurons are sensitive to, and may underlie, the STR of information (e.g., Jensen et al., 2002; Sauseng et al., 2009; Uhlhaas et al., 2009; Fuentemilla et al., 2010; Palva and Palva, 2011). Because the relationship between the BOLD signal and the broad band of frequencies at which different neural systems can oscillate is poorly understood, and certainly indirect, current fMRI methods are poorly suited to measure neural oscillatory dynamics. Therefore, we (LaRocque et al., 2013) designed this follow-up study to replicate the critical features of Lewis-Peacock et al. (2012), with the exception that we concurrently measured neural activity with the electroencephalogram (EEG), rather than with fMRI. In addition to being sensitive to neuronal oscillations across a broad, physiologically relevant range of frequencies, EEG has the additional property of affording greater temporal resolution than fMRI, which could permit more nuanced interpretations of the time course of the activation and deactivation of neural representations.

A one-sentence summary of the LaRocque et al. (2013) results is that they replicate the principal finding from Lewis-Peacock et al. (2012): MVPA of the EEG signal failed to find evidence that information that was outside the focus of attention, but nonetheless in STM (i.e., UMIs), was retained in an active state (Fig. 3). An additional analysis also ruled out the possibility that a neural representation is represented differently when being retained as a UMI vs. when being retained as a AMI. (If this were the case, MVPA of data trained on AMIs from Phase 1 might be expected to fail to detect UMIs during

Phase 2). This was achieved by implementing MVPA by training and testing on data from Delay 2 (i.e., following the first retrocue, when there was one AMI and one UMI) via the leave-one-out cross-validation procedure. These results qualitatively replicated those from the train-on-Phase-1-test-on-Phase-2 analysis.

In order to characterize, in the EEG classification data, the time course of the removal of memory items from the focus of attention we focused on the first retrocue, because this cue initiates the "unloading" of uncued items from the focus of attention (Oberauer, 2005; LaRocque et al., 2013). The estimate was made simply by determining the time point following retrocue onset at which evidence for an active neural representation of the UMI was lost (i.e., the time point at which classifier evidence for the UMI did not differ statistically from classifier evidence for the category that was not presented on that trial, an empirically derived baseline). The estimate derived from the EEG data of the time required for the neural representation of a single UMI to fall to baseline was 1.25 s (LaRocque et al., 2013).

The principal finding from the Lewis-Peacock et al. (2012) and the LaRocque et al. (2013) studies is that UMIs are not maintained in an active state. This is at odds with theoretical models positing one or more intermediate levels of activation between the focus of attention and unactivated LTM (Cowan, 1988; McElree, 2001; Oberauer, 2002; Olivers et al., 2011). Additionally, there are empirical studies that, in contrast to the studies reviewed here, have reported data derived from signal intensity-based analyses that could be interpreted as neural evidence for an intermediate state of activation for UMIs (e.g., Lepsien and Nobre, 2007; Nee and Jonides, 2008; Peters et al., 2012). Thus, resolving these empirical inconsistencies will be important for progress to be made on the theoretical front. In the following section it will be argued that these inconsistencies may be attributable to inferential limitations of the signal-intensity assumption.

Implications of MVPA for ROI-based analyses

One corollary of the signal-intensity assumption -- the assumption of the category specificity of neuroimaging signal measured from ROIs that have been defined either anatomically or functionally -- has underlain many cognitive neuroscience studies of working memory (including many performed by this author). This section will examine this assumption in the context of the AMI/UMI distinction, and will conclude that it can lead to incorrect inference. Note that although the argumentation will be framed in the relatively narrow terms of this question in the working memory literature, its logic may generalize more broadly to many domains of cognitive and systems neuroscience in which signal from "domain-specific ROIs" or "stimulus-selective neurons" is interpreted.

There are compelling intuitive and theoretical (Cowan, 1988; McElree, 2001; Oberauer, 2002; Olivers et al., 2011) reasons to posit an intermediate level of activation occupied by UMIs relative to AMIs and inactive LTM representations. This section will draw on two studies whose results can be seen as consistent with these multiple-levels models, and thus at odds with the findings from Lewis-Peacock et al. (2012) and LaRocque et al. (2013) that were reviewed in the previous section. One of these studies comes from Nobre and colleagues, who have been studying this question from the perspective that attention may modulate internal representations in a manner similar to its influence on perceptual processing (Griffin and Nobre, 2003; Nobre et al., 2004). In particular, Lepsien and Nobre (2003) used a procedure very similar to that described for Lewis-Peacock et al. (2012) and LaRocque et al. (2013): Subjects viewed the serial presentation of two stimulus items, one a face and one a scene, and during the ensuing delay period viewed a retrocue indicating one of the two stimulus categories and then, roughly 5 sec later, a second cue that instructed subjects to "switch" their attention to the previously uncued category or to "stay", the combination of cues being 100% informative about which stimulus would be tested with a recognition probe. Of principal relevance for this chapter are the modulations of activity measured from ROIs determined independently to respond more strongly to faces than scenes (and located in fusiform gyrus), or more strongly to scenes than faces (and located in parahippocampal gyrus). Signal intensity within these ROIs was seen to increase and decrease in a manner congruent with the category indicated by each cue (Griffin and Nobre, 2003). Although the authors make no explicit claims about multiple levels of activation in working memory, this aspect of their results can be interpreted as evidence for such models.

A second group, Nee and Jonides, has explicitly tested multiple-levels models (Nee and Jonides, 2008; Nee and Jonides, 2011). In one study they followed the rapid serial visual presentation of three words with a recognition probe, reasoning that the level of activation to recognition probes matching the most recently presented item (the "*-1 item*"), which was assumed to be held in a 1-item focus of attention, would differ from probes matching the other two serial positions, which, being outside the focus of attention, were presumed to make greater demands on retrieval processes (Nee and Jonides, 2008). Their results showed that probes matching the *-1 item* evoked higher-magnitude responses in inferior temporal cortex relative to probes matching the *-2* and *-3 items*, whereas the latter increased evoked higher-magnitude responses in the medial temporal lobe and left mid-ventrolateral prefrontal cortex (Nee and Jonides, 2008).

What is particularly germane for this chapter is that Lepsien and Nobre (2003), in summarizing their findings, characterize their fusiform and parahippocampal ROIs as "involved in maintaining representations of faces and scenes respectively" (p. 2072), and that Nee and Jonides (2008) characterize the inferior temporal cortex as "inferior temporal representational cortex" (p. 14228)[v]. That is, both of these publications articulate the assumption (pervasive within the cognitive neuroimaging literature) that, under certain circumstances, activity in carefully selected regions can be interpreted as being specific to a particular function. These are examples of reasoning based on reverse inference, reasoning "backward" from the presence of elevated activity in a certain area of the brain (in this case, e.g., a fusiform gyrus ROI), to the engagement of a particular cognitive function (e.g., the STR of face information)[vi].

The practice of reverse inference is widespread in the cognitive neuroimaging literature, despite the fact that one often hears it being disparaged in conversation with cognitive neuroscientists. And it is certainly true that, in some cases, reverse inference can lead to incorrect inferences. For example, it can be incorrect to conclude from an observation of elevated activity in the PFC during a task that working memory was engaged during the task. The reason for this is that "human prefrontal cortex is not specific for working memory" (D'Esposito et al., 1998)– it can support many functions other than working memory. An influential paper by Poldrack (2006) has cogently made the point that the validity of reverse inference depends critically on the specificity of the pattern of activity that is being interpreted. To illustrate this point quantitatively, he evaluated the proposition that "activation in Broca's area implies engagement of language function" by computing a Bayes factor (ratio of the posterior odds to the prior odds) with data from the BrainMap database (tabulating, for 'language studies' and 'not language studies' in the database, the number of contrasts for which Broca's area was 'activated' vs. 'not activated'). The resultant Bayes factor of 2.3 was considered only weak evidence for this inference. In a concluding section of his paper, however, Poldrack (2006) notes that "There are two ways in which to improve confidence in reverse inferences: increase the selectivity of response in the brain … or increase the prior probability of the cognitive process in question." (p. 62). The former factor is critical for this chapter: MVPA can support stronger reverse inferences than can univariate techniques because it measures high-dimensional neural representations that have markedly higher selectivity than do univariate activation peaks. Early evidence for this was demonstrated for face perception and the FFA by Haxby and colleagues (2001), who showed that faces could be discriminated from six other visual categories from the multivariate signal from ventral temporal cortex even when the FFA was excluded from the MVPA. Additionally, they showed a high level of discriminability for all categories (except shoes) when the MVPA was restricted to only the FFA (Haxby et al., 2001). Against this backdrop, we now return to the question of whether UMIs are represented at an intermediate level of activity between that of AMIs and inactive LTM.

In part to explore the contradictory findings regarding the physiological state in which UMIs are maintained, Jarrod Lewis-Peacock performed a head-to-head comparison of the results of univariate vs. multivariate analyses on the data from Lewis-Peacock et al. (2012). To do so, he first identified voxels that were selectively sensitive (in terms of univariate signal intensity) to the STR of just one of the three visual categories used in the experiment (visual, semantic, phonological). (This chapter will focus on what we called "exclusive" ROIs, defined as voxels that showed elevated delay-period signal intensity (vs. ITI baseline) for only one of the three categories of stimulus.) Second, he observed that the fluctuations of signal intensity levels in these regions followed the pattern that would be predicted by multiple-state models: Higher delay-period activity when a region's "specific" category was an AMI; lower, but above-baseline, delay-period activity when a region's "specific" category was a UMI. When the same ROI-specific signal was submitted to MVPA, however, a different pattern emerged: The multivariate patterns of activity in putatively category-specific ROIs reflected which stimulus category was currently in the focus of attention, regardless of whether or not that category was the one for which the ROI might be assumed to be "selective" (Figure 4). For example, multivariate patterns of activity in the voxels whose GLM-defined activity was specific to STM for phonological stimuli reliably conveyed, on trials featuring a semantic and a visual stimulus, which was the current AMI (Lewis-Peacock and Postle, 2012).

This finding has two implications that will be emphasized here. First, with regard to the narrower question of the neural representation of items in STM, it demonstrates empirically that a univariate "read out" of signal intensity from an ROI, despite its univariate specificity, cannot be used as a reliable index of the state of stimulus information in STM. (Further, it reinforces the idea that sustained activity need not be the neural basis for the STR of information (Lewis-Peacock and Postle, 2012; LaRocque et al., 2013; Stokes and Duncan, in press).) Second, and more broadly, it raises concerns about the practice of defining functional localizers in cognitive neuroimaging. The Lewis-Peacock and Postle (2012) data demonstrate empirically a point that has been made, either explicitly or implicitly, in many recent discussions of MVPA (e.g., Haxby et al., 2001; Haynes and Rees, 2006; Kriegeskorte et al., 2006; Norman et al., 2006; Pereira et al., 2009) (as well as of reverse inference (Poldrack, 2006)), which is that one can not reliably infer the active representation of information by simply inspecting the level of signal intensity in a particular region of the brain. A reason for this is that neural representations, widely assumed to be high-dimensional and anatomically distributed, are almost guaranteed to be poorly characterized by the essentially 1-dimensional time-varying signal from an ROI. Stated differently, interpretation of time-varying signal from an ROI requires the assumption that locally homogeneous activity (often on the order of a square centimeter, or more) can be construed as supporting a single mental function, and as doing so independent of other brain areas (for more elaboration of this argument, see Lewis-Peacock and Postle, 2012; Riggall and Postle, 2012). Because these latter assumptions are almost certainly not true under most conditions, it follows that elements within a functional ROI defined with univariate statistics are certain to be involved in many more functions than just the one that the experimenters had in mind when they set out to define this "functionally specific ROI". Similarly, the empirical results of applications of MVPA to multiunit extracellular recordings from the monkey (e.g.,Meyers et al., 2008; Crowe et al., 2010; Stokes and Duncan, in press) indicate that caution should be taken in inferring the function of activity in individual neurons that may have been defined as having "selective" properties.

Limitations and outstanding questions

This final section is organized by themes under which are reviewed caveats that need to accompany some of the arguments made earlier in this chapter, as well as some future directions that MVPA-based working memory research might take.

Necessity

Successful MVPA decoding means that patterns of activity are reliably different in two or more stimulus conditions, but it doesn't necessarily follow from this that a region from which one can decode is necessary for the active representation of that information. An alternative possibility is that the differential patterns observed in the area in question may be "echoes" of the critical stimulus-representing activity that is occurring elsewhere in the brain. This may explain the seeming contradiction of the lack of stimulus specificity of the FFA as demonstrated by MVPA of activity from this region (Haxby et al., 2001) versus the relative specificity of perceptual deficits that arise from damage to (as reviewed, e.g., by Farah, 1990) or stimulation of (e.g., Parvizi et al., 2012) this region. That is, it is possible that the ability to predict from FFA signal whether the stimulus being viewed is a cat, a bottle, or a chair (Haxby et al., 2001) may be due to the fact that differential levels of activity from the regions whose activity *is* necessary for perceiving stimuli from these categories also bias the state of activity in the FFA.

Another illustration of this idea that relates to the AMI/UMI distinction comes from Peters and colleagues (2012), who have argued for a distinction between *search templates* that are held in working memory to guide a visual search (and are analogous to the AMIs emphasized in this chapter) and *accessory memory items* that can be simultaneously held in working memory, but are irrelevant to the search (analogous to UMIs, e.g., Olivers et al., 2011). In an fMRI study of visual search from a rapid serial visual presentation (RSVP) of superimposed face and house stimuli, they were able to decode the category of the UMI, during the RSVP stream, from 17 discrete regions that were located in all four lobes of the brain and in neostriatum, in both hemispheres (Peters et al., 2012). Although it is possible that subjects solved their task by maintaining 17 discrete "copies" of the UMI, a more likely account is that the category-discriminating activity of some of these regions was a byproduct of task-critical neural representation occurring elsewhere. (Note that one cannot appeal to anatomically distributed regions for this particular argument, because the "roaming seachlight" technique used to identify these areas only used signal arising locally within the area covered by the searchlight at any single point in the procedure. This question of the "aperture" of MVPA is addressed again further along in this section.)

If successful decoding of information need not imply that the region in question is representing that information in a meaningful way, what about the converse? Can one posit that successful MVPA decoding is a necessary, if not sufficient, condition for determining that a region contributes to the representation of a particular stimulus or category of information? In a strict sense, the answer has to be 'no', because failure to decode stimulus (or category) identity is a failure to reject the null hypothesis. However, questions such as this one can sometimes be profitably reframed in terms of sensitivity.

Sensitivity

We can not definitively rule out the possibility that, e.g., the failure of Riggall and Postle (2012) to decode the remembered direction of motion from PFC was due to PFC representing this information at a spatial scale too fine-grained for fMRI to detect, or in a neural code to which fMRI is insensitive (e.g., oscillatory synchrony, Salazar et al., 2012). Having acknowledged this inherent limitation of interpreting negative findings with MVPA, however, it is also important to acknowledge that MVPA methods are unequivocally much more sensitive than univariate methods. Sticking with STM for motion, there is no univariate method (e.g., comparing intensity of BOLD evoked response from area MT and/or other regions; "fMRI adaptation"; etc.) that would allow one to reliably predict the remembered direction of motion on a particular set of trials. Thus, although there are legitimate concerns that one can raise about overinterpreting, say, a failure to decode motion direction from PFC, this does not alter the fact that such a conclusion can be drawn with higher confidence than can be the conclusion from elevated BOLD signal in PFC that this activity corresponds to the active representation of stimulus information. This is for the reason that multivariate patterns boast, almost as a rule, quantitatively superior specificity than do voxels defined by univariate contrasts.

Localization vs. anatomically distributed

One curious fact about MVPA is that one of the more common procedures by which it is applied to neuroimaging data sets is via the so-called roaming searchlight approach. This is implemented by defining a sphere with a given volume (e.g., Peters et al. (2012) used a radius of 6 mm), locating it on a particular voxel, and assessing multivariate changes within the voxels in the sphere as a function of task condition. The sphere is then moved by one voxel and the process repeated, iteratively across every voxel in the data set. I say "curious", because this procedure imposes strict localizationist constraints on a method inherently well suited to detect neural representations that are anatomically broadly distributed. Thus, if a particular representation depended on functional connectivity between distal regions of, say, parietal and temporal cortex, a searchlight analysis would not detect this representation. A practical reason for implementing a roaming searchlight is that is avoids some of the complications associated with whole-brain classification. These include concerns about overfitting the model solution if every gray-matter voxel is included in the analysis, or about introducing bias via the feature-selection step that one implements to select just a subset of voxels over which to perform MVPA. Some of these concerns are currently being addressed by applying sparse machine learning methods to MVPA[vii]. For an overview of machine learning approaches to MVPA, Pereira et al. (2009) is a good place to start. For the purposes of this chapter, I'll sum up this section by observing that an important goal for future cognitive neuroscience research is developing a better understanding of how the constraints of different implementations of MVPA can limit, or bias, what a particular analysis is telling us about the neural representation of information. This will allow researchers to implement the MVPA procedure that is best suited for the scientific question being addressed.

Conclusion

This chapter has summarized how a relatively recent methodological development, the introduction of MVPA to the analysis of neuroimaging data sets, has led to new insights about the neural bases of the STR of information. It may be that the sustained, elevated signal that has long been accepted as the neural correlate of STM relates to processes being carried out in "real time", but not to the storage, per se, of information. An important next step in validating this idea will be to assess what was previously referred to as the "gold standard" of evidence for STR, the sensitivity of particular regions to the amount of information being held in STM (i.e., "load sensitivity"). Preliminary data from the author's laboratory suggests that this, too, is not a reliable indicator of storage. Although this possibility may seem to fly in the face of "everything that we know" about STM, there are precedents for it in the literature. For example, it is widely accepted that the oscillation of circuits in sensory systems in the alpha band is a mechanism for suppressing the function of those circuits(e.g., Buzsaki, 2006). The finding of load-sensitivity of delay-period alpha-band power in verbal STM has therefore been interpreted as evidence for (stimulus-nonspecific) inhibition increasing as a function of load (Jensen et al., 2002). More recently, it has been suggested that the load-dependent changes in the "contralateral delay activity" (CDA) that are often interpreted as a neural correlate of storage in visual STM (e.g., Vogel and Machizawa, 2004; Reinhart et al., 2012) may also reflect changes in the dynamics of alpha-band oscillations (van Dijk et al., 2010). By this view, the CDA would be a by-product of general state changes in the visual system, rather than an index of storage, per se.

An additional literature that is implicated in the results and ideas reviewed here is that of extracellular recordings from awake, behaving monkeys performing tests of working memory and STM. Here, too, the recent application of multivariate methods has begun to challenge traditional assumptions about the function of sustained, elevated firing rates. For example, Meyers et al. (2008) have shown with MVPA that the representation of stimulus category information in a delayed-match-to-category task "is coded by a nonstationary pattern of activity that changes over the course of a trial with individual neurons [in inferior temporal cortex and PFC] containing information on much shorter time scales than the population as a whole" (p. 1407). That is, this information does not seem to be carried in the sustained activity of individual neurons. A similar principle has been reported for the representation of spatial information in parietal cortex (Crowe et al., 2010).

These are exciting times to be studying the neural and cognitive bases of working and short-term memory, and MVPA techniques are among the many recent methodological innovations that hold promise for important discoveries in the coming years.

References

Bisley, J., Zaksas, D., Droll, J. A. and Pasternak, T. (2004). Activity of neurons in cortical area MT during a memory for motion task. *Journal of Neurophysiology* **91**: 286-300.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10**: 186-198.

Buzsaki, G. (2006). *Rhythms of the Brain*. New York, Oxford University Press.

Christophel, T. B., Hebart, M. N. and Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *The Journal of Neuroscience* **32**: 2983–12989.

Cohen, M. X. (2011). It's about time. *Frontiers in Human Neuroscience* **5**: doi: 10.3389/fnhum.2011.00002.

Courtney, S. M., Ungerleider, L. G., Keil, K. and Haxby, J. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex* **6**: 39-49.

Courtney, S. M., Ungerleider, L. G., Keil, K. and Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386**: 608-611.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin* **104**: 163-171.

Crowe, D. A., Averbeck, B. B. and Chafee, M. V. (2010). Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. *The Journal of Neuroscience* **30**: 11640-11653.

D'Esposito, M., Ballard, D., Aguirre, G. K. and Zarahn, E. (1998). Human prefrontal cortex is not specific for working memory: a functional MRI study. *NeuroImage* **8**: 274-282.

Farah, M. J. (1990). *Visual Agnosia*. Cambridge, MA, MIT Press.

Freedman, D. J. and Assad, J. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature* **443**: 85-88.

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

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N. and Düzel, E. (2010). Theta-coupled periodic replay in working memory. *Current Biology* **20**: 606-612.

Fuster, J. M. and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science* **173**: 652-654.

Gnadt, J. W. and Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research* **70**: 216-220.

Griffin, I. C. and Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience* **15**: 1176-1194.

Harrison, S. A. and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**: 632-635.

Haxby, J. V., Gobini, M. I., Furey, M. L., Ishai, A., Schouten, J. L. and Pietrini, P. (2001). Distributed and overlapping representatinons of faces and objects in ventral temporal cortex. *Science* **293**: 2425-2430.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* **7**: 523-534.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY, John Wiley & Sons, Inc.

Jensen, O., Gelfand, J., Kounios, J. and Lisman, J. E. (2002). Oscillations in the Alpha Band (9-12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex* **12**: 877-882.

Jha, A. and McCarthy, G. (2000). The influence of memory load upon delay interval activity in a working memory task: an event-related functional MRI study. *Journal of Cognitive Neuroscience* **12, suppl. 2**: 90-105.

Jonides, J., Smith, E., Koeppe, R., Awh, E., Minoshima, S. and Mintum, M. (1993). Spatial working memory in humans as revealed by PET. *Nature* **363**: 623-625.

Kanwisher, N., McDermott, J. and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience* **17**: 4302-4311.

Kriegeskorte, N., Formisano, E., Sorger, B. and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Science (USA)* **104**: 20600-20605.

Kriegeskorte, N., Goebel, R. and Bandettini, P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Science (USA)* **103**: 3863-3868.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A., Oberauer, K. and Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *Journal of Cognitive Neuroscience* **25**: 127-142.

Lebedev, M. A., Messinger, A., Kralik, J. D. and Wise, S. P. (2004). Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biology* **2**: 1919-1935.

Lepsien, J. and Nobre, A. C. (2007). Attentional modulation of object representations in working memory. *Neuropsychologia* **17**: 2072-2083.

Lepsien, J., Thornton, I. and Nobre, A. C. (2011). Attention and short-term memory: Crossroads. *Neuropsychologia* **49**: 1569-1577.

Leung, H.-C., Seelig, D. and Gore, J. C. (2004). The effect of memory load on cortical activity in the spatial working memory circuit. *Cognitive, Affective, & Behavioral Neuroscience* **4**: 553-563.

Lewis-Peacock, J. A., Drysdale, A., Oberauer, K. and Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience* **23**: 61-79.

Lewis-Peacock, J. A. and Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *The Journal of Neuroscience* **28**: 8765-8771.

Lewis-Peacock, J. A. and Postle, B. R. (2012). Decoding the internal focus of attention. *Neuropsychologia* **50**: 470-478.

Linden, D. E. J., Oosterhof N.N., Klein C. and Downing, P. E. (2012). Mapping brain activation and information during category-specific visual working memory. *Journal of Neurophysiology* **107**: 628–639.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **27**: 817-835.

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology* **100**: 1407-1419.

Nakamura, K. and Kubota, K. (1995). Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. *Journal of Neurophysiology* **74**: 162– 178.

Nee, D. E. and Jonides, J. (2008). Neural correlates of access to short-term memory. *Proceedings of the National Academy of Science (USA)* **105**: 14228-14233.

Nee, D. E. and Jonides, J. (2011). Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: evidence for a 3-state model of memory. *NeuroImage* **15**: 1540-1548.

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R. and Mesulam, M. M. (2004). Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience* **16**: 363-373.

Nobre, A. C. and Stokes, M. G. (2011). Attention and short-term memory: Crossroads. *Neuropsychologia* **49**: 1391-1392.

Norman, K. A., Polyn, S. M., Detre, G. J. and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* **10**: 424-430.

O'Craven, K. M., Downing, P. E. and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature* **401**: 584-587.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **28**: 411-421.

Oberauer, K. (2005). Control of the contents of working memory—A comparison of two paradigms and two age groups. *Journal of Experimental Psychology: Learning, Memory, & Cognition* **31**: 714-728.

Olivers, C. N. L., Peters, J., Houtkamp, R. and Roelfsema, P. R. (2011). Different states in visual working memory: when it guides attention and when it does not. *Trends in Cognitive Sciences* **15**: 327-334.

Owen, A. M., Stern, C. E., Look, R. B., Tracey, I., Rosen, B. R. and Petrides, M. (1998). Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. *Proceedings of the National Academy of Sciences, USA* **95**: 7721-7726.

Palva, S. and Palva, J. M. (2011). Functional roles of alpha-band phase synchronization in local and large-scale cortical networks. *Frontiers in Psychology* **2**: doi: 10.3389/fpsyg.2011.00204.

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S. and Grill-Spector, K. (2012). Electrical stimulation of human fusiform face-selective regions distorts face perception. *The Journal of Neuroscience* **32**: 14915-14920.

Pereira, F., Mitchell, T. and Botvinick, M. M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* **45**: S199-S209.

Peters, J., Roelfsema, P. R. and Goebel, R. (2012). Task-relevant and accessory items in working memory have opposite effects on activity in extrastriate cortex. *The Journal of Neuroscience* **32**: 17003-17011.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* **10**: 59-63.

Polyn, S. M., Natu, V. S., Cohen, J. D. and Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science* **310**: 1963-1966.

Postle, B. R. (2006). Distraction-spanning sustained activity during delayed recognition of locations. *NeuroImage* **30**: 950-962.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience* **139**: 23-38.

Postle, B. R. (2011). What underlies the ability to guide action with spatial information that is no longer present in the environment? Spatial Working Memory. A. Vandierendonck and A. Szmalec. Hove, U.K., Psychology Press**:** 897-901.

Postle, B. R., Berger, J. S. and D'Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences (USA)* **96**: 12959-12964.

Postle, B. R., Druzgal, T. J. and D'Esposito, M. (2003). Seeking the neural substrates of working memory storage. *Cortex* **39**: 927-946.

Postle, B. R. and Hamidi, M. (2007). Nonvisual codes and nonvisual brain areas support visual working memory. *Cerebral Cortex* **17**: 2134-2142.

Ranganath, C., DeGutis, J. and D'Esposito, M. (2004). Category-specific modulation of inferior temporal activity during working memory encoding and maintenance. *Cognitive Brain Research* **20**: 37-45.

Reinhart, R. M., Heitz, R. P., Purcell, B. A., Weigand, P. K., Schall, J. D. and Woodman, G. F. (2012). Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *The Journal of Neuroscience* **32**: 7711-7722.

Riggall, A. C. and Postle, B. R. (2012). The relation between working memory storage and elevated activity, as measured with fMRI. *The Journal of Neuroscience* **32**: 12990-12998.

Salazar, R. F., Dotson, N. M., Bressler, S. L. and Gray, C. M. (2012). Content-specific fronto-parietal synchronization during visual working memory. *Science* **338**: 1097-1100.

Sauseng, P., et al. (2009). Brain oscillatory substrates of visual short-term memory capacity. *Current Biology* **19**: 1846-1852.

Serences, J. T., Ester, E. F., Vogel, E. K. and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science* **20**: 207-214.

Soon, C. S., Brass, M., Heinze, H. J. and Haynes, J.-D. (2008). Unconscious determinants of free decision in the human brain. *Nature Neuroscience* **11**: 543-545.

Stokes, M. and Duncan, J. (in press). Dynamic brain states for preparatory attention and working memory. Oxford's Handbook of Attention. S. Kastner and A. C. Nobre. Oxford, U.K., Osford University Press**:** 897-901.

Super, H., Spekreijse, H. and Lamme, V. A. F. (2001). A neural correlate of working memory in the monkey primary visual cortex. *Science* **293**: 120-124.

Suzuki, W. A., Miller, E. K. and Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *Journal of Neurophysiology* **78**: 1062-1081.

Swaminathan, S. K. and Freedman, D. J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nature Neuroscience* **15**: 315-320.

Takeda, K. and Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. *Journal of Neurophysiology* **87**: 567-588.

Takeda, K. and Funahashi, S. (2004). Population vector analysis of primate prefrontal activity during spatial working memory. *Cerebral Cortex* **14**: 1328-1339.

Takeda, K. and Funahashi, S. (2007). Relationship between prefrontal task-related activity and information flow during spatial working memory performance. *Cortex* **43**(1): 38-52.

Todd, J. J. and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* **428**: 751-754.

Tsujimoto, S. and Sawaguchi, T. (2004). Properties of delay-period neuronal activity in the primate prefrontal cortex during memory- and sensory-guided saccade tasks. *European Journal of Neuroscience* **19**(2): 447-457.

Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D. and Singer, W. (2009). Neural synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience* **3**: doi: 10.3389/neuro.3307.3017.2009.

van Dijk, H., van der Werf, J., Mazaheri, A., Medendorp, W. P. and Jensen, O. (2010). Modulations of oscillatory activity with amplitude asymmetry can produce cognitively relevant event-related responses. *Proceedings of the National Academy of Science (USA)* **107**: 900-905.

Vogel, E. K. and Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature* **428**: 748-751.

Xu, Y. and Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature* **440**: 91-95.

Zaksas, D. and Pasternak, T. (2006). Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *The Journal of Neuroscience* **26**: 11726-11742.

Zarahn, E., Aguirre, G. K. and D'Esposito, M. (1997). A trial-based experimental design for fMRI. *NeuroImage* **6**: 122-138.

Zhou, Y. D. and Fuster, J. M. (1996). Mnemonic neuronal activity in somatosensory cortex. *Proceedings of the National Academy of Sciences (USA)* **93**: 10533-10537.

Acknowledgments

I thank Drs. Mark Stokes and Jarrod Lewis-Peacock for helpful comments on an earlier draft of this chapter. This work was supported by NIH grant MH064498.

Figure Legends

Figure 1.a. Behavioral task from Riggall and Postle (2012). Subjects maintained the direction and speed of a sample motion stimulus over a 15 sec-long delay period. Midway through the delay period, they were cued as to the dimension on which they would be making an upcoming comparison against the remembered sample, either direction or speed. At the end of the delay period, they were presented with a probe motion stimulus and had to indicate with a button press whether it did or did not match the sample stimulus on the cued dimension.
b. BOLD and MVPA time courses from four ROIs, Sample presentation occurred at 0 sec, and at 8 sec subjects were cued that either the direction or speed of sample motion would be tested on that trial. (*A-D*) Average ROI BOLD activity. Data from direction-cued trials use solid lines and speed-cued trials use dashed-lines, bands cover average standard error across subjects. (*E-H*) ROI stimulus-direction decoding results and (*I-L*) ROI trial-dimension decoding results. Each waveform represents the mean direction-decoding accuracy across subjects (n = 7) for a classifier trained with data limited to a single time point in the trial and then tested on all time points in the hold out trials (e.g., the green line illustrates the decoding time course from a classifier trained on only data from time point 4, indicated by the small green triangle along the x-axis.) Horizontal bars along the top indicate points at which the decoding accuracy for the corresponding classifier was significantly above chance ($p < 0.05$, permutation test). Schematic icons of trial events are shown at the appropriate times along the x-axis. Data are unshifted in time.
The journal of neuroscience : the official journal of the Society for Neuroscience by SOCIETY FOR NEUROSCIENCE Copyright 2012 Reproduced with permission of SOCIETY FOR NEUROSCIENCE in the format Other Book via Copyright Clearance Center.

Figure 2.a. Behavioral tasks from Experiment 2 of Lewis-Peacock et al. (2012). (A) In the first phase, participants performed short-term recognition of a pseudoword (phonological STM), a word (semantic STM), or two lines (visual STM). (B) In the second phase, during the same scanning session, participants performed short-term recognition with two stimuli (between-category combinations of pseudowords, words, and lines). On half of the trials, the same memory item was selected as behaviorally relevant by the first and second cues (repeat trials), and on the other half of trials the second cue selected the previously uncued item (switch trials).
b. Classifier decoding from Experiment 2 of Lewis-Peacock et al. (2012). Results are shown separately for repeat (left) and switch (right) trials. Classifier evidence values for phonological, semantic, and visual were relabeled and collapsed across all trials into three new categories: *cued* (red, the category of the memory item selected by the first cue), *other* (blue, the category of the other memory item), and *irrel* (gray, the trial-irrelevant category). The colored shapes along the horizontal axis indicate the onset of the targets (red and blue circles, 0 sec), the first cue (red triangle, 10 sec), the first recognition probe (red square, 18 sec), the second cue (red or blue triangle, 22 sec), and the final recognition probe (red or blue square, 30 sec). Data for each category are shown as ribbons whose thickness indicate ±1 SEM across participants, interpolated across the 23 discrete data points in the trial-averaged data. Statistical comparisons of evidence values focused on within-subject differences. For every 2-sec interval throughout the trial, color-coded circles along the top of each graph indicate that the classifier's evidence for the *cued* or *other* categories, respectively, was reliably stronger ($p < .002$, based on repeated measures t tests, corrected for multiple comparisons) than the evidence for the *irrel* category. Reprinted with permission from Jarrod A. Lewis-Peacock, Andrew T. Drysdale, Klaus Oberauer, and Bradley R. Postle, 'Neural Evidence for a Distinction between Short-term Memory and the Focus of

Attention', Journal of Cognitive Neuroscience, 24:1 (January, 2012), pp. 61-79. © 2012 by the Massachusetts Institute of Technology.

Figure 3. Classifier decoding from EEG study of LaRocque et al. (2013). Results are shown separately for cue repeat (left) and cue switch trials (right). Graphical conventions are same as from Fig. 2, with exception that width of the brackets surrounding significance markers denotes extent of delay period used for statistical analysis; *p < 0.05, **p < 0.005.
Reprinted with permission from LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K. & Postle, B.R., 'Decoding Attended Information in Short-Term Memory: An EEG Study', Journal of Cognitive Neuroscience, 25:1 (January, 2013), pp. 127-142. © 2013 by the Massachusetts Institute of Technology.

Figure 4. Signal intensity vs. MVPA in "category selective" ROIs. Trial-averaged decoding of Phase 2 switch trials from Lewis-Peacock et al. (2012), using the GLM method (top row) and the MVPA method (bottom row). Data for each category (*1st*, the first cued category; *2nd*, the second cued category; *irrel*, the trial-irrelevant category) are shown as ribbons whose thickness indicate +/-1 SEM across participants. The colored shapes along this horizontal axis indicate the onset of the targets (green and purple circles), the first cue (green triangle), the first probe (green square), the second cue (purple triangle), and the second probe (purple square). Statistical comparisons focused on within-subject differences: For every 2-s interval throughout the trial, color-coded bars along the top of each graph indicate that the signal intensity (GLM) or classifier evidence (MVPA) for each category was above baseline. Activation baseline is mean signal intensity during rest, whereas information baseline is mean classifier evidence for *irrel* at each time point. Circles inside and outside these bars indicate that the value for one trial-relevant category was stronger than the value for the other trial-relevant category (small circles: p < 0.05; big circles: p < 0.002, Bonferroni corrected).
Neuropsychologia by Elsevier Science Ltd. Reproduced with permission of Elsevier Science Ltd in the format reuse in a book/textbook via Copyright Clearance Center.

Figure 1

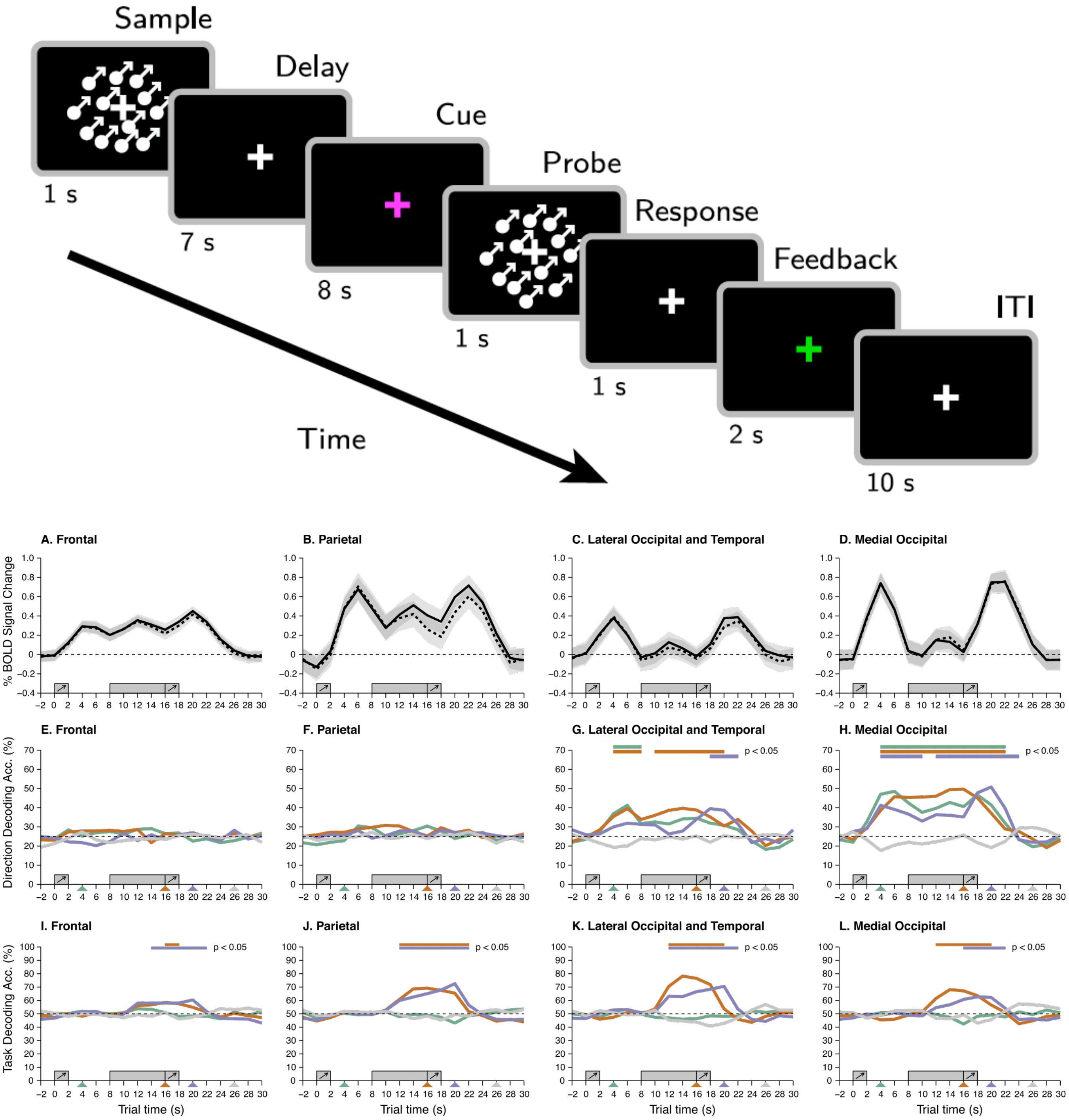
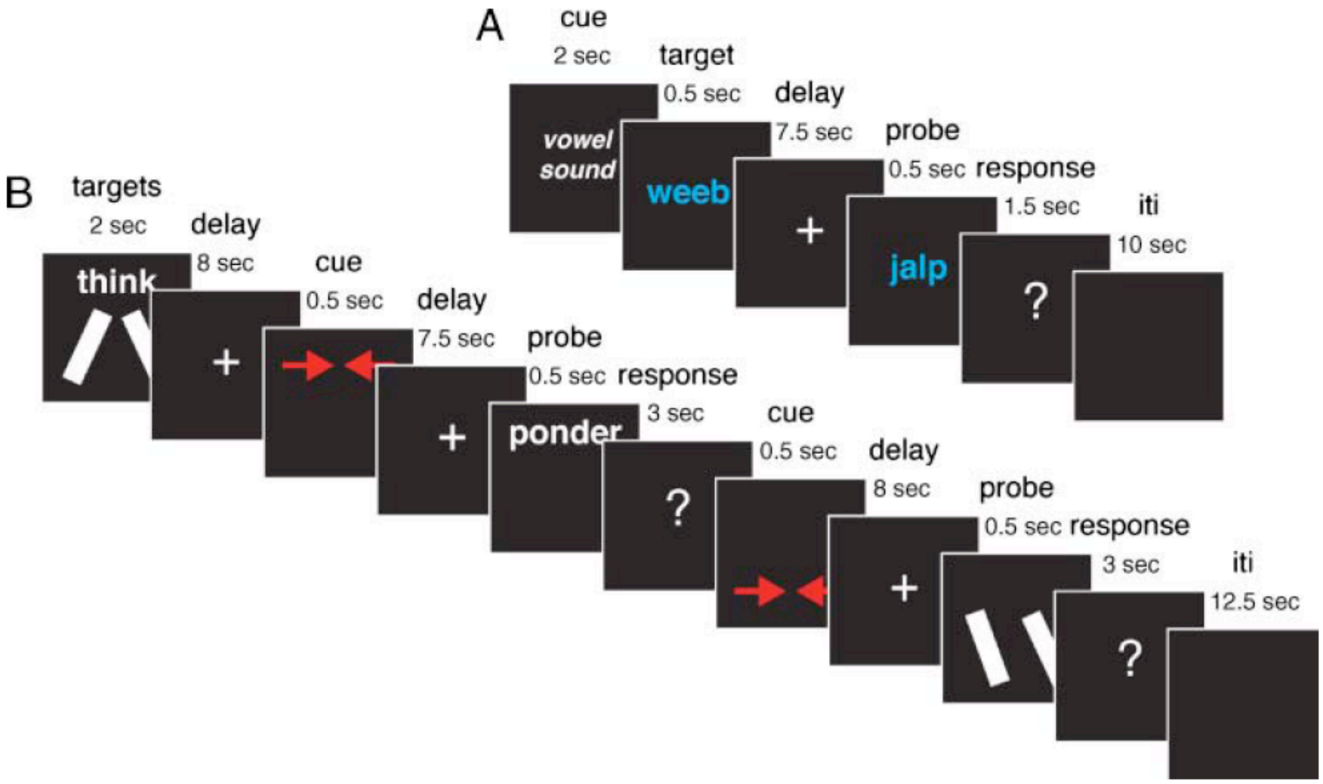

Sample
1 s
Delay
7 s
Cue
8 s
Probe
1 s
Response
1 s
Feedback
2 s
ITI
10 s
Time
A. Frontal
B. Parietal
C. Lateral Occipital and Temporal
D. Medial Occipital
% BOLD Signal Change
E. Frontal
F. Parietal
G. Lateral Occipital and Temporal
H. Medial Occipital
Direction Decoding Acc. (%)
p < 0.05
I. Frontal
J. Parietal
K. Lateral Occipital and Temporal
L. Medial Occipital
Task Decoding Acc. (%)
Trial time (s)
−2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30

Figure 2


A
cue
2 sec
vowel sound
target
0.5 sec
weeb
delay
7.5 sec
+
probe
0.5 sec
jalp
response
1.5 sec
?
iti
10 sec
B
targets
2 sec
think
delay
8 sec
+
cue
0.5 sec
delay
7.5 sec
+
probe
0.5 sec
ponder
response
3 sec
?
cue
0.5 sec
delay
8 sec
+
probe
0.5 sec
response
3 sec
?
iti
12.5 sec

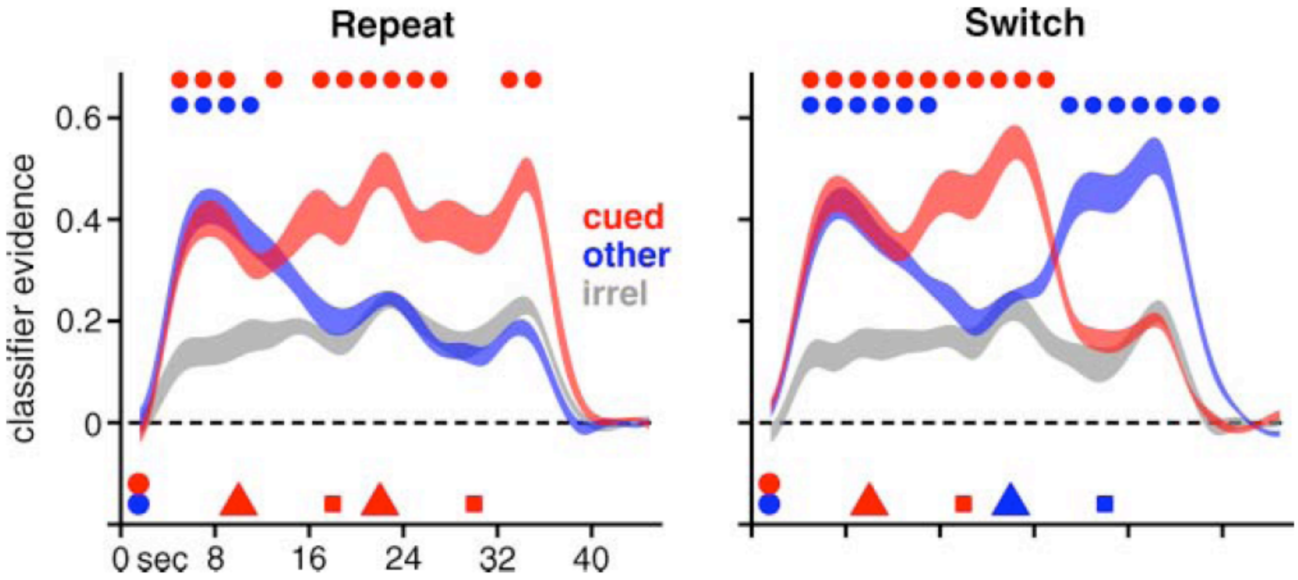

Repeat
Switch
classifier evidence
0.6
0.4
0.2
0
cued
other
irrel
0 sec
8
16
24
32
40

Figure 3

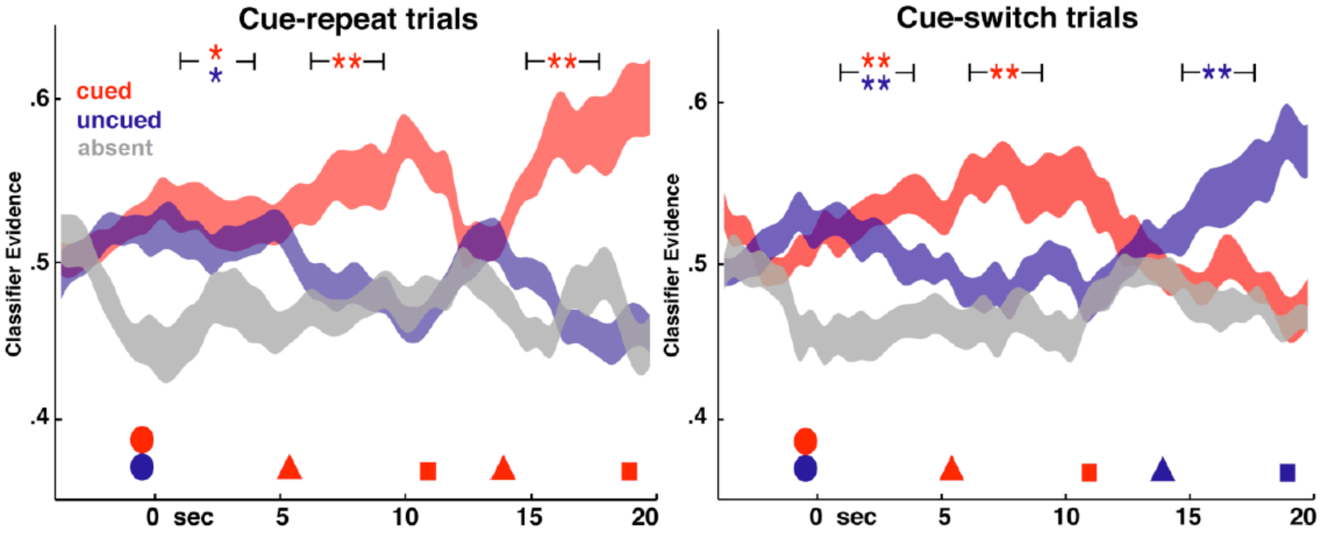
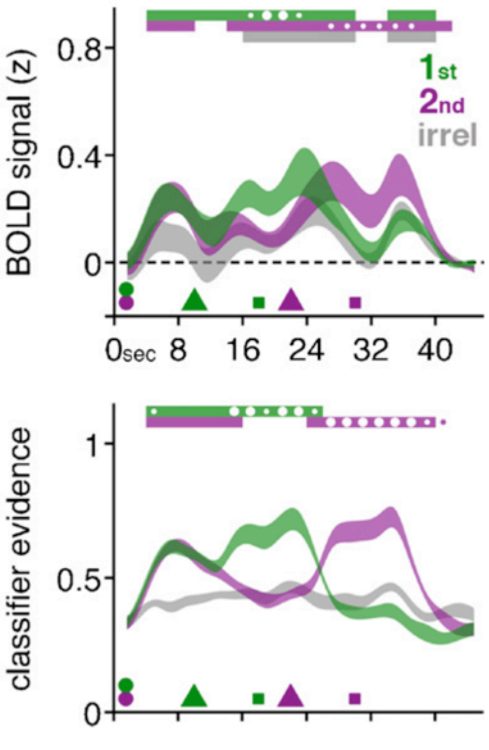


Figure 4



Endnotes

[ii] To be sure, connectivity analyses are increasingly common in the neuroimaging literature. Many studies employing such analyses, however, nonetheless draw on assumptions of signal intensity and modularity. It is also the case that MVPA can be applied in a manner that makes the assumption of modularity. This will be addressed in the section on *Implications and practical considerations*.

[ii] As an aside, it is worthy of note that although MVPA has been applied successfully to sensory processing in topographically organized cortex (e.g., as in the decoding of orientation (Harrison and Tong, 2009; Serences et al., 2009)), it has also been successfully applied to "higher level" processing in polymodal cortex. Thus, for example, MVPA has demonstrated contextual reinstatement during episodic memory retrieval (Polyn et al., 2005), the recognition of individual faces (Kriegeskorte et al., 2007), and neural correlates of free choice(Soon et al., 2008), all entailing the decoding of information from polymodal temporal, parietal, and/or frontal cortex.

[iii] All-the-while acknowledging, of course, that the blood oxygen level-dependent (BOLD) signal reflects a hemodynamic response to the cellular activity that we really care about.

[iv] Behavioral data from Oberauer (2005) and from LaRocque et al. (2013) indicate that subjects remove items from the focus of attention even during the second delay period of this multi-step task, when they know that $p = .5$ that the uncued memory item will be cued by the second retrocue.

[v] Note that in a subsequent study from this group, Lepsien and colleagues (2011) observed that delay-period signal in ROIs "preferentially responsive to face or scene stimuli" -- in fusiform and parahippocampal gyri, respectively – was not sensitive to memory load, leading them to call into question the role of these regions in STM maintenance, per se.

[vi] Here again I'll note that this author is among the many, many, cognitive neuroscientists who have published studies applying this logic (e.g., Postle et al., 2003; Postle, 2006; Postle and Hamidi, 2007). Thus, I want to make clear that my intent here is not to create the impression that the two groups whose papers are being examined here are in any way noteworthy for applying the signal-intensity assumption in their neuroimaging research. It is simply the case that the their papers have high thematic overlap with those of Lewis-Peacock et al. (2012) and LaRocque et al. (2013), which form the backbone of this chapter.

[vii] A complication with applying sparse machine learning techniques to neuroimaging datasets, in turn, is that the weighting of features (for fMRI, voxels) that satisfy sparcity goals of a particular algorithm may produce a solution that distorts the "true" anatomical distribution of information in the brain. That is, care must be taken when interpreting the anatomical distribution of "importance maps" that result from MVPA.