How does the brain keep information "in mind"?

Journal:	*Current Directions in Psychological Science*
Manuscript ID	CDPS-15-0288.R1
Manuscript Type:	Invited Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Postle, Bradley; University of Wisconsin Madison, Psychology and Psychiatry
Keywords:	working memory, fMRI, multivariate pattern analysis, prefrontal cortex, short-term memory
Abstract:	Working memory, the ability to hold information"in mind", to transform it as needed, and to use it to guide behavior, is critical for many domains of cognition, including planning, problem solving, and language production and comprehension. Like many aspects of cognition, working memory does not map to one or a few anatomically localizable "systems", nor, indeed, to any neurally derived signals that are readily accessible to "first order" analyses, such as the visual inspection of time series data. The relatively recent advent of applying of multivariate analysis methods to working memory data sets is providing strong evidence that such mechanisms as "sensorimotor recruitment" and the "temporary activation of representations from long-term memory" provide the best accounts of how the brain keeps information in mind.

How does the brain keep information "in mind"?

Bradley R. Postle

Departments of Psychology and Psychiatry

University of Wisconsin-Madison

1202 West Johnson St.

Madison, WI 53726

postle@wisc.edu

The ability to "hold information in mind" – to think about it – in the absence of steady input from the outside world, is central to our conscious experience. It also gives rise to working memory, the ability to flexibly use this information to guide behavior, often after it has been juggled or otherwise transformed. Working memory is critical for many domains of cognition, including planning, problem solving, and language production and comprehension. Working memory performance is also an important factor underlying individual differences across a broad spectrum of experimental and "real world" measures of human performance, and its dysfunction is characteristic of many psychiatric and neurological diseases. For all these reasons, understanding the neural bases of the short-term retention (STR) of information (a.k.a. "working memory storage") has been a priority for decades.

A point of emphasis for this review will be the importance of distinguishing between the *STR* of information, a phenomenon that is, in principle, measurable in the brain, and the construct of *working memory*, which is a label referring to a class of behavior (and its associated experimental tasks) that draws on many cognitive processes, among them the STR of information, but also selective attention, decision making, and many others. Indeed, as we shall see, the assumption that working memory maps to one or a few anatomically localizable "systems", or, indeed, to a neurally derived *signal* that is readily accessible to visual inspection, can lead to a lack of conceptual clarity and, sometimes, to erroneous inferences about neural functioning.

The concept of "activation" in working memory

Contemporary thinking about how the brain accomplishes the STR of information has focused on the phenomenon of persistent elevated neuronal activity, an idea that can be traced back at least as far as Hebb (1949), who postulated that reverberatory activity between the neurons involved in the perception of information is necessary for the STR of that information until it can be

encoded via synaptic reorganization into a long-term trace. Physiological evidence for persistent activity that could be the basis for such a transient trace began to emerge from recordings from the prefrontal cortex (PFC) in the 1970s (reviewed in Postle, 2015c). Although such activity could, in principle, correspond to many different functions (e.g., Curtis and Lee, 2010), its explicit linkage with the construct of working memory has proven to be potently influential during the past quarter century of cognitive neuroscience research.

A Synthesis of Cognitive Theory and Neuroscience data

Contemporaneous with, but independent of, developments in neurophysiology, experimental psychologists were developing a model of *working memory* as a multicomponent cognitive system comprised of storage buffers for different kinds of information, and a Central Executive that controlled the access of information to the buffers, and the interactions of the buffers with other cognitive systems (e.g., Baddeley, 1986). The model for working memory that has dominated the past quarter century of cognitive neuroscience research came about when neurobiologist Patricia Goldman-Rakic proposed that sustained delay-period activity in the PFC of monkeys performing delay tasks, and the storage buffers of the multicomponent model, were cross-species manifestations of the same fundamental mental phenomenon (e.g., Goldman-Rakic, 1987). Seminal studies from Goldman-Rakic's group typically followed a two-stage procedure characteristic of sensory neuroscience: First, determine the tuning properties of a neuron (e.g., *to what locations in the visual field does its firing rate increase?*); second, study that neuron's delay-period activity during trials when the animal is remembering that neuron's preferred vs. non-preferred information. The canonical finding was that neurons in PFC had the property of "memory fields", an analogy to "receptive fields" of sensory neurons, suggesting that each neuron was tuned for the STR of information of a particular kind. Furthermore, the data suggested a

topography of memory fields that mirrored the functional organization of the visual system, such that dorsolateral PFC was proposed to be the neural substrate of working memory for "where" to-be-remembered information was located, and ventrolateral PFC the neural substrate for "what" information was being remembered.

Neuroimaging of working memory, 1990 – 2010, and the "signal intensity assumption"

As neuroimaging methods became available to cognitive neuroscientists, they adopted the core assumptions that governed the study of working memory in nonhuman primates. Of particular relevance here is what can be called the "signal intensity assumption", which boils down to the reasoning that one can infer the active representation of a particular kind of information from the signal intensity in an area of the brain whose function is known a priori. This is perhaps most clearly illustrated in the widespread practice of using "functional localizers" to identify putatively category-specific regions of the brain. For example, the "fusiform face area" (FFA) is 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, whereas the converse is true for the "parahippocampal place area" (PPA). A working memory study might take advantage of this knowledge by first identifying the FFA and PPA with a scan that presents alternating visual presentation of faces vs. houses, then assessing how activity in these regions of interest (ROIs) varies during a test of working memory for faces vs. houses. In such a study, the neural correlates of the STR of face vs. scene information would be inferred from the fact that delay-period activity in a FFA ROI was greater on trials requiring the STR of face information, and the converse would be true for the PPA ROI (see (Postle, 2015a) for more detail).

Even stronger evidence was inferred from "load sensitivity", when the delay-period activity in an ROI scaled monotonically with the amount of information being retained. The logic here was

simply that when a system has to retain more information, it must have to "work harder", and this should be reflected in a level of activity that increases and decreases with each additional item that is added to or taken from the set that must be retained. The property of load sensitivity has been central to debates over the role of various regions in working memory functions (reviewed by Postle, 2006).

Accommodating the high dimensionality of brain function raises challenges for the "signal intensity assumption", and for memory-systems models of working memory

The past few years have witnessed dramatic and fast-moving changes in our understanding of the neural bases of the STR of information, many of these driven by the introduction to cognitive neuroscience of methods from statistical machine learning, often variants of multivariate pattern analysis (MVPA). As a result, many presumed physiological "signatures" of the STR of information are being reinterpreted as more general, state-related changes that can accompany cognitive-task performance, and theoretical models are being rethought.

Conceptual problems with the "signal intensity assumption"

Most neuroscientists would endorse the broad generalization that neural representations are high-dimensional, and supported by anatomically distributed, dynamic computations. Prior to the past decade, however, data from most human neuroimaging and nonhuman neurophysiological studies have been analyzed within a fundamentally univariate framework that, in retrospect, is at odds with how we think that the brain works. The "functional localizer" approach, for example, often leads to the identification of elevated (or decreased) signal intensity in voxels occupying a several-cubic-millimeter (or larger) volume of tissue, and, in effect, assumes that these voxels are all "doing the same thing". Furthermore, the interpretation of this cluster of similarly activated voxels often entails a third, albeit often implicit, assumption, which is that this

locally homogenous activity can be construed as supporting a mental function independent of what's happening in other parts of the brain. One can see from this summary that a signal intensity-based analysis is constrained, a priori, to only be capable of supporting hypotheses that brain functions (like working memory) are organized in a modular manner. It follows from this that models of working memory as a cognitive system supported by the PFC may be largely a consequence of how the data from working memory experiments have been analyzed.

A direct comparison of MVPA vs. Signal Intensity-Based analyses

To illustrate how MVPA can lead to a different conclusion about how working memory is supported in the brain, let's consider a task in which, on each trial, a subject is asked to remember two items drawn from two of three possible categories: words, nonwords, and visual patterns. (At the end of each trial, the subject will be asked to decide whether a probe matches the meaning (for words), the phonology (for nonwords), or the shape (for patterns) of one of that trial's sample stimuli.) For a signal intensity-based analysis, one would first define brain areas (i.e., functional ROIs) presumed to be specific to the processing of each category by identifying voxels whose response to, say, nonwords, is statistically greater than is their response to words or to patterns. Next, one might ask the (inherently 1-D) question of "does the signal intensity within his "nonword-specific" ROI increase above baseline levels during the delay period of working-memory-for-nonwords trials?" Importantly, because the signals from all the voxels within an ROI are pooled together, there is only one spatially averaged signal whose variation is being assessed. MVPA, on the other hand, doesn't assume that an element in the data set (i.e., a voxel or a neuron) "only does one thing". Rather, in our example it assigns each voxel a 3-D value, each dimension corresponding to that voxel's level of activity for each of the three conditions of the experiment. Then, rather than treating the ROI as a single entity, it assesses whether the pattern of activity across all the voxels in the ROI is statistically discriminable for words vs. nonwords vs. patterns.

(Successful discrimination is often referred to as "decoding".)

The consequences of treating an ROI as a high-dimensional versus a low-dimensional data set were illustrated empirically by Lewis-Peacock and Postle (2012) with a task that required working memory across two delay periods. The presentation of two stimuli (say, a word and a nonword) was followed by a retrocue indicating which of the two would be the first to be probed. After the first probe, a second retrocue indicated (with equal probability) whether the second probe would test working memory for the same item or the other one. Thus, our task assessed the effects of switching attention between items held in working memory. The striking finding from the MVPA was that the STR of any two stimulus categories could be decoded from the ROI that was putatively "specific" for the third (Figure 1). This means that the MVPA's superior sensitivity demonstrated that the presumption of specificity of the signal intensity-based analysis was invalid.

Multivariate Analyses Break the Conceptual Link between Elevated Delay-Period Activity and "Storage"

The theoretical consequences of MVPA and other multivariate methods have been dramatic. In the domain of visual working memory, the successful decoding of delay-period stimulus identity from early visual cortex, including V1, despite the absence of above-baseline delay-period activity, has given strong support to *sensorimotor recruitment* models, whereby the same systems and representations that are engaged in the perception of information can also contribute to its STR. Importantly, related studies have also generated evidence that is difficult to reconcile with memory systems models. For example, although the STR of specific directions of motion is decodable from medial and lateral occipital regions (despite the absence of elevated

delay-period activity), this information is not decodable from regions of intraparietal sulcus and frontal cortex (including PFC) that nonetheless evinced robust, load-sensitive delay-period activity (Emrich et al., 2013). Because MVPA features superior sensitivity and specificity, it must be the case that working memory-related fluctuations in signal intensity that are observed PFC and parietal regions reflect some process(es) other than memory storage per se, perhaps attentional control, or some other more general aspect of neurophysiological state, such as cortical excitability or inhibitory tone (reviewed in Postle, 2015b).

The above-summarized study of Lewis-Peacock and Postle (2012) was drawn from a series of studies assessing a model of the STR of information that is distinct from, and mutually compatible with, sensorimotor recruitment: the *temporary activation of representations from long-term memory (LTM)*. In the first study in this series, Jarrod Lewis-Peacock trained MVPA classifiers to discriminate neural activity associated with judgments that required accessing information from LTM about three categories: the likability of famous individuals; the desirability of visiting famous locations; the recency with which familiar objects had been used. Next, outside the scanner, subjects were taught arbitrary paired associations among items in the stimulus set. Finally, subjects were scanned a second time, but this time while performing delayed recognition of paired associates (i.e., see one item from the LTM memory set at the beginning of the trial, and indicate whether or not the trial-ending probe is that item's associate). The finding was that multivariate pattern classifiers trained on data from the first scanning session, when subjects were accessing and thinking about information from LTM, were successful at decoding the category of information that subjects were holding in working memory in the second scanning session

(Lewis-Peacock and Postle, 2008). Such an outcome could only be possible if the working memory task and the LTM task drew on the same neural representations.

Lessons from the past, directions for the future

Although the idea of PFC-as-working-memory-buffer has been potent and enduring, with the benefit of hindsight, it can be seen to be flawed on at least two levels. *Theoretically*, it is a conflation of the buffering with the Central Executive functions of the multicomponent model. Prior to the 1970s, decades of lesion studies had established that PFC was not necessary for the STR, per se, of information, but rather for controlling behavior, including when guided with information held in memory. Thus, for example, the integrity of the PFC was known to be crucial for controlling perseveration on tasks that periodically changed reward contingencies, for minimizing susceptibility to distraction and interference, and for mentally transforming information from the format in which it had been presented (reviewed in Postle, 2015c). Therefore, to record in the PFC was to "listen in on" the Central Executive, not a memory buffer. *Analytically*, this idea is flawed due to its reliance on the signal-intensity assumption. Although sustained delay-period activity in individual neurons (or voxels, or collections of EEG eletrodes) may correspond to our intuitions about how a STR signal should behave, multivariate population-level analyses are demonstrating analytically that our intuitions can mislead us. Indeed, in recent years, the application of "retrospective multivariate" analyses to datasets that were collected under the signal-intensity assumption have yielded reinterpretations of PFC delay-period activity that, tellingly, align better with conceptualizations of PFC that were prevalent in the 1960s, prior to the advent of contemporary neurophysiological and neuroimaging methods, than with those that have held sway over much of the past quarter century (reviewed in Postle, 2015b).

The approach taken here may have broader implications for the study of brain-behavior relations. From a theoretical perspective, the emphasis from neuroscience on dynamical systems analysis (e.g., (Buzsaki, 2006, Shenoy et al., 2013); toward which MVPA can be construed as baby steps) suggests that many classes of behavior might be better construed as emergent, rather than the output of systems that are dedicated to an a priori defined set of functions. Analytically, it is clear that first-order, intuition-based interpretations of fluctuating signal intensity will often be misleading.

References:

Baddeley, A. D. (1986). Working Memory. London, Oxford University Press.
Buzsaki, G. (2006). Rhythms of the Brain. New York, Oxford University Press.
Curtis, C. E. and D. Lee (2010). "Beyond working memory: the role of persistent activity in decision making." Trends in Cognitive Sciences **14**: 216-222.
Emrich, S. M., A. C. Riggall, J. J. Larocque and B. R. Postle (2013). "Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory." The Journal of Neuroscience **33**(15): 6516-6523. PMCID: 3664518
Goldman-Rakic, P. S. (1987). Circuitry of the prefrontal cortex and the regulation of behavior by representational memory. Handbook of Neurobiology. V. B. Mountcastle, F. Plum and S. R. Geiger. Bethesda, American Physiological Society**:** 373-417.
Hebb, D. O. (1949). The Organization of Behavior: A Neuropsychological Theory. New York, NY, John Wiley & Sons, Inc.
Lewis-Peacock, J. A., A. T. Drysdale, K. Oberauer and B. R. Postle (2012). "Neural evidence for a distinction between short-term memory and the focus of attention." Journal of Cognitive Neuroscience **24**: 61-79. PMCID: 3222712
Lewis-Peacock, J. A. and B. R. Postle (2008). "Temporary activation of long-term memory supports working memory." The Journal of Neuroscience **28**: 8765-8771. PMCID: 2699183
Lewis-Peacock, J. A. and B. R. Postle (2012). "Decoding the internal focus of attention." Neuropsychologia **50**: 470-478. PMCID: 3288445
Postle, B. R. (2006). "Working memory as an emergent property of the mind and brain." Neuroscience **139**: 23-38. PMCID: 1428794
Postle, B. R. (2015a). Activation and information in working memory research. The Wiley-Blackwell Handbook on the Cognitive Neuroscience of Memory. A. Duarte, M. Barense and D. R. Addis. Oxford, U.K., Wiley-Blackwell**:** 21-43.
Postle, B. R. (2015b). "The cognitive neuroscience of visual short-term memory." Current Opinion in Behavioral Sciences **1**: 40-46; doi:10.1016/j.cobeha.2014.1008.1004.
Postle, B. R. (2015c). Neural bases of the short-term retention of visual information. Mechanisms of Sensory Working Memory: Attention & Performance XXV. P. Jolicoeur, C. LeFebvre and J. Martinez-Trujillo. London, U.K., Academic Press**:** 43-58.

Shenoy, K. V., M. Sahani and M. M. Churchland (2013). "Cortical control of arm movements: A dynamical systems perspective." Annual Review of Neuroscience **36**: 337-359.

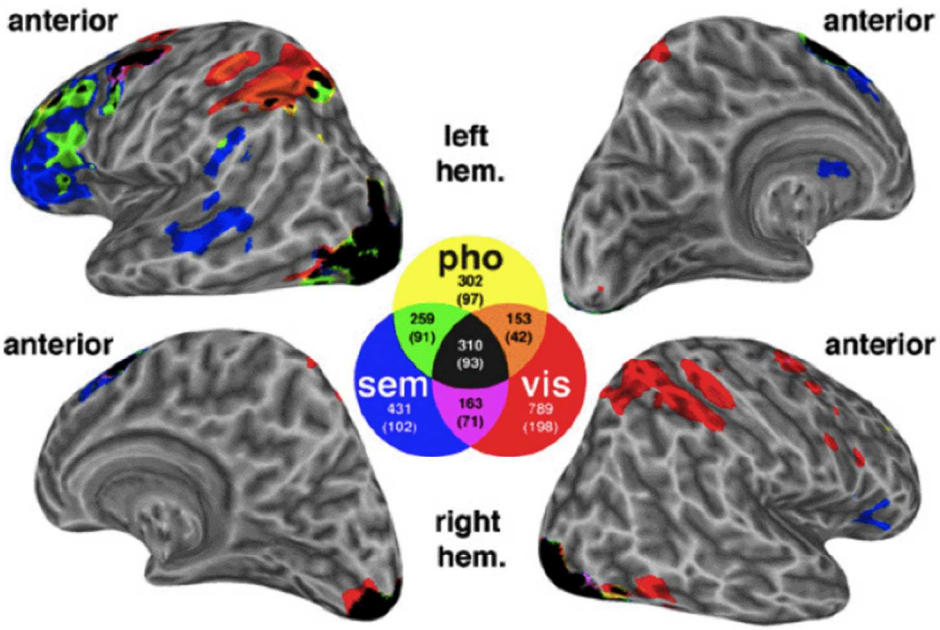

Figure 1.a. Functional ROIs derived from delay-period activation maps, for nonword (a.k.a. phonological, "pho"), word (a.k.a. semantic, "sem"), and visual (vis) stimuli, from data first reported in Lewis-Peacock et al., (2012). Although this composite, thresholded, group-average mapwas generated for illustration purposes, the analyses illustrated in Figure 1.b. were carried out in subject-specific ROIs. The color-coded venn diagram illustrates the mean number of voxels for each category (plus the overlapping voxels between categories).
296x203mm (72 x 72 DPI)

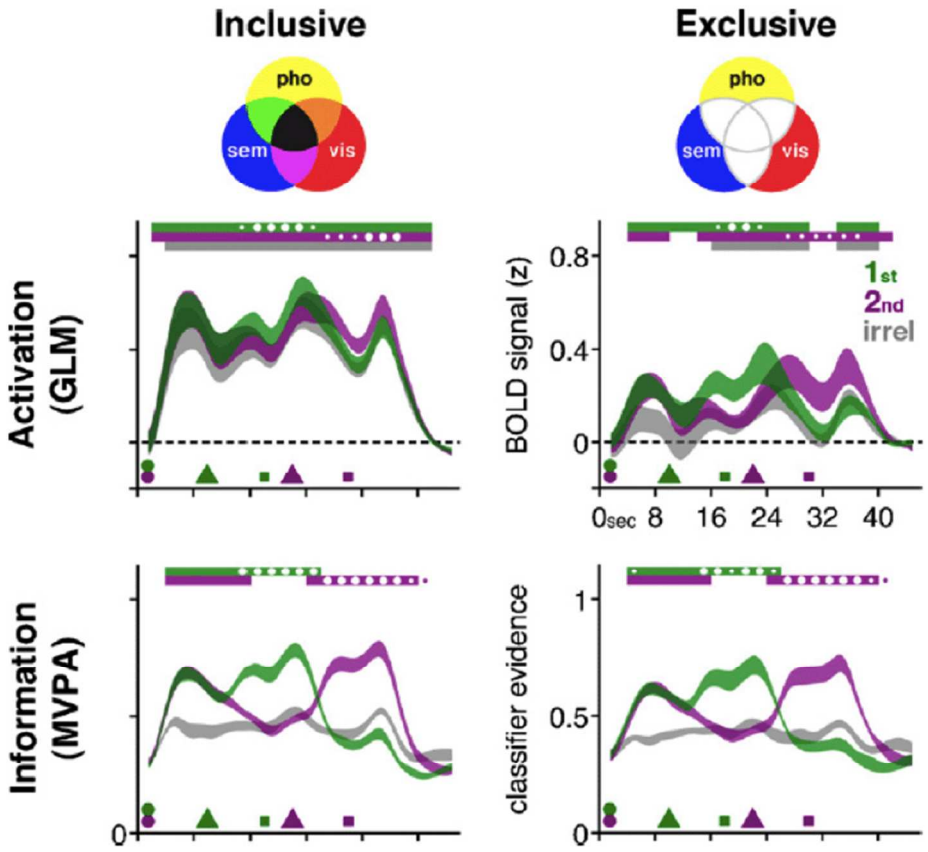

Figure 1.b. Time series data from the functional ROIs from Figure 1.a. Venn diagrams in top row illustrate the construction of "inclusive" and "exclusive" ROIs. Trial-averaged BOLD "activation" data (from these univariate general linear model ("GLM")-defined ROIs) are illustrated in the middle row. Each trial presented an item from each of two of the three categories, and the data have been collapsed across stimulus category and are plotted as a function of the category that was cued as relevant for the first ("1st") memory probe, for the second ("2nd") memory probe, or that was not presented on that trial (and was, therefore, irrelevant ("irrel") on that trial). Trial-averaged MVPA decoding of these same data is illustrated in the third row. The time series data are displayed as ribbons whose thickness indicates +/- 1 SEM across subjects. Symbols along the time axis correspond to stimulus presentation (circles), retrocuing (triangles) and probes (squares). 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 when the value for the corresponding time series differs from baseline; and circles inside and outside these bars indicate when the value for one trial-relevant category is higher than the value for the other trial-relevant category (small circles: $p < 0.05$; big circles: $p < 0.002$, Bonferroni corrected). For "Activation" time series data, baseline is mean signal intensity during intertrial interval; for "Information" time series data, baseline is mean classifier evidence for irrel category at each time point. Note that although subjects performed an equal number of trials in which the second retrocue unpredictably cued the initially cued or the initially uncued memory item, trials from only the latter condition are shown here. Versions of these figures first appeared in Lewis-Peacock & Postle (2012).
296x257mm (72 x 72 DPI)