# Revisiting the role of persistent neural activity during working memory

**Kartik K. Sreenivasan**[1], **Clayton E. Curtis**[2], **and Mark D'Esposito**[3]

[1] Division of Science and Mathematics, New York University Abu Dhabi, 19 Washington Square North, New York, NY 10011, USA
[2] Department of Psychology, and Center for Neural Science, New York University, 6 Washington Place, New York, NY 10003, USA
[3] Helen Wills Neuroscience Institute, and Department of Psychology, University of California, Berkeley, 132 Barker Hall, Berkeley, CA 94720, USA

**What are the neural mechanisms underlying working memory (WM)? One influential theory posits that neurons in the lateral prefrontal cortex (lPFC) store WM information via persistent activity. In this review, we critically evaluate recent findings that together indicate that this model of WM needs revision. We argue that sensory cortex, not the lPFC, maintains high-fidelity representations of WM content. By contrast, the lPFC simultaneously maintains representations of multiple goal-related variables that serve to bias stimulus-specific activity in sensory regions. This work highlights multiple neural mechanisms supporting WM, including temporally dynamic population coding in addition to persistent activity. These new insights focus the question on understanding how the mechanisms that underlie WM are related, interact, and are coordinated in the lPFC and sensory cortex.**

## Introduction

WM comprises the set of operations that support the active retention of behaviorally relevant information over brief intervals. Given the central role of WM in goal-directed behavior, establishing the neural basis of WM has been a priority of neuroscience research. Early WM studies observed that selective increases in neural activity during the presentation of a to-be-maintained sample item persisted throughout the blank 'delay' interval of a WM delay task, bridging the temporal gap between the sample and the subsequent contingent response [1,2]. This work inspired the theoretical framework that has predominated in the field: neurons or neuronal populations that are selectively tuned to the to-be-remembered information hold this information in an active state through persistent activation [3]. We refer to this model, which emphasizes stable persistent neural activity (see Glossary) in selective neurons as the fixed-selectivity model. Motivated by this model, functional MRI (fMRI) studies in humans and electrophysiological studies in monkeys have consistently identified persistent neural activity in the lPFC, leading many to conclude that the lPFC stores representations of WM memoranda.

A decade ago, we provided a critique of the literature on persistent activity in the context of contemporary models of prefrontal cortical function [4]. We hypothesized that, in contrast to existing theories of WM, persistent lPFC activity signifies attention directed to internal representations maintained in sensory cortices. Viewed through the lens of the fixed-selectivity model, evidence for this proposal is limited. Studies of sensory and motor function, however, suggest that information is likely to be represented through the combined activity of neural populations with diverse tuning properties rather than individual highly-tuned neurons [5,6]. This notion offers a promising framework for understanding WM.

In recent years, analytic and methodological advances (Box 1) have expanded researchers' ability to capture the multivariate nature of population coding and the causal relationships between neural activity and behavior. The findings generated using these approaches underscore the need for a revision of existing views of WM. In light of these results, we revisit the issue of how information remains active during WM. The studies we discuss here focus on

### Glossary

**Delay tasks:** the experimental paradigm typically used to study the neural basis of working memory (WM). A trial in a delay task begins with a brief presentation of a sample item. The subject encodes this item into WM and maintains this item over a blank 'delay' period of a few to several seconds. At the end of the delay period, a probe stimulus appears and the subject initiates a behavioral response contingent on the WM representation of the sample item. A key feature of delay tasks is that they temporally segregate subcomponents of WM such as stimulus encoding, storage, and retrieval/response.

**Persistent neural activity:** above-baseline neural activity that remains stable and elevated during a trial of a delay task. Persistent neural activity begins during the sample presentation and persists throughout the delay period, returning to baseline at the end of the trial. According to the fixed-selectivity model of WM (see main text), persistent neural activity in neurons selective for WM memoranda is the mechanism by which WM information is actively maintained.

**TMS:** transcranial magnetic stimulation.

**Voxel:** the spatial unit for measuring changes in blood-oxygenation-level dependent (BOLD) signal with fMRI. A voxel is a 3D volumetric pixel, typically of the order of 3 mm$^3$. BOLD signal within a voxel is an indirect measure of the summed activity of many tens of thousands of neurons. A single whole-brain fMRI image can comprise 60,000–100,000 voxels.

**Working memory:** the set of operations that support the ability to maintain information in an active state, to manipulate that information, and to use that information to guide behavior. WM is essential for several aspects of

---

*Corresponding author:* Sreenivasan, K.K. (kartik.sreenivasan@nyu.edu).

*Keywords:* working memory; prefrontal cortex; top down; decoding; forward encoding; MVPA; fMRI.

1364-6613/$ – see front matter

http://dx.doi.org/10.1016/j.tics.2013.12.001

visual WM, but the general principles discussed herein apply to WM in other modalities.

### Box 1. Methodological advances

Here we briefly describe analytic and methodological advances that have furthered our understanding of the neural basis of WM. The reader is encouraged to seek out some of the excellent reviews on these approaches (referenced below) for more details.

*Decoding analysis*

Unlike standard univariate analyses, which independently examine data from individual neurons or voxels for differences across conditions, multivariate decoding methods consider data from several neurons or voxels at once to identify patterns of activity that encode task-related information [10–13]. This technique uses machine learning algorithms to decode, or categorize, unlabeled test data using labeled training data. Successful (above-chance) decoding signifies that the activity pattern entered into the algorithm differs between the categories of interest, implying that the underlying neural activity encoded information about these categories. The chief advantage of this approach is potentially increased sensitivity [78]. However, patterns of neural activity may reliably distinguish between conditions for various reasons, some of which are not anticipated by the experimental design [18,79]. Caution is therefore required when interpreting the nature of the information identified via decoding analysis [80].

*Forwarding encoding models*

Conversely, encoding models predict fMRI activity from task conditions [81–83]. These models rely on *a priori* assumptions about the features of task conditions that will result in changes in the hemodynamic response. In WM studies, forward encoding models of visual cortex have been constructed using knowledge about tuning for visual features [17]. Neural activity in hypothetical populations of neurons (channels) tuned to different values in feature space can be reconstructed from training data by estimating the degree to which each voxel's response contributes to a given channel. The critical advantage over decoding analyses is that this approach can predict fMRI responses to novel stimuli [14]. Encoding approaches are potentially more powerful for identifying information encoded in neural activity, but are constrained by the validity of the underlying assumptions of the model.

*TMS*

TMS uses magnetic fields to focally modulate cortical excitability [84]. In WM studies, TMS is used either offline to modulate cortical function for the duration of the experiment or online to modulate activity during specific epochs of a task. TMS effects on behavior or neural activity in distal regions can support strong causal inferences about the functional role of the regions targeted with TMS. Attenuation of TMS effects as a function of distance from the coil imposes restrictions on which brain regions can be targeted.

## Evidence for persistent WM representations in visual cortex

Neurons in visual cortex are selectively tuned to visual stimulus features and are consequently well suited for maintaining high-fidelity representations of visual information in the service of WM [7]. Yet, from the perspective of the fixed-selectivity model, evidence for sustained WM representations in visual cortex has been equivocal. Although sustained responses have been observed in temporal cortex [8], studies typically describe transient neural responses to sample stimuli without any subsequent sustained activation. Studies of early visual regions routinely note an absence of persistent activity [9].

Contemporary multivariate encoding and decoding statistical analyses (Box 1), however, consistently demonstrate that visual cortex does retain sensory WM representations. Decoding analysis applied to fMRI or electrophysiological data can identify activity distributed across neurons or neural populations that encodes task-relevant information [10–13]. By contrast, forward encoding models take advantage of assumptions about neural population tuning to reconstruct the response of hypothetical channels from fMRI voxels that represent the weighted sum of subpopulations of neurons tuned to these channels [14]. Both approaches can test whether feature or item information is encoded in the multivariate patterns of activity during WM, regardless of whether this activity exhibits sustained stimulus-selective responses during sample presentation that persist across the blank delay interval of the WM task. Studies incorporating these methods find that patterns of delay period activity in early visual cortex contain information about simple visual features held in WM [15–19] (Figure 1A,B). Similarly, delay patterns in occipital and temporal regions specialized for object representation encode actively maintained visual objects [20–24], consistent with studies that inferred a role for temporal cortex in WM storage on the basis of persistent neural activity in these regions [8].

Moreover, this work establishes four key properties of population coding of WM information in visual cortex. First, decoding and forward encoding analyses have extracted information specific to the contents of WM from visual cortex activity across multiple timepoints during the delay period [15,17,24,25], indicating that visual cortical WM representations persist throughout the period separating the visual stimulus and the contingent behavioral response. Second, given the limited capacity of WM [26], neural coding of sensory representations should prioritize task-relevant over task-irrelevant information [27]. Selectivity for task-relevant information was illustrated in a study where decoding based on the multivoxel pattern of delay period activity in early visual cortex was successful only for the task-relevant feature (orientation or color) of the memoranda [16]. Similarly, other work has shown that multivoxel patterns of delay period activity encode only items cued in the sample display as task relevant [15,24]. Third, and in contrast to the fixed-selectivity model, information about items maintained in WM can be encoded by neural populations that are not highly selective for the maintained stimuli. A recent fMRI study examined the degree to which decoding information about items maintained in WM was dependent on voxels that were highly selective for the WM items. The key finding was that removing highly selective voxels from the analysis did not substantially reduce the ability to decode information about the WM items [24]. These results are in line with studies demonstrating that perceptual [28] and motor [5] information is distributed across neural populations with diverse tuning preferences.

Fourth, and perhaps most importantly, fMRI measures of sensory representations in visual cortex are tied to the precision of WM representations. For example, one study found decreases in the ability to decode maintained directions of motion from multivoxel delay period activity as the number of to-be-maintained motion directions increased, possibly due to interference between spatially overlapping representations [29]. Reductions in the ability to decode the contents of WM predicted decrements in the precision