How do neural processes give rise to cognition?

Simultaneously predicting brain and behavior with a dynamic model of visual working memory

Aaron T. Buss[1], Vincent A. Magnotta[2], Will Penny[5], Gregor Schöner[3], Theodore J. Huppert[4], & John P. Spencer[5]

[1]Department of Psychology, University of Tennessee, Knoxville

[2]Department of Radiology, University of Iowa

[3]Institute for Neurocomputing, Ruhr University, Bochum, Germany

[4]Department of Radiology, University of Pittsburgh

[5]School of Psychology, University of East Anglia, Norwich, UK

Corresponding author information:

Aaron Buss
301C Austin Peay
Department of Psychology
University of Tennessee-Knoxville
Knoxville, TN 37917
865-974-3818

John P. Spencer
0.09 Lawrence Stenhouse Building
School of Psychology
University of East Anglia
Norwich NR4 7TJ
United Kingdom
+44-1603-593968

Abstract

There is consensus that activation within distributed functional brain networks underlies human thought. The impact of this consensus is limited, however, by a gap that exists between data-driven correlational analyses that specify where functional brain activity is localized using fMRI, and neural process accounts that specify how neural activity unfolds through time to give rise to behavior. Here, we show how an integrative cognitive neuroscience approach may bridge this gap. In an exemplary study of visual working memory, we use multi-level Bayesian statistics to demonstrate that a neural dynamic model simultaneously explains behavioral data and predicts localized patterns of brain activity, outperforming standard analytic approaches to fMRI. The model explains performance on both correct trials and incorrect trials where errors in change detection emerge from neural fluctuations amplified by neural interaction. Critically, predictions of the model run counter to cognitive theories of the origin of errors in change detection. Results reveal neural patterns predicted by the model within regions of the dorsal attention network that have been the focus of much debate. The model-based analysis suggests that key areas in the dorsal attention network such as the intraparietal sulcus play a central role in change detection rather than working memory maintenance, counter to previous interpretations of fMRI studies. More generally, the integrative cognitive neuroscience approach used here establishes a framework for directly testing theories of cognitive and brain function using the combined power of behavioral and fMRI data.

Keywords: visual working memory; change detection; fMRI; dynamic field theory

Although great strides have been made in understanding the brain using data-driven methods (Smith et al., 2009), to understand the brain's complexity, psychological and brain sciences will need sophisticated theories (Gerstner, Sprekeler, & Deco, 2012). *But what would a good theory of brain function look like?* (This question was posed in a July 11, 2014 *New York Times* Opinion Page by Gary Marcus: http://www.nytimes.com/2014/07/12/opinion/the-trouble-with-brain-science.html). Addressing this question requires theories that bridge the disparate scientific languages of neuroscience and psychology: we must create psychological explanations for behavior using neural process accounts, and neuroscientific theories of brain function that make sense of behavior. In short, bridge theories must explain what the brain is doing in real-time to generate specific patterns of neural *and* behavioral data (for related ideas see, O'Reilly, 2006).

Bridging brain and behavior may seem like a central goal in the psychological and brain sciences; however, this goal has rarely been directly realized. Many theories in psychology focus on cognitive processes with a primary goal of explaining behavioral data (Anderson et al., 2004; Bays, Catalao, & Husain, 2009; Brady & Tenenbaum, 2013). Other theories focus on neural processes with a primary goal of explaining neural data (Brunel & Wang, 2001; Deco, Rolls, & Horwitz, 2004; Domijan, 2011; Edin, Macoveanu, Olesen, Tegnér, & Klingberg, 2007; Raffone & Wolters, 2001). Rarely is the same model used to generate both behavioral and neural data, that is, simultaneously integrating both cognitive and neural processes (Wijeakumar, Ambrose, Spencer, & Curtu, 2016). This level of explanation is arguably the most critical, however, because it can explain *how neural processes give rise to cognition and behavior* (see Palmeri, Turner, & Love, 2017 for a special issue devoted to this topic).

To illustrate, consider the current state of theory within the domain of visual working

memory (VWM). VWM is central cognitive system used to remember visual information during short term delays and compare visual items that cannot be simultaneously foveated (for a review see Luck & Vogel, 2013). For instance, VWM is often probed in the change detection task (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988). In this task, participants are shown a memory array consisting of 1-8 objects (e.g., colored squares). After a brief delay (e.g., 1s), participants are shown a test array and asked to determine whether all the items are the same or different. Results from this task have revealed that VWM has a highly limited capacity. Although estimates vary across studies, it is generally accepted that people can store only 2-4 items in VWM at one time (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988; Rouder, Morey, Morey, & Cowan, 2011).

According to one prominent view, these capacity limits reflect the functioning of a memory system that stores a limited number of fixed-resolution representations in independent memory 'slots' (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988; Zhang & Luck, 2008). An alternative view holds that VWM is better conceived of as a shared resource that can be flexibly distributed among the items making up a scene, with no fixed upper limit on the number of items that can be stored (Bays et al., 2009; Bays & Husain, 2008; Wilken & Ma, 2004). There have been a host of recent modeling efforts designed to contrast these two perspectives using Bayesian approaches (e.g., Brady & Tenenbaum, 2013; Donkin et al., 2013; Kary et al., 2016; Rouder et al., 2008; Sims et al., 2012) and efforts to expand these views using drift diffusion models (Sewell, Lilburn, & Smith, 2016). In all cases, these studies use mathematical models to instantiate conceptual claims about VWM and test these claims at the level of behavior, typically using proportion correct, although some recent papers have also examined reaction times (Donkin et al., 2013; Sewell et al., 2016), VWM confidence (van den Berg, Yoo, & Ma, 2017),

feature chunking (Brady & Tenenbaum, 2013), and psychometric functions for difference detection (Sims et al., 2012) or feature estimation with models that do not have strict limits on slots or resources (Oberauer & Lin, 2017; Swan & Wyble, 2014). None of these models have been used to explain patterns of neural data, nor were they designed to do so.

Other theories of VWM have focused on the neural bases of this cognitive system. fMRI research shows that a distributed network of frontal and posterior cortical regions underlies change detection performance. VWM representations are thought to be actively maintained in the intraparietal sulcus (IPS), the DLPFC, the ventral-occipital (VO) cortex for color stimuli, and the lateral-occipital complex (LOC) for shape stimuli (Todd & Marois, 2004, 2005). In addition, there is suppression of the temporo-parietal junction (TPJ) during the delay interval, and activation of the ACC during the comparison phase (Mitchell & Cusack, 2008; Todd, Fougnie, & Marois, 2005). Moreover, there is greater activation of this network on change versus no change trials, and the hemodynamic response on error trials tends to be less robust (Pessoa, Gutierrez, Bandettini, & Ungerleider, 2002; Pessoa & Ungerleider, 2004).

Efforts to understand the theoretical bases of VWM at the neural level have focused on the biophysical properties that give rise to sustained activation—the putative neural basis of VWM representations (Constantinidis & Steinmetz, 1996; Fuster & Alexander, 1971; Miller, Erickson, & Desimone, 1996; Moody, Wise, di Pellegrino, & Zipser, 1998). There have been quite detailed biophysical accounts of how networks of neurons give rise to sustained activation. These models have been used to explain both neurophysiological data (Brunel & Wang, 2001; Compte, Brunel, Goldman-Rakic, & Wang, 2000) and, in some cases, aspects of fMRI signals (Deco et al., 2004; Domijan, 2011; Edin et al., 2007). Other models have explored the possibility that VWM representations are encoded in terms of neural synchrony across neuronal assemblies

(Raffone & Wolters, 2001), while recent work has also raised the possibility that working memory performance reflects the reactivation of representations from 'memory-silent' neural codes (Rose et al., 2016; Sprague, Ester, & Serences, 2016; cf., Schneegans & Bays, 2017). Although these models explain how neural processes can encode and maintain visual information, they have not been used to capture any behavioral data from VWM paradigms. This is not surprising. Biophysical models are computationally complex; thus, simulating behavioral performance across many iterations of the model is often not a realistic goal.

There are some models that have the potential to bridge the gap between brain and behavior. These models use variants of neuronal dynamics. For instance, Swan and Wyble (2014) proposed a model of VWM with some neural dynamics; however, these dynamics were discrete and activation levels were updated in one-shot steps at encoding and retrieval making a direct link to real-time neural measures not possible. Similarly, Oberauer and Lin (2017) proposed a model inspired by a connectionist network using the concept of neural activation; however, there was no attempt to simulate real-time neural dynamics directly. In both of these papers, the focus was solely on simulating behavioral data.

In summary, then, although understanding how the brain gives rise to behavior is clearly an important goal, this goal has been rarely addressed within the domain of visual working memory. We contend that research on VWM is not unique in this regard. Creating theories that bridge between these levels of analysis is fundamentally challenging as highlighted in a recent special issue on model-based fMRI (Turner, Forstmann, Love, Palmeri, & Van Maanen, 2016). Model-based fMRI is a promising approach to understanding human cognitive neuroscience that uses computational models of cognitive processes to link brain and behavior. Turner and colleagues reviewed the current state of the literature, highlighting many exciting approaches,

but they also revealed a fundamental challenge: very few approaches create a direct mapping between brain and behavior. This is what they call *integrative cognitive neuroscience* (ICN). The goal of ICN is to develop a model where one can tune parameters to achieve good fits to both brain and behavior and, reversely, that brain and behavioral measures can feed back to inform the quality of the model/theory.

We pursue an ICN approach here within the domain of VWM. We begin with a Dynamic Field Theory (DFT) of VWM that has shown promise by generating novel, *a priori* behavioral predictions that run counter to other cognitive models of visual working memory (Johnson, Ambrose, van Lamsweerde, Dineva, & Spencer, submitted; Johnson, Spencer, Luck, & Schöner, 2009). Critically, this theory also simulates neural population activation on a millisecond timescale and explains how neural activation in the brain is turned into a behavioral decision on each trial. This is not done using an algorithmic mapping of activation to behavioral measures; rather, the model actively generates a decision on each trial via the activation of a neural decision system engaged during the comparison process. Thus, in DFT there is not brain at one level and behavior at another. Rather, brain measures and behavioral outcomes both arise from neural population dynamics. The result is an integrative cognitive neuroscience (ICN) model that directly simulates both neural activation and behavior.

The goal of the paper is to test the DF model of VWM with fMRI. We do this first by simulating previous fMRI findings from the literature, simultaneously fitting the model to both behavioral and fMRI data. This yields an initial set of model parameters we can use to generate novel neural predictions. It also leads to a discovery: what was thought to be a neural signature of working memory – an asymptote at high memory loads – may actually be a neural signature of brain regions coupled to working memory rather than a signature of working memory *per se*.

Our model also explains why this asymptote does *not* occur in paradigms using a longer memory delay.

Next, we test a set of novel neural predictions generated by the DF model. One of the unique features of the model is that it specifies the neural processes that underlie both correct and incorrect trials in the change detection task (Johnson, Simmering, & Buss, 2014). Consequently, an optimal way to test the model is in a change detection task that has high numbers of correct and incorrect trials. Thus, we created a novel experiment that optimized participants' performance so they generated many errors, but maintained performance at above-chance levels. We then used this paradigm in a task-based fMRI study conducted using a 3T MRI scanner.

But how do we know if the DF model provides a good account of these data? Ideally, we would test the model against a competing theory of VWM; however, as our review above indicates, no other theory of VWM simultaneously predicts both neural and behavioral data. Thus, we tested the model against a standard statistical model. The idea here was simple: typically, fMRI data are analyzed using a general linear modelling (GLM) approach with regressors for each factor in the experiment. In order for the DF model to be useful, it should – at the very least – capture more variance than the standard statistical model. To evaluate this, we used Bayesian linear multi-variate modeling to evaluate the DF model's ability to capture data from 23 regions of interest (ROIs) relative to different variants of a task-based GLM. A Variational Bayes algorithm (Roberts & Penny, 2002) was then used to estimate the model evidence which takes into account model fit but also penalizes models for their complexity (Bishop, 2006). Finding the best model over a group of subjects was then implemented using Random Effects Bayesian Model Selection (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). Results show that the DF model

outperforms the standard statistical model. Further, the mapping of model components to ROIs provides a novel functional picture of how the brain implements VWM across a distributed network. Critically, this analysis reveals not only where VWM lives in the brain, but which brain areas implement which functions.

The paper is organized as follows. We first describe the theory we test, including background on the larger theoretical framework this theory is embedded within, Dynamic Field Theory. Next, we derive a mapping from neural activity in the model to hemodynamic responses measured with fMRI and contrast this with other approaches to model-based fMRI. Our objective here is to highlight how the dynamic field approach is an example of integrative cognitive neuroscience (Turner et al., 2016). We then ask if this approach yields useful information by simulating – for the first time – a key finding from the literature using a neural process model. We then generate a set of novel predictions and test them in an fMRI experiment, using a GLM-based approach to model testing. We conclude with an evaluation of our integrative cognitive neuroscience approach—have we achieved a model that effectively bridges between brain and behavior? We address this question by placing our approach within the context of the theoretical literature on VWM and contrasting our model with other psychological and neuroscience models in the field.

A Dynamic Field Theory of Visual Working Memory

The model we evaluate was developed within the framework of Dynamic Field Theory (Schoner, Spencer, & DFT Research Group, 2016). Thus, we begin with a brief review of the concepts of DFT. This theoretical framework has a long history in psychology and neuroscience dating back almost 30 years (Buss & Spencer, 2014, 2018; Buss, Wifall, Hazeltine, & Spencer, 2014; Erlhagen & Schöner, 2002; Kopecz & Schöner, 1995; Perone, Molitor, Buss, Spencer, &

Samuelson, 2015; Perone, Simmering, & Spencer, 2011; Schöner & Thelen, 2006; Schutte & Spencer, 2009; Schutte, Spencer, & Schoner, 2003; Simmering, 2016; Simmering & Spencer, 2008; Thelen, Schöner, Scheier, & Smith, 2001). Readers are referred to our recent book for a more complete introduction (Gregor Schoner et al., 2016).

Activity within populations of cortical neurons is hypothesized to be the best neural correlate of behavioral performance (Cohen & Newsome, 2008). Thus, we anchor our approach at this level. In particular, the theory we evaluate—a dynamic field theory (DFT) of VWM (Johnson, Spencer, Luck, et al., 2009; Johnson, Spencer, & Schöner, 2009)—simulates the activity of neural populations from millisecond-to-millisecond as the neural dynamic network engages in a particular working memory task.

A central issue in neural population dynamics is stability—how does a neural population stabilize a particular pattern through time (Amari, 1977; Grossberg, 1982; Wilson & Cowan, 1972). This can be formalized using the language of dynamical system theory. Specifically, one can think about how the activity of a neural population, u, changes through time, $\dot{u}$, as a function of its current state and other inputs to the population. These dynamics can be formalized as follows:

$$\dot{u} = -u + h \quad (1)$$

where $\dot{u}$ is the rate of change in activation through time, u is the current state of activation, and h is a collection of inputs to the field that, when summed, modulate the resting level of the population.

If we plot the phase portrait of this system, that is, a plot of the system in the space u by $\dot{u}$, we see that the system is a linear dynamical system (see red line in Figure 1A). There is a special place in this linear plot where $\dot{u} = 0$. If activation, u, is set to this value, then the rate of

change is 0 and the system will stay put—it won't change through time. This special place in the phase portrait is called an attractor. In equation 1, h is the attractor state – when activation reaches this value, the rate of change in activation is zero (if $u = h$, then $\dot{u} = 0$).

If we plot the behavior of this neural dynamic system through time, we can see that it stays near this attractor position. This is readily apparent when we add some neural noise to the equation, $\xi(t)$. For instance, in Figure 1B, we start the neural population at a random value near h and simulate the dynamics through time, adding a random value to the system at each time point (see x-axis). For the first 250 time steps, we keep h at the value -4 (see green line), and the system randomly wanders up and down, but always stays near h. After 250 time steps, we then boost h to the value -2 (see the magenta line in Figure 1A). This is like boosting the overall excitability of the neural population (a common form of neural interaction in the brain, see Bastian, Riehle, Erlhagen, & Schöner, 1998). The system jumps up to the activation value -2 (see Figure 1B), quickly finding the new attractor state. After another 500 time steps, we return h to the value -4. Again, the activation quickly moves to the new attractor state and stays around this value.

Although this captures some features of neural population dynamics, this simple dynamical system fails to capture that neural populations are inherently non-linear. For instance, neural populations often require a robust input to 'turn on', and once they are 'on', they are often 'sticky' – they stay 'on' even when there is relatively little input (e.g., see Hock, Kelso, & Schöner, 1993). This type of non-linearity can be captured by adding a sigmoidal function to the equation:

$$\dot{u} = -u + h + c*g(u) + \xi(t) \qquad (2)$$

Where

$$g(u) = 1 / (1 + \exp(-\beta(u))) \quad (3)$$

The sigmoidal function, $g(u)$, has 'output' that varies between 0 and 1. β defines the steepness of the transition from 0 to 1, and this function is typically centered around a threshold value of 0 activation. Thus, as activation, u, increases from a negative 'resting' level toward 0, the sigmoidal function starts producing positive output. At an activation value of 0, the sigmoidal function outputs a value of 0.5. And at higher positive activation values, the sigmoidal function saturates at an output of 1.0. Note that the 'output' of the sigmoidal function is multiplied by a connection strength, c, in equation 2.

To understand the consequence of this sigmoidal function, consider the phase portrait of this new system in Figure 1C when $h = -4$ (red line). Notice the S-shaped bend in the system as it approaches the value $u = 0$ (the threshold value). We can see that at negative values of u (when $g(u) = 0$), the system follows the equation $\dot{u} = -u + h$, while at large positive values of u (when $g(u) = 1$), the system follows the equation $\dot{u} = -u + h + c$. Importantly, however, there is still only a single attractor state at $h = -4$ (see black square). Consequently, this system will always stay near this attractor state. This is shown in Figure 1D. Note how the system behaves just like the linear system for the first 250 time steps.

Critically, when we boost h from -4 to -2 as before, the non-linear system goes through a bifurcation, that is, the attractor layout changes (see magenta line in Figure 1C). Now the system has two attractor states – one near -2 (the new 'resting' level defined by h) and one at +3 (the value $h + c$, where $c = 5$ in this example). Moreover, in between these two attractors is a repeller indicated by the diamond. Figure 1D shows that this changes how the neural population behaves through time. When the excitability of the neural population is boosted by raising h to -2, the system quickly moves to this new attractor state. However, after another 250 time steps (around

time point 500) the system jumps to the value $h+c$ and remains stably activated in this 'on' state through time. The behavior of this system inspires an analogy—the neural population has detected the presence of a weak input, and the system has kicked itself into an 'on' state. Note that this state is stable, but not permanent. For instance, once we decrease h back to the initial 'resting' value at time step 750 (see green line in Figure 1D), the activation eventually settles back to the original attractor state. This is reflected in Figure 1C—recall that at a low h value, there is only one stable attractor state.

This non-linear dynamical system captures several key properties of neural population dynamics (e.g., bi-stability; see Tegnér, Compte, & Wang, 2002); however, the system can only represent that something is present or absent (i.e., that activation is high or low). To enrich the system, we need to think about how to represent the *dimensions* within which the neural system is embedded. In DFT, this is done by thinking about the tuning curves of neurons in a population. Neurons in cortex are sensitive to particular types of information, typically in a graded way. For instance, some neurons are 'tuned' to spatial dimensions (Constantinidis & Steinmetz, 2001)—they prefer stimuli, say, to the left side of the retina. Other neurons are 'tuned' to color dimensions (Matsumora, Koida, & Komatsu, 2008; Xiao, Wang, & Felleman, 2003)—they like blue hues. Importantly, these tuning functions are typically quite broad (Wachtler, Sejnowski, & Albright, 2003); this means a color neuron will respond really vigorously to blue hues, but also quite a bit to cyan, and maybe even a bit to pink as well.

How do we incorporate these tuning functions into the neuronal dynamics picture? We can integrate these concepts using dynamic fields (DFs) where each neuron contributes its tuning curve weighted by its current firing rate to an activation field (Erlhagen et al., 1999). This tuning of neural units creates a direct link between activation fields in DFT and task dimensions varied

in experiments that has predicted a wide range of behavioral data (Buss & Spencer, 2014; Buss et al., 2014; Johnson, Spencer, Luck, et al., 2009). To make this concrete, let's start with 100 neural sites instead of just one. Each site will have the same neural dynamics as before; however, now that we have 100 neural sites, we have to think about how they are connected to one another across the cortical field. We will wire them up using a canonical lateral connectivity pattern with local excitation and surround inhibition (Amari, 1977; Compte et al., 2000; Wilson & Cowan, 1972), and the 'ordering' of sites along the represented dimension will be based on their tuning curves. This means that neurons that 'like' similar spatial locations or similar colors will pass strong, reciprocal excitation to one another because they are close together in the field, while neural sites that 'like' very different locations or colors will share reciprocal inhibition because they are far apart in the field. Mathematically, this can be summarized as follows (Amari, 1977; Wilson & Cowan, 1972):

$$\tau_e \dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int c_e(x - x')\, g\left(u(x',t)\right) dx' - \int c_i(x - x')\, g\left(u(x',t)\right) dx' + \xi(x',t) \quad (4)$$

Note the similarities to the neuronal dynamics in equation 2; however, now activation is distributed over the behavioral dimension, *x* (e.g., color). Similarly, inputs, *s(x,t)*, are distributed over *x*; thus, a red input (*x* = 25) is different from a blue input (*x* = 60). The laterally excitatory connections are defined by c_e (an excitatory Gaussian connection matrix), while the inhibitory connections are defined by c_i (an inhibitory Gaussian connection matrix). As before, these are convolved with the sigmoidal function, *g(u)*. This means that only above-threshold sites in the field contribute to neural interactions, that is, to local excitation and surround inhibition. Neural interactions for each location, x, are evaluated relative to every other position in the field, x'. Lastly, τ_e specifies the timescale over which excitation evolves in the field.

To understand the consequences of the lateral connectivity in a dynamic field – how neural sites talk to one another based on their neural tuning – it is useful to first plot activation with connectivity and the sigmoidal function turned off. Figure 1E shows the same type of simulation as in Figure 1A-B where we start with low excitability, then boost excitation locally, and then return to a lower resting level. Now, however, we do the boosting by giving a color input to the field centered at value 25 (see grey 'shadow' along the feature axis). Specifically, the input is off for 250 time steps, then on for 500 time steps, and then weaker for the last 250 time steps. As can be seen in Figure 1E, the activation in the dynamic field just mimics the input through time (see light grey 'shadow' projected along the back wall of the image). Thus, without any lateral connectivity or sigmoidal modulation, the activation is feed-forward / input-driven.

Figure 1F shows the same input sequence, but now with lateral connectivity and sigmoidal modulation switched on (akin to the simulation in Figure 1C-D). Initially, the cortical field is stably at rest, that is, at the value defined by *h*. At time 250, the color is presented and sites that are 'tuned' to red are activated. Around time step 500, noise fluctuations boost several sites around color value 25 into the 'on' state – they go above-threshold as defined by the sigmoidal function. Consequently, these neural sites start passing activation to their 'neighbors'. The result is the large 'peak' of activation centered over color value 25. The shadow along the feature axis shows the structure of this peak – one can see strong local excitation with inhibitory 'troughs' on either side of the peak.

Peaks in dynamic fields are the basic unit of representation accounting for detection, selection, and working memory cognitive states. Peaks are a stable attractor state of the neural population. Note how the peak in Figure 1F retains it shape through time, even amidst the neural noise evident in this simulation. This attractor state is not permanent, however; once the strength

of input is reduced, the peak reduces in strength, eventually relaxing back to the original resting level. More interestingly – as we show below – we can increase the strength of neural interactions in the field by increasing the strength of local excitation and surround inhibition and activation peaks show a form of working memory: peaks of activation can be stably maintained through time even when the input is removed (Fuster & Alexander, 1971).

Recent work has offered more biophysically detailed models of these base functions (Deco et al., 2004; Durstewitz, Seamans, & Sejnowski, 2000; Wei, Wang, & Wang, 2012), showing how spiking networks together with synaptic dynamics can reproduce, for instance, a sustained activation 'peak' (often called a 'bump' attractor). Although these newer models are computationally more detailed, we can ask: is all of this detail necessary for linking brain and behavior? Critically, there are drawbacks to this level of detail: the link of biophysical models to behavioral data is much weaker than for DFT, and the number of parameters and range of dynamical states are much larger. Thus, we do not anchor our account at this level. Nevertheless, there are links between DFT and biophysical models: under simplified assumptions, the population-level neural dynamics of DFT may be obtained from the Mean Field approximation (Faugeras, Touboul, & Cessac, 2009). We leverage this understanding here to derive a relationship between DFT and fMRI, adapting biophysical accounts for how neural activity gives rise to the BOLD signal (Deco et al., 2004; Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001).

The link between DFT and the Mean Field approximation establishes that there is a theoretical connection between neural population dynamics in DFT and theories of spiking network activity. We can also ask if this connection extends beyond theory to practice—can we directly measure properties of neural population dynamics captured by DFT in real brains? This

issue was initially explored using multi-unit neurophysiology in the 1990s. In several studies of neural activity in premotor cortex, results showed that predictions of DF models of motor planning were evident in multi-unit recordings from premotor cortical neurons (Bastian et al., 1998; Bastian, Schoner, & Riehle, 2003; Erlhagen et al., 1999; Jancke et al., 1999). More recently, this connection has been explored using voltage-sensitive dye imaging in visual cortex (Markounikau, Igel, Grinvald, & Jancke, 2010). Again, properties of neural population dynamics in DF models such as slowing of neural responses due to laterally-inhibitory interactions were evident in cortical recordings. From these examples, we conclude that DFT offers a good approximation of the dynamics of populations of neurons in cortex. This sets the stage to expand this line of work to human cognitive neuroscience techniques such as fMRI.

We have now reviewed the basic concepts of neural population dynamics in cortical fields that underlie DFT. The next step is to couple multiple DFs together to create a neural architecture that implements specific cognitive processes in a neural way. In the next section, we describe a neural architecture designed to capture how people encode and consolidate features in VWM, how they remember these features during a delay, and how they compare these remembered features with the features in a test array to generate ‘same’ and ‘different’ decisions.

A Dynamic Field model of VWM

We situate the dynamic field (DF) model within the canonical task used to study VWM—the change detection task (Luck & Vogel, 1997). Participants are shown a sample array with multiple objects. After a delay, a test array is displayed and participants decide whether the sample and test arrays are the ‘same’ or ‘different’. Previous work has focused on encoding and maintenance in this task, resulting in debates about whether VWM consists of fixed-resolution “slots” (Luck & Vogel, 1997) or a distributed resource (Bays & Husain, 2008). Other work has

investigated the biophysical properties of neural networks that give rise to sustained activation in VWM (Wei et al., 2012). Critically, detecting change requires that encoding and maintenance be integrated with comparison. *The DF model provides the only formal account that specifies how this integration occurs in a neural system to generate 'same' and 'different' responses* (Johnson et al., 2014; Johnson, Spencer, Luck, et al., 2009).

Figure 2 shows the architecture of the DF model (see Supplemental Information for model equations and parameters). The model consists of four components that are interconnected yet serve particular functional roles (see Tables S1-S2 and Supplemental Information). The contrast field (CF) and WM layers have populations of color-sensitive neurons that build 'peaks' of activation through local-excitatory connections reflecting the presented colors (see also Engel & Wang, 2011). Inputs are presented strongly to the CF layer which leads to the formation of peaks of activation within this field during stimulus presentation. These peaks then send activation to the WM field which also builds peaks of activation at the location of the inputs (see peaks in WM layer in Figure 2). Both fields pass inhibition to one another through a shared inhibitory layer (not visualized in Figure 2 for simplicity). Through this pattern of coupling, the model dynamics operate such that CF becomes suppressed (see inhibitory profile in CF layer in Figure 2) once items are consolidated within the WM field and the inputs are removed. When items are re-presented at test, inputs that match peaks in WM will be suppressed in CF, while non-matching inputs will build peaks in CF. During this phase of the trial, the model engages in a winner-take-all comparison process by boosting the 'same' and 'different' nodes close to threshold (via activation of a 'gate' node; see Figure 2). The 'different' node receives input from CF; the 'same' node receives input from WM. Consequently, if the model detects non-matching inputs at test, 'different' will win the competition; if, however, no or few non-matching inputs

are detected, 'same' will win the competition due to strong input from WM. It is important to point out that the input to the 'same' node is effectively normalized by input from the inhibitory layer to enable equitable comparisons with the 'different' node as the set-size (SS) increases (see equation 7 in the supplemental information). That is, as the SS increases, more items will be activated in WM, generating more input to the 'same' node. This would create a large asymmetry between activation in the 'same' and 'different' systems, making it hard to detect differences at high SS. To help compensate for this asymmetry, the Inhib layer also sends inhibitory output to the 'same' node, effectively balancing the increase in excitation from WM at high SS with an increase in inhibition from Inhib (which also increases at high SS).

Before describing the dynamics of the model in detail, it is useful to first consider the following dynamic field equation that defines the neural population dynamics of the CF layer to connect to the concepts introduced in the previous section:

$$\begin{aligned}\tau_e \dot{u}(x,t) = & -u(x,t) + h + s(x,t) + \int c_{uu}(x-x')\, g\left(u(x',t)\right) dx' \\ & - \int c_{uv}(x-x')\, g\left(v(x',t)\right) dx' - a_{uv_{global}} \int g\left(v(x',t)\right) dx' \\ & + \int c_r(x-x')\, \xi(x',t)\, dx' + a_{ud}\, g(d(t)) - a_{um}\, g(m(t))\end{aligned} \tag{5}$$

Activation, *u*, in CF evolves over the timescale determined by the *τ* parameter (see Supplemental Information). The first three terms term in Equation 5 are the same as in Equation 4. Next is local excitation, $\int c_{uu}(x-x')\, g\left(u(x',t)\right) dx'$, which is defined as the convolution of a Gaussian local excitation function, $c_{uu}(x-x')$, with the sigmoided output, $g(u(x',t))$, from the CF layer. CF receives inhibition from an inhibitory layer, *v*. Lateral inhibitory contributions are specified by, $-\int c_{uv}(x-x')\, g\left(v(x',t)\right) dx'$, which is defined as the convolution of a Gaussian

surround inhibition function and the sigmoided output from an inhibitory layer (*v*). There is also a global inhibitory contribution specified by, $-a_{uv\_global} \int g\left(v(x',t)\right)dx'$, which is applied homogenously across the field. These two inhibitory terms give rise to inhibitory troughs that surround local excitatory peaks in the contrast layer. The next term specifies spatially correlated noise, $\int c_r(x-x')\xi(x',t)dx'$, which is defined as the convolution of a Gaussian kernel and a vector of white noise. This simulates a set of noisy inputs to CF reflecting neural noise impinging upon this local neural population. The last two terms specify inputs from the decision nodes (see Figure 2). Both of these inputs are modulated by the sigmoidal function (*g*). The 'different' node (*d*) globally excites CF, $a_{ud}g\big(d(t)\big)$, while the same or "match" node (*m*) globally inhibits CF, $-a_{um}g(m(t))$. These excitatory and inhibitory inputs help maintain peaks in CF if a difference is detected, and help suppress activation in CF if 'sameness' is detected (see 'crossing' inhibitory connections between the decision nodes and CF/WM in Figure 2). Note that there is no direct input from WM to CF.

Figure 3 shows an exemplary simulation of a single change detection trail to show how activation changes through time as the model encodes items into memory, maintains memory representations during a delay, and then detects a difference in a subsequently presented stimulus array. Figure 3A shows activation across the feature space in CF and WM through time. Figure 3B shows the node activations through time. The remaining panels show time slices through CF and WM at particular points during the simulation indicated by the boxes in Figure 3A (see also downward arrows marking the same time points in Figure 3B).

At 100 ms into the simulation, 3 colored stimuli (3 Gaussian inputs) are presented to the model. Initially, this is associated with large increases in activation in CF; a bit later, peaks build in the WM layer (see Figure 3A). As activation builds in WM, activation in CF becomes

suppressed. After 600 ms into the simulation, the stimulus array is turned off. Now, activation within CF is strongly suppressed (see troughs in Figure 3C). However, activation in WM is sustained in the absence of the input throughout the delay period (see Figure 3D) due to strong recurrent interactions within this layer. At 1800 ms into the simulation, a second array of stimuli is presented to the model. At presentation of the test array, the gate node is activated (Figure 3B); this boosts the activation of the 'same' and 'different' nodes. At the same time, the presentation of the novel color (C4) leads to the formation of a new peak in CF (Figure 3E). This peak increases the activation of the 'different' node and this node goes above threshold (Figure 3B) leading to a 'different' decision on this trial.

A key innovation of the DF model is that the model captures what happens on both correct *and* incorrect trials. Figure 4 shows exemplary simulations of instances in which the model performs correctly or incorrectly on each trial type in the change detection task. Figure 4A shows a correct rejection trial – correctly responding 'same' on a 'same' trial. Note that we are using terminology from the literature on visual change detection here (Cowan, 2001; Pashler, 1988). A sample array of four colors is presented at the start of the simulation, generating peaks in CF. Peaks in CF drive the consolidation of the peaks in the WM field, after which activation within CF becomes suppressed. This is shown in the lower left panels of Figure 4A: at the offset of the memory array, 4 peaks are being actively maintained in WM while there is a profile of inhibitory troughs in CF. During the memory delay, activation is maintained within WM via recurrent interactions. When the same four colors are presented at test, no peaks are built in CF (see asterisks above CF input locations in Figure 4A). The decision nodes are plotted at the top. At the end of the trial, the 'same' decision is above threshold indicating the that the model has correctly generated a 'same' response.

Figure 4B shows a simulation of a hit trial – correctly detecting a change on a 'different' trial. The dynamics during the presentation of the memory array are comparable. In particular, at the offset of the memory array, four peaks are being actively maintained in WM, with a profile of inhibitory troughs in CF. During the test array, a new item is presented (C5) along with 3 of the original inputs (C2-C4); the new input generates a peak in CF at this color value because there is not enough inhibition at this site to prevent the peak from emerging (see asterisk above CF in Figure 4B). The peak in CF passes strong input to the 'different' node such that by the end of the trial, the 'different' node is above threshold indicating that the model has correctly generated a 'different' response.

The bottom two panels in Figure 4 show the model's performance on error trials. Figure 4C shows a false alarm trial – incorrectly generating a 'different' response on a no change trial. False alarms are likely to arise in the model when a peak fails to consolidate in WM. This is shown in the lower left panels of Figure 4C: after presentation of the memory array, one peak fails to consolidate (fails to go above threshold; see asterisk) and activation at this site returns to baseline levels during the delay. Consequently, when the same colors are presented at test, the model falsely detects a change (see asterisk above CF in the right column of Figure 4C). In contrast to other models (Cowan, 2001; Pashler, 1988), therefore, false alarms reflect a failure of consolidation / maintenance rather than a guess.

A 'miss' trial is shown in Figure 4D – incorrectly generating a 'same' response on a change trial. This simulation shows a typical state of the neural dynamics after presentation of the memory array, with four peaks being maintained in WM and an inhibitory profile in CF. Note, however, the strong inhibitory suppression on the left side of the feature space as there are three WM peaks relatively close together. Consequently, when a different color is presented in

that region of feature space, a weak activation bump is generated in CF (see asterisk above CF in Figure 4D). This bump is too weak to drive a ‘different’ response and the ‘same’ node wins the decision-making competition (see top panel in Figure 4D). Thus, in contrast to assumptions of other models (Cowan, 2001; Pashler, 1988), comparison is not a perfect process in the DF model; misses occur even when all items are remembered. This aspect of the DF model is consistent with more recent work illustrating how comparison errors can impact performance on WM tasks (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007).

Note that errors in the DF model are impacted by stochastic noise in the equations--a realistic source of neural noise that is evident in actual neural systems. These fluctuations are amplified by local excitatory / inhibitory neural interactions and can influence the macroscopic patterns -- peaks in the model -- that impact different behavioral outcomes such as 'same' and 'different' decisions. Notice, for instance, that the inputs across all four panels in Figure 4 are identical; the parameters of the model are identical as well. Thus, the only thing that differs is how the activation dyanmics unfold through time in the context of neural noise. Of course, noise is not the only factor that influences whether the model makes an error. The number of inputs plays a large role as does the metric similarity of the items. With more peaks to maintain, there is more competition among peaks as well as more global inhibition. Consequently, the likelihood of a false alarm increases because neighboring peaks might fail to consolidate in WM. At the same time, with more peaks in WM, there is also a greater overall suppression of CF and stronger input to the ‘same’ node. Consequently, the likelihood of a miss increases as well.

Are there unique neural signatures of the processes illustrated in Figure 4? If so, that would provide a way to test our account of the origin of errors in change detection. To examine this question here, we used an integrative cognitive neuroscience approach initially developed in

Buss et al. (2014) and Wijeakumar et al. (2016). We describe this approach next.

Turning neural population activation in DFT into hemodynamic predictions

In this section, we describe a linking hypothesis derived from the model-based fMRI literature that directly links neural dynamics in DFT to hemodynamics that can be measured with fMRI. This requires consideration of multiple factors, including what is measured by fMRI both in terms of hemodynamics and spatially in patterns of BOLD within voxels through time. Here, we make several simplifying assumptions which we discuss. The end product is a direct link—millisecond by millisecond—between neural activation in the DF model and fMRI measures through time as well as to behavioral decisions on each trial. Although the timescale of fMRI does not allow for millisecond precision, the model is specified at that fine-grained timescale and, therefore, could be mapped to other technologies such as ERP in future work (we return to this issue in the General Discussion). Critically, this approach extends beyond previous model-based approaches (Ashby & Waldschmidt, 2008; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). Specifically, this approach specifies mechanisms that directly give rise to behavioral and neural responses; consequently, any modifications to these mechanisms directly impact the resultant behavioral and neural resposnes predicted by the model. To illustrate, we contrast our approach with model-based fMRI examples using the adaptive control of thought - rational (ACT-R) framework. We conclude that the DF-based approach is an example of an integrative cognitive neuroscience approach to fMRI (Turner et al., 2016).

Our approach builds from the biophysiological literature examining the basis of the neural blood flow response. Logothetis and colleagues (2001) demonstrated that the local-field potential (LFP), a measure of dendritic activity within a population of neurons, is temporally correlated with the blood oxygen level dependent (BOLD) signal. Furthermore, the BOLD

response can be reconstructed by convolving the LFP with an impulse response function which specifies the time course of the blood flow response to the underlying neural activity. Deco and colleagues followed up on this work using an integrate-and-fire neural network to demonstrated that an LFP can be simulated by summing the absolute value of all of the forces that contribute to the rate of change in activation of the neural units (Deco et al., 2004). Attempts to simulate fMRI data using this approach were equivocal—some hemodynamic patterns produced by the network did qualitatively mimic fMRI data measured in experiment; however, no efforts were made to quantitatively evaluate the fit of the spiking network model to either the behavioral or fMRI data.

Here, we adapt this approach to construct an LFP signal for each component of the DF model. To describe how we transform the real-time neural activation in the model into a neural prediction that can be measured with fMRI, re-consider the equation that defines the neural population dynamics of the CF layer (reproduced here for convenience):

$$\begin{aligned}\tau_e \dot{u}(x,t) = \; & -u(x,t) + h + s(x,t) + \int c_{uu}(x-x')\, g\left(u(x',t)\right) dx' \\ & - \int c_{uv}(x-x')\, g\left(v(x',t)\right) dx' - a_{uv_{global}} \int g\left(v(x',t)\right) dx' \\ & + \int c_r(x-x')\, \xi(x',t)\, dx' + a_{ud}\, g\big(d(t)\big) - a_{um}\, g\big(m(t)\big)\end{aligned} \tag{6}$$

To simulate hemodynamics, we transformed this equation into an LFP equation that we could track in real time (millisecond by millisecond) for each component of the model (see Equations 9-14 in Supplemental Information). This time-course was then convolved with an impulse response function to give rise to hemodynamic predictions that could be compared to BOLD data. To illustrate, equation 7 specifies the LFP for the contrast field: we summed the absolute value of all terms contributing to the rate of change in activation within the field, excluding the stability term, *-u(x,t)*, and the neuronal resting level, *h*. We also excluded the

stimulus input, *s(x,t)*, because we applied inputs directly to the model rather than implementing these in a more neurally realistic manner (e.g., by using simulated input fields as in Lipinski et al., 2012). The resulting LFP equation was as follows:

$$u_{LFP}(t) = \frac{\left| \iint c_{uu}(x - x')\, g\left(u(x', t)\right) dx\, dx' \right|}{\eta} + \frac{\left| \iint c_{uv}(x - x')\, g\left(v(x', t)\right) dx\, dx' \right|}{\eta} + \left| a_{uv_{global}} \int g\left(v(x', t)\right) dx' \right| + \frac{\left| \iint c_r(x - x')\, \xi(x', t) dx\, dx' \right|}{\eta} + \left| a_{ud}\, g(d(t)) \right| + \left| a_{um}\, g(m(t)) \right| \quad (7)$$

It is important to note several simplifying assumptions here. First, neural activity in the CF field was aggregated into a single LFP (representing a single neural region). We consider this a starting point for explorations of this model-based fMRI approach. An alternative would be to use several basis functions to sample different parts of the field and then explore the mapping of these localized LFPs to voxel-based patterns in the brain. Later in the paper, we quantitatively map hemodynamic predictions from the DF model to BOLD signals measured from 1cm^3 spheres centered at regions of interest from a meta-analysis of the fMRI VWM literature (Wijeakumar, Spencer, Bohache, Boas, & Magnotta, 2015). At this resolution (1cm^3), slight variations in hemodynamics due to which part of the field we are sampling from probably make little difference. By contrast, if we were studying population dynamics in visual cortex with a 7T scanner in different laminar layers, the use of basis functions to sample the field would be an interesting alternative to explore.

Similarly, in equation 7 we normalized each contribution to the LFP by dividing by the number of units in that contribution, either by 1 (e.g., for the 'same' node) or by η, the field size.

This way, contributions to the CF LFP from, say, the different node were of comparable magnitude to contributions from local excitatory interactions. Again, this is a simplifying assumption that can be explored in future work. For instance, there is an emerging literature examining how excitatory versus inhibitory neural interactions differentially contribute to the BOLD signal (Lee et al., 2010). It would be possible to differentially weight these types of contributions to the LFP in future work as clarity emerges on this front. In the simulations reported below, we down-weighted all inhibitory LFP components by a factor of 0.2.

Once an LFP has been calculated from each component of the DF model – one LFP for CF, one for WM, one for 'different', and one for 'same' – a hemodynamic response can then be calculated by convolving u_{LFP} with an impulse response function that specifies the time-course of the slow blood-flow response to neural activation (see Equation 15 in the Supplemental Information). The simulated hemodynamic time course for each component was computed as a percent signal change relative to the maximum intensity across the run. Average responses for each trial-type within each component were then computed within the relevant time window (14s for the simulations of the Todd & Marois data and 20s for the Magen et al. data) as the amount of change relative to the onset of the trial (see Supplemental Information for full details). A group average for each trial type was then computed across the group of runs.

Figure 5 shows an exemplary simulation of the model for a series of 8 trials with a memory load -- or set size (SS) -- of 2 items for the first two trials and 4 items for the subsequent six trials. Panels A-C show neural activation of the decision nodes and associated LFPs / hemodynamic predictions through time. In particular, panel C shows the activation of the decision and gate nodes, highlighting the evolution of decisions that reflect the overt *behavior* of the model. Going from left to right, the model makes 8 decisions in sequence (see labels at the

bottom of the figure): (1) “different” (correct), (2) “same” (correct), (3) “different” (correct), (4) “different” (incorrect), (5) “same” (incorrect), (6) “same” (correct), (7) “different” (correct), and (8) “same” (correct). Note that the long delays in-between trials accurately reflects the typical delays between trials in a neuroimaging experiment. We have fixed this time interval here to make it easier to see the hemodynamic response associated with each trial (which is delayed by several seconds reflecting the slow hemodynamic response); critically, however, we can match these inter-trial intervals precisely to reflect the actual timings used in experiment.

Panels A and B in Figure 5 show the LFP and hemodynamic responses for the ‘same’ and ‘different’ nodes, respecitvely. In general, the decision node hemodynamics are strongly influenced by the inhibition at test evident in the winner-take-all competition. For instance, the first trial is a ‘different’ (correct) trial. Here, the ‘different’ node wins the competition, but notice that the ‘same’ (Figure 5A) hemodynamic response is stronger than the ‘different’ hemodynamic response (Figure 5B); even though ‘different’ wins the competition with strong excitatory activation, the ‘same’ hemodynamic response is stronger due to the inhibitory input to this node. This is counterintuitive – the node with the stronger hemodynamic response is actually the one that *loses* the competition. We test this prediction using fMRI later in the paper.

Note that it is possible we could reverse the counterintuitive decision-node prediction in the model in two ways. First, the magnitude of the inhibitory contribution to the decision node dynamics could be reduced via parameter tuning. This would be tricky to achieve, however, because the decision system dynamics have to balance ‘just right’ such that the full pattern of behavioral data are correctly modeled. If, for instance, inhibition is too weak, the model might respond ‘same’ at high memory loads simply because there are so many peaks in WM and, therefore, strong input to the ‘same’ node at test. Thus, there are strong constraints in model

parameters – if we try to 'tune' the neural / hemodynamic predictions so they make more intuitive sense, the model might no longer accurately fit the behavioral data.

That said, there is a second way we could modify the hemodynamic predictions of the decision nodes more directly, making them less dominated by inhibition: we could down-weight the inhibitory contributions within the LFP equation itself. Doing so would be more akin to a 'two-stage' approach as outlined by Turner et al. (2016) in which separate parameters are used to generate behavioral responses and neural responses. However, by doing so we could implement the hypothesis that inhibitory contributions to LFPs are weaker than excitatory contributions, a hypothesis that could be explored using optogenetics (e.g., Lee et al., 2010). To do this, we could add a new inhibitory weighting parameter to equation 7 to reduce the strength of the inhibitory contributions (i.e., the second, third, and sixth terms in the equation). Note that this would have to be applied to all inhibitiory terms in the full model; consequently, inhibition would have less of an effect on the decision-node hemodynamics, but it would also have less of an effect on the CF and WM hemodynamics as well. We explore this sense of parameter tuning in the first simulation experiment.

Panels E and G in Figure 5 show the activation of CF and WM, respectively. Note that all of the activation dynamics highlighted in the field activities in Figure 3A still occur here; however, these dynamics are compressed in time as we are showing a sequence of 8 trials with relatively long inter-trial intervals. That said, on each trial, the sequence of stimulus presentations is evident in CF at the start and end of each trial (see peaks at the onset and offset of each inhibitory period in Figure 5E), while the active maintenance of peaks in WM is also readily apparent (Figure 5G).

Panels D and F in Figure 5 show the LFP and hemodynamic predictions for CF and WM.

CF is influenced by whether the trial is ‘same’ or ‘different’, with a slightly stronger response in CF on ‘different’ trials (see, for instance, the large first and third hemodynamic peaks; we show this more clearly later in the paper when we aggregate LFPs across many simulation trials versus the individual simulations as shown here). WM is most strongly influenced by how many items are maintained during the delay; thus, this layer shows relatively weaker responses on the first two trials when the memory load is 2 items compared to the subsequent trials when the memory load is 4 items.

In summary, Figure 5 illustrates over a series of trials how the model generates a complex pattern of predictions associated with the neural processes that undelie encoding and consolidation of items in WM, the maintenanace of those items during the memory delay, and decision-making and comparison processes at test. Importantly, LFPs and hemodyanmic responses are extracted from the same patterns of neural activation that drive neural function and behavioral responses on each trial. In this way, distinct neural dynamics are engaged across components of the model as different types of decisions unfold in the context of the change detection task and these directly lead to hemodynamic predictions. The distinctive nature of these simulated neural responses is important for being able to use the model to shed light on the functional role of different brain regions in VWM. For instance, if we find a good correspondence between model hemodynamics and hemodynamics measured with fMRI, this uniqueness gives us confidence that we can infer different functions are being carried out by those brain regions.

Comparisons with other model-based fMRI approaches

Beyond the literature on VWM, other model-based approaches to fMRI analysis have been implemented that bridge the gap between brain and behavior (see Turner et al. 2017 for an

excellent summary and classification of different approaches). In our previous paper exploring a model-based fMRI approach using DFT (Buss et al., 2014), we compared the DFT approach to the model-based fMRI approach using ACT-R. Comparing these approaches is a useful starting point as there are similarities in the broader goals of DFT and ACT-R.

Anderson and colleagues have developed a technique for simulating fMRI data with the ACT-R framework (Anderson, Albert, & Fincham, 2005; Anderson et al., 2008; Anderson, Qin, Sohn, Stenger, & Carter, 2003; Borst & Anderson, 2013; Borst, Nijboer, Taatgen, van Rijn, & Anderson, 2015; Qin et al., 2003). ACT-R is a production system model that explains behavioral data based on the duration of engagement of processing modules and differential engagement of these modules across conditions. Specifically, ACT-R models posit a cognitive architecture consisting of separate modules that are recruited sequentially in a task. This generates a 'demand' function for each module through time – a time course of 0s and 1s with 1s being generated when a module is active. The 'demand' function can then be convolved with an HRF for each module to generate a predicted BOLD signal for each component of the architecture. The predicted hemodynamic pattern can then be compared against brain activity measured with fMRI in specific brain regions to determine the correspondence between modules in the model and brain regions.

This approach is similar to the DFT-based approach used here. Both ACT-R and DFT build architectures to realize particular cognitive functions. Both measure activation through time for each part of the larger architecture. These activation signals are then convolved with an impulse response function to generate predicted BOLD signals for each component. By comparing these predicted signals to fMRI data, the components can be mapped to brain regions and function can be inferred from this mapping. This can be done by qualitatively comparing

properties of the predicted brain response through time to measured HRFs (e.g., Buss et al., 2014; Fincham, Carter, van Veen, Stenger, & Anderson, 2002). We adopt this approach in the first simulation experiment here. Model-predicted data can also be quantitatively compared to measured fMRI data using a general linear modeling approach (e.g., Anderson et al., 2007). We adopt this approach in the subsequent simulation experiment.

In the review of model-based fMRI approaches by Turner and colleageus (Turner et al., 2016), they used the ACT-R approach as an example of integrative cognitive neuroscience (ICN). Recall that the goal of ICN is to develop a single model capable of predicting both neural and behavioral measures. Formally, ICN approaches use a single model with a single set of parameters, θ, that jointly explain both neural and behavioral data. Consequently, such models must make a moment-by-moment prediction of neural data, and a trial-by-trial prediction of the behavioral data. One can see why ACT-R might be a good example of ICN: the model specifies the activation of each module in real time, and this activation affects the model's neural predictions because it changes the 'demand' function (the vector of 0s and 1s through time). Differences in activation also affect behavior, for instance, modulating reaction times.

Given the similarities between ACT-R and DFT, we can ask if DFT rises to the level of ICN as well. Like with ACT-R, DFT proposes a specific integration of brain and behavior. In particular, there are not separate neural vs. behavioral parameters; rather, there is one set of parameters in the neural model and changes in these parameters have direct consequences for both neural activity – the LFPs generated for each component – and for the behavioral decisions of the model – whether the 'same' or 'different' node enter the 'on' state and when in time this decision is made (yielding a reaction time for the model).

These examples highlight that in DFT, brain and behavior do not live at different levels.

Instead, there is one level – the level of neural population dynamics. This level generates neural patterns through time on a millisecond timescale. This level also generates macroscopic decisions on every trial via the neural population activity of the 'same' and 'different' nodes. When one of these nodes enters the 'on' attractor state at the end of each trial, a behavioral decision is made. In this sense, we contend that DFT – like ACT-R – is an example of an ICN approach.

Given the many similarities in these two approaches to model-based fMRI, we can ask the next question: are there key differences? The most substantive difference is in how the two frameworks conceptualize 'activation' and, relatedly, how they implement processes through time. As demonstrated in Figures 3-5, the activation patterns measured in each neural population in the DF model are more than just an index of the engagement of the population; rather, activation has meaning—it represents the colors presented in the task. This was emphasized in our introduction to DFT. Although 'activation' and, in particular, the neural dynamics that govern activation, are key concepts in DFT, we moved beyond the *level* of activation to think about what activation represents by modelling *activation in a neural field distributed over a feature dimension*.

Critically, by grounding activation in a specific feature space we also had to specify the neural processes through time that do the job of consolidating features in WM, maintaining those features through time, and then comparing the features in WM with the features in the test array. Thus, our model not only specifies what activation means; *it also specifies the neural processes that underlie behavior*, that is, the neural processes that give rise to the macroscopic neural patterns that underlie same/different decisions on each trial. Importantly, the details of this neural implementation have consequences for the activation patterns produced by the model. If we, for

instance, changed how encoding and consolidation were done by adding new layers to the model to separate visual encoding from shifts of attention to each item (Schneegans, Spencer, & Schöner, 2016), the model would generate different activation patterns through time and, consequently, different hemodynamic predictions.

By contrast, activation in ACT-R is abstract. Each module takes a specific amount of time which creates differences in the 'demand' or 'activation' function, but the modules in ACT-R typically do not actually implement anything; rather, they instantiate how long the process would take if it were to implement a particular function. Sometimes modules are actually implemented (Jilk, Lebiere, O'Reilly, & Anderson, 2008), but this has not been done with any fMRI examples.

Is this difference in how 'activation' is conceptualized important? To evaluate this question, consider a recent model of VWM using ACT-R (Veksler et al., 2017). At face value, this model sets up an ideal contrast—in theory, we could contrast the model-based fMRI prediction of our DF model with model-based fMRI predictions derived from the Veksler et al. ACT-R model. To explain why we cannot do this, it is useful to first describe the Veksler et al. model.

The Veksler et al. model uses the ACT-R memory equation to implement a variant of VWM. Each item in the display is associated with an activation level in the memory module that is a function of whether it was fixated/encoded, how recently it was fixated/encoded, a decay rate, a base-level offset for activation, and logistically distributed noise with a mean of 0 and a specific *SD*. To place this model in the context of change detection, we must first make some decisions about how encoding works. For instance, in many change detection experiments, fixation is held constant, so we could assume that a specific number of items start off at a

baseline activation level. We could also hypothesize that each item takes a certain amount of time to encode and then let the model encode as many items as it can in the time allowed.

After encoding, the next key issue is which items are still remembered after the delay when memory is tested. Concretely, the memory module specifies an activation value of each item through time. If that activation value is above a threshold when memory is tested, that item is remembered. If the activation is below threshold at test, that item is forgotten.

The challenging question is what to do in this model at test. Each item is only represented by an activation level—there is no content. Consequently, *it's not clear how to do comparison*. One idea is to assume that comparison is a perfect process. This is similar to assumptions in the original models of VWM by Pashler (1988) and Cowan (2001). Thus, if an item is remembered, we always get a correct response. If an item is forgotten, then we could just have the model randomly guess. Sometimes the model will generate a lucky guess. Other times the model will guess incorrectly, generating a false alarm or a miss.

Although this approach sounds reasonable, it does not actually do a good job modelling behavior because performance varies as a function of whether the test array is the 'same' or 'different'. In particular, adults are typically more accurate on 'same' trials than 'different' trials (Luck & Vogel, 1997); interestingly, children and aging adults show this effect more dramatically (Costello & Buss, 2018; Simmering, 2016; Wijeakumar, Magnotta, & Spencer, 2017). If the model has a perfect comparison process, it's not clear how to account for such differences unless one simply builds in a bias in the guessing rate with more 'same' guesses than 'different' guesses. More importantly, this approach to comparison does not generate any predictions about the activation level on 'guess' trials when an item is forgotten because the underlying demand function would be the same on all guess trials. This doesn't match empirical

data because we know that fMRI data vary on 'correct' vs. 'incorrect' trials, as well as on false alarms vs. misses (Pessoa & Ungerleider, 2004).

In sum, when we try to implement change detection in the ACT-R VWM model, we run into a host of questions with no clear solutions. Critically, many of these questions are centered on the main contrast with DFT that, in ACT-R, there is 'activation' but no details about what activation represents. This example also highlights how important the comparison process is to predicting neural activation. On this front, we re-emphasize that to our knowledge, DFT is the only model of VWM that specifies a mechanism for how comparison is done. This observation will have consequences below—although there are many models of VWM, because none of them specify how comparison is done this means that no other models make hemodynamic predictions that we can contrast with DFT *where comparison is part of the unfolding hemodynamic response*. Instead, we opt for a different model-testing strategy by contrasting DFT with a standard statistical model.

Simulations of Todd & Marois (2004) and Magen et al. (2009)

The goal of this paper is to examine whether DFT is a useful bridge theory, *simultaneously capturing both neural and behavioral data* to directly address the neural mechanisms that underlie cognitive processes (Buss & Spencer, 2018; Buss et al., 2014; Wijeakumar et al., 2016). Here we ask whether the model can simulate two findings from the fMRI literature that describe different relationships between intraparietal sulcus (IPS) and VWM performance. One set of data show that neural activation as measured by BOLD asymptotes as people reach the putative limit of working memory capacity. In particular, Todd and Marois (2004) reported that the BOLD signal in the intraparietal sulcus (IPS) increases as more items must be remembered with an asymptote near the capacity of VWM. This suggests that the IPS

plays a direct role in VWM. This basic effect has been reported in multiple other studies as well (Todd & Marois, 2004; Xu & Chun, 2006; for related ideas using EEG, see Vogel & Machizawa, 2004). In contrast, a second set of results shows that the BOLD response in the IPS does *not* asymptote when the memory delay is increased in duration (Magen, Emmanouil, McMains, Kastner, & Treisman, 2009). From this observation, Magen and colleagues proposed that the posterior parietal cortex is more involved with the rehearsal or attentional processes that mediate VWM, rather than being the site of VWM directly. Here we ask if the DF model can shed light on these differing brain-behavior relationships, explaining the seemingly contradictory set of results.

These initial simulations serve two functions. First, they provide an initial exploration of whether the LFP-based linking hypothesis generates hemodynamics from the DF model that are qualitatively similar to measured BOLD responses. This is a non-trivial step because simulating both brain and behavior requires integrating the neural processes that underlie encoding, consolidation, maintenance, and comparison. The present experiment explores whether we get this integration approximately right. Second, this experiment serves to fix parameters of the DF model. Specifically, we allowed for some parameter modification here as we attempted to fit behavioral data from Todd and Marois (2004). We then fixed the model parameters when simulating data from Magen et al (2009) as well as in a subsequent experiment where we generated novel, *a priori* neural predictions that could be tested with fMRI.

Methods

Simulations were conducted in Matlab 7.5.0 (Mathworks, Inc.) on a PC with an Intel® i7 3.33 GHz quad-core processor (the Matlab code is available at www.dynamicfieldtheory.org). For the purposes of mapping model dynamics to real-time, 1 time-step in the model was equal to

2 ms. For instance, to mimic the experimental paradigm of Todd and Marois (2004), the model was given a set of Gaussian inputs (e.g., 3 colors = 3 Gaussian inputs centered over different hue values) corresponding to the sample array for a duration of 75 time-steps (150 ms). This was followed by a delay of 600 time-steps (1200 ms) during which no inputs were presented. Finally, the test item was presented for 900 time-steps (1800 ms). For the simulation of the Magen et al. (2009) task (Experiment 3), the sample array was presented for 250 time-steps (500 ms), followed by a delay of 3000 time-steps (6000 ms) and a test array that was presented for 1200 time-steps (2400 ms). For both simulations, the response of the model was determined based on which decision node became stably activated during the test array (see Figures 3-5). Recall that the local-excitation/lateral-inhibition operating on the decision nodes gives rise to a winner-take-all dynamics that generates a single active (i.e., above 0) decision node at the end of every trial.

The central question here was whether the neural patterns generated by the model mimic the differing BOLD signatures reported by Todd and Marois (2004) and Magen et al. (2009). To examine this question, we first used the model to simulate the behavioral data from Todd and Marois (2004). We initialized the model using the parameters from Johnson et al. (2009a), then modified parameters iteratively until the model provided a good quantitative fit to the behavioral patterns from Todd and Marois (2004). For example, the resting level of the CF component had to be increased to accommodate for the shorter duration of the memory array in the Todd and Marois study. To compensate for the increased excitability of this component, we also had to reduce the strength of its self-excitation (see Appendix for full set of parameters and differences from the Johnson et al. 2009a model). We implemented the model to match the number of participants from the target studies to facilitate statistical comparison of the datasets. Specifically, we simulated the model 17 times in the Todd and Marois (2004) task to match the

17 participants in this study, and 12 times in the Magen et al. (2009) task to match the 12 participants in their study. We adminstered 60 same and 60 different trials at each set size for each simulation run. Group data were then computed to compare with group data from these studies.Once the model provided a good fit to the Todd and Marois (2004) behavioral data, we then assessed whether components of the model produced the asymptote in the IPS hemodynamic response observed in the original report. This was indeed the case. These model parameters were then used to simulate data from Magen at al. (2009) as well as in the subsequent fMRI experiment to test novel predictions of the model.

Results

As shown in Figure 6A, the model captured the behavioral data from Todd and Marois (2004) well overall with RMSE = 0.063. It is important to note that the model was able to reproduce these data even though there were many differences in the behavioral task between this study and the study by Johnson et al (2009a) that was used to generate the model. The duration of the memory array was shorter in the Todd and Marois task (100ms compared to 500ms in Johnson et al.) and the memory delay was longer (1,200 ms compared to 1,000 ms in Johnson et al.). To highlight these differences, Table 1 summarizes the different versions of the change detection task that have been previously modeled using DFT.

Critically, the model showed a pattern of differences between activation over SS that reproduced the asymptote effect in CF (shown in panel B of Figure 6 along with fMRI data from IPS from Todd & Marios, 2004). Thus, the CF component replicated the pattern of activation reported by Todd and Marois from IPS. Comparing SS1-4 with each other, there was a significant increase in the average time course of the hemodynamic response for the contrast layer as SS increased (all $p<.01$). As reported by Todd and Marois (2004), there was not a

significant difference in the hemodynamic time course between SS4 and SS6 ($t(16)= 0.1187$, $p=.907$) or SS4 and SS8 ($t(16)= 0.5188$, $p=.611$). These data show a good correspondence between the neural dynamics from CF and the measured hemodynamic responses of IPS.

To examine whether the asymptotic effect was unique to CF, we examined the hemodynamic patterns produced by the other model components (Figure 6C). The 'same' node also produced evidence of an asymptote in the simulated hemodynamic response (comparing SS1 through SS4: $p < .001$; SS4 v SS6: $t(16)= 0.2589$, $p= .799$). However, a decrease in activation was observed between SS4 and SS8 ($t(16)= -7.927$, $p< .001$). The WM field and the 'different' node did not produce a statistical asymptote in activation. The WM field showed a systematic increase in the HDR over set sizes (all $t(16) >16.1290$, $p< .001$). The 'different' node showed a decrease in activation from SS1 to SS4 ($t(16)>3.8783$, $p<.002$), a trending difference between SS4 and SS6 ($t(16) = 2.024$, $p= .06$), and an increase in activation between SS6 and SS8 ($t(16)= 7.3788$, $p< .001$). These results illustrate that different components of the model can yield distinct patterns of hemodynamics based on how these components are activated over the course of a task.

We next examined whether the same model with the same parameters could also simulate behavioral and IPS data from Experiment 3 in Magen et al. (2009). Simulation results this task are shown in Figure 7. As can be seen, the model approximated the behavioral data well (now presented as capacity, Cowan's *K*, instead of percent correct) with an overall RMSE = 0.477 (Figure 7A). The hemodynamic data from the model did not show a double-humped pattern; however, none of the model components showed an asymptote in this long-delay paradigm, consistent with the steady increase in activation evident in data from posterior parietal cortex from Magen et al. (2009). In particular, activation increased across set sizes for the CF, WM, and

'same' node components (all $t(11) > 3.031$, $p < .02$). The hemodynamic response produced by the 'different' node decreased in amplitude between SS1 and SS3 ($t(11) = -10.817$, $p < .001$) and from SS3 to SS5 ($t(11) = -5.6792$, $p < .001$). The amplitude of the hemodynamic response did not differ between SS5 and SS7 ($t(11) = 0.006$, $p = .995$).

Discussion

These results represent an important step in model-based approaches to fMRI. To our knowledge, this is the first demonstration of a fit to both behavioral *and* fMRI data from a neural process model in a working memory task. Simultaneously integrating behavioral and neural data within a neurocomputational model is an important achievement (Turner et al., 2016). This points to the utility of DFT as a bridge theory in psychology and neuroscience.

The DF model is also the first neural process model to quantitatively reproduce the asymptotic pattern from IPS reported by Todd and Marois (2004). Interestingly, the asymptote in the HDR was observed most robustly in the CF component. The asymptote in CF was due to the dynamics that give rise to the inhibitory filter within this field. As more items are added to the WM field, each item carries weaker activation due to the buildup of lateral inhibition. Consequently, less inhibition is passed from the Inhib layer to CF as the set size increases. An asymptote was also partially observed in the 'same' node. In this case, the asymptote was due to the effect of inhibition weakening the average synaptic output per peak within the WM field.

Interestingly, the hemodynamics within the WM field grew at each increase in set-size due to the combined influence of inhibitory and excitatory synaptic activity. Strictly speaking, the model does have a carrying capacity in terms of the number of peaks that can be simultaneously maintained (Spencer, Perone, & Johnson, 2009). The model is capacity-limited for two reasons. First, there are crowding effects: each new color peak that is added to the field

has an inhibitory surround that can suppress the activation of metrically similar color values (see, Franconeri et al., 2010). Second, each peak increases the amount of global inhibition across the field; consequently, it becomes harder to build new peaks at high set sizes (for detailed discussion, see Spencer et al., 2009). Importantly, however, there is not a direct correspondence in the model between the number of peaks that it can maintain and the capacity estimated by its performance, that is, the maximum number of peaks in WM is not the same as capacity estimated by *K* (see Johnson et al., 2014). In this sense, the continued increase in WM-related activation across set sizes evident in Figure 6 simply reflects that the model has not yet hit its *neural* capacity limit.

This set of results challenges prior interpretations of neural activation in VWM. That is, a hypothesized signature of working memory – the asymptote in the BOLD signal at high working memory loads – is not directly reflected in cortical fields that serve a working memory function; rather, this effect is reflected in cortical fields *directly coupled to working memory* (CF and the 'same' node in the case of the DF model) via the shared inhibitory layer. More concretely, the primary synaptic output impinging upon CF is the inhibitory projection from Inhib. As peaks are added to WM, activation saturates in this field as does the amount of activation within the inhibitory layer. Thus, the asymptotic effect is a signature of neural populations *coupled* to WM systems *rather than the site of WM itself*.

Multiple empirical papers have reported evidence of an asymptote in IPS in VWM tasks, some using fMRI (Ambrose, Wijeakumar, Buss, & Spencer, 2016; Magen et al., 2009; Todd & Marois, 2004, 2005; Xu, 2007; Xu & Chun, 2006) and some using EEG (Sheremata, Bettencourt, & Somers, 2010; Vogel & Machizawa, 2004). Although the asymptote effect is consistent, there is variability in the details of the asymptote effect across studies and associated neural indices.

Several papers have reported that the asymptote effect varies systematically with individual differences in behavioral estimates of capacity (see Todd et al., 2005). For instance, Vogel and Machizawa (2004) showed that increases in the contralateral delay amplitude in parietal cortex from a memory load of 2 to 4 items correlates with individual differences in capacity measured with Cowan's *K*. Other studies, however, have not replicated this link to individual differences. Xu and Chun (2006) found correlations between *K* and increases in brain activity in IPS for simple features, but no significant correlation for complex features. Magen et al. (2009) reported a divergence between behavioral estimates of capacity and brain activity in IPS. Similarly, Ambrose et al. (2016) found no robust correlations between behavioral estimates of capacity and brain activity across manipulations of colors and shapes.

Other studies have used the asymptote effect to investigate the type of information stored in IPS. Xu (2007) reported that IPS activation varies with the total amount of featural information people must remember. Xu and Chen (2006) modified this conclusion, suggesting that superior IPS activity varies with featural complexity while inferior IPS activity varies with the number of objects that must be remembered. Variation with featural complexity was also reported by Ambrose et al. (2016), but this effect extended to multiple areas including ventral occipital cortex and occipital cortex. More recently, data from Sheremata et al. (2010) suggest that left IPS remembers contralateral items, but right IPS contains two populations, one for spatial indexing of the contralateral visual field and another involved in nonspatial memory processing.

Critically, all of these studies adopt the same perspective – that the asymptote effect points towards a role for IPS in memory maintenance. We found one exception to this view: Magen et al. (2009) suggest that IPS activity may reflect the attentional demands of rehearsal

rather than capacity limitations per se as activation increases above capacity in some conditions. The perspective offered by the DF model may be most in line with Magen et al. (2009) in that our findings suggest IPS does not play a central role in maintenance but rather comparison.

Critically, we showed that the same model could reproduce the pattern of hemodynamic responses reported by both Todd and Marois (2004) and Magen et al. (2009). In particular, the model showed an asymptote in the Todd and Marois short-delay paradigm as well as the absence of a asymptote in the Magen et al. long-delay paradigm. Why are there these differences? In large part, this comes down to the relative coarseness of the hemodynamic response. In the short-delay paradigm, activation differences in CF at high set sizes are relatively short-lived and, therefore, fail to have a big impact on the slow hemodynamic response. In the long delay condition, by contrast, activation differences in CF at high set sizes extend across the entire delay; consequently, these differences are reflected even in the slow hemodynamic response.

Although the DF model did a good job capturing the magnitude of the hemodynamic response in IPS, simulations of data from Magen et al. (2009) failed to capture the shape of the hemodynamic response – the double-humped hemodynamic response that has been observed across multiple studies (Todd, Han, Harrison, & Marois, 2011; Xu & Chun, 2006). We examined this issue in a series of exploratory simulations and found that the details of the HDR played a role in the non-optimal fit. In particular, if we re-run our simulations with a narrower HDR that starts later and lasts for less time (see blue line in Supplemental Figure 1A), we still effectively simulate IPS data from both studies and see more of a double-humped hemodynamic response for simulations of data from Magen et al. (with 'humps' at the right points in time). That said, we were not able to show the dramatic dip in CF hemodynamics around 12s that is evident in the data. We suspect that this could be achieved by down-weighting the inhibitory contributions to

the LFP more strongly. This highlights a key direction for future work that adopts a two-stage approach to optimizing DF models – a first stage of getting the fits to neural data approximately right and a second stage where parameters of the HDR and the LFP→HDR mapping are iteratively optimized to fit neural data.

More generally, the present simulations show how neural process models can usefully contribute to a deeper understanding of what particular fMRI signatures like the asymptote effect actually indicate. To our knowledge, the asymptote effect has only been simulated using abstract mathematical models (see Bays, 2018 for a recent comparison of plateau vs. saturation models). While this can be useful, it can be difficult to adjudicate between competing theories at this level as the myriad papers contrasting slot and resource models can attest (e.g., Brady & Tenenbaum, 2013; Donkin et al., 2013; Kary et al., 2016; Rouder et al., 2008; Sims et al., 2012). Our results show that neural process models can shed new light on these debates, clarifying why particular neural and behavioral patterns are evident in some experiments and not others.

In the next section, we seek more direct evidence of the neural processes implemented in the DF model. Importantly, the model not only simulates the asymptote in activation observed in IPS, but makes quantitative predictions regarding neural dynamics on both correct and incorrect trials. Thus, we describe an fMRI study optimized to test hemodynamic predictions of the DF model. We then use our integrative cognitive neuroscience approach combined with general linear modeling to create a mapping from the neural dynamics in the DF model to neural dynamics in the brain.

Testing novel predictions of the DF model: An fMRI study of VWM

Having fixed the model parameters via simulations of data from Todd and Marois (2004), we examined our central question—whether the DF model predicts the localized neural

dynamics measured with fMRI as people engage in the change detection task on both correct and incorrect trials. Because the model generates specific neural patterns on every type of trial (see Figure 4), an optimal way to test the model is in a task where each trial type occurs with high frequency. Thus, we developed a change detection task that would yield many correct and incorrect trials for analysis, but above-chance responding (ensuring that participants were not guessing). Below we describe the task and details of the fMRI data collection. We then present behavioral data from a preliminary behavioral study and the fMRI study along with behavioral simulation results from the DF model. This sets the stage for a detailed examination of whether the hemodynamic patterns predicted by the model are evident in the fMRI data and whether such patterns are localized to specific brain regions that can be said to implement the particular neural processes instantiated by model components.

Materials and Methods

Participants

Nineteen participants completed the fMRI study; data from three of these participants were not included in the final analyses due to equipment malfunction and unreadable fMR images (distribution of the final sample: 7 males; *M* age = 25.7 yrs, *SD* age = 4.2 yrs). Nine additional participants completed a preliminary behavioral study (3 males; *M* age = 23.4 yrs, *SD* age = 2.2 yrs). Informed consent was obtained from all participants and all research methods were approved by the Institutional Review Board at the University of Iowa. All participants were right-handed, had normal or corrected to normal vision, and did not have any medical condition which would interfere with the MR machine.

Behavioral Task

Each trial began with a verbal load (two aurally presented letters lasting for 1000 ms; see

Todd & Marois, 2004). Then an array of colored squares (24 x 24 pixels; 2° visual angle) was presented for 500 ms (randomly sampled from CIE*Lab color-space at least 60° apart in color space). Squares were randomly spaced at least 30° apart along an imaginary circle with a radius of 7° visual angle. Next was a delay (1200 ms) followed by the test array (1800 ms). Trials were separated by a jitter of either 1.5s, 3s, or 5s selected in a pseudorandom order in a ratio of 2:1:1 ratio, respectively. On 'same' trials (50%), items were re-presented in their original locations. On 'different' trials, items were again re-presented in the original locations but the color of a randomly-selected item was shifted 36° in color space (see Figure 8A). Participants responded with a button press. On 25% of trials, the verbal load was probed (adding 500 ms to the trial; see Todd and Marois, 2004; *M* correct = 75%; *SD* = 13%). This ensured that participants could not use verbal working memory to complete the task (because verbal working memory was occupied with the letter task). Participants completed 5 blocks of 120 trials (3 blocks at SS4; 1 block each of SS2, SS6) in one of two orders (2,4,6,4,4; 6,4,2,4,4). Each block was administered in an individual scan that lasted for 1,040 s. A robust number of error trials were obtained at SS4 (FA: *M*=28.7, *SD*= 10.4; Miss: *M*=65.8, *SD*= 15.3) and SS6 (FA: *M*=12.9, *SD*=4.5; Miss: *M*=31.1, *SD*=6.4).

fMRI Acquisition

The fMRI study used a 3T Siemens TIM Trio system using a 12-channel head coil. Anatomical T1 weighted volumes were collected using an MP-RAGE sequence. Functional BOLD imaging was acquired using an axial 2D echo-planar gradient echo sequence with the following parameters: TE=30ms, TR=2000ms, flip angle=70°, FOV=240x240mm, matrix=64x64, slice thickness/gap=4.0/1.0mm, and bandwidth=1920Hz/pixel.

fMRI Preprocessing

Standard preprocessing was performed using AFNI (version 18.2.12) which included slice timing correction, outlier removal, motion correction, and spatial smoothing (Gaussian FWHM=5mm). The time series data were transformed into MNI space using an affine transform to warp the data to the common coordinate system. The T1-weighted images were used to define the transformation to the common coordinate system. T1 images were registered to the MNI_avg152T1+tlrc template. The coordinates for the regions of interest described by Wijeakumar et al. (2015) were used to define the centers of 1 cm^3 spheres. Since the time series data was mapped to a common coordinate system, the average time course for each participant was then estimated using the defined sphere.

Simulation Methods

Simulations were conducted as described above with the inputs modified to reflect the timing and stimuli properties (e.g., color separation) in the task given to participants. Initial observations indicated that the small metric changes in the task made detecting changes difficult in the model. Thus, to obtain better fits to the behavior data we changed one model parameters governing the resting level of the "different" node. For the previous simulations this value was -9, but for our version of the task with small metric changes we increased this value to -5 to be closer to threshold.

Behavioral Results and Discussion

Figure 8 shows the behavioral data from the preliminary behavioral study (Figure 8B), from the fMRI study (Figure 8C), and from the model (Figure 8D). Note that error bars were generated by running multiple iterations of the model and calculating standard deviation across runs. A two-way ANOVA (SS x Change trial) on the behavioral data from the fMRI study revealed main effects of SS ($F(2,15)=153.06$, $p< .001$) and Change trial ($F(1,16)=88.90$, $p< .001$)

and an interaction between SS and Change trial ($F(2,15)=10.98$, $p< .001$). Follow up t-tests showed that participants performed significantly better on SS2 compared to both SS4 ($t(16)= 16.29$, $p< .001$) and SS6 ($t(16)= 14.00$, $p< .001$), and better on SS4 compared to SS6 ($t(16)= 7.31$, $p< .001$). Participants performed better on Same trials compared to Different trials at SS2 ($t(16)= 3.843$, $p< .001$), SS4 ($t(16)= 8.47$, $p< .001$), and SS6 ($t(16)= 8.13$, $p< .001$). Importantly, all participants performed better than chance suggesting that they were not simply guessing (all t values > 4.5, $p < .001$).

The DF model that simulated data from Todd and Marois (2004) and Magen et al. (2009) also captured the data from the fMRI study and the preliminary behavioral study well (*RMSE* = 0.11 across both datasets) demonstrating that the model generalizes to behavioral differences across tasks (see Table 1). In summary, behavioral data from the present study show that participants generated many correct and incorrect responses, yet remained above-chance in all conditions. This provides an optimal data set, therefore, to test the neural predictions of the DF model regarding the origin of errors in change detection. The model did a good job reproducing these behavioral data with a single modification to a parameter across simulations (changing the resting level of the "different' node for our metric version of the task). This sets the stage to test the neural predictions of the DF model to determine whether the model can simultaneously capture both brain and behavior.

Testing Predictions of the DF Model with GLM

To test the hemodynamic predictions of the model, we adapted a general linear model (GLM) approach. As noted previously, it would be ideal to test the DF model against a competitor model, but no such competitor exists that predicts both brain and behavior. Instead, we asked whether the DF model out-performs the standard statistical modeling approach to fMRI

data using GLM.

In conventional fMRI analysis, a model of brain activity that has been parameterized for each stimulus condition is estimated via linear regression. A set of parametric maps for each condition is then constructed and used to infer locations in the brain where these model coefficients are statistically non-zero or different between conditions. The proposed innovation is to use the DF model to reparametize the GLMs because the DF model predicts the expected patterns across conditions. *The DF model in this case constitutes a task-independent and transferable bridge theory with the ability to make simultaneous task-specific predictions of both brain and behavior*. Note that this approach is novel relative to existing fMRI methods such as dynamic causal modeling (DCM; Penny, Stephan, Mechelli, & Friston, 2004) in that most common variants of DCM use deterministic state-space models while the DF model is stochastic (but see Daunizeau et al., 2012). Moreover, the DF model provides a direct link to behavioral measures while DCM does not (but see Rigoux and Daunizeau, 2015 for steps in this direction). More generally, DF and DCM have different goals with DCM using fMRI data to make hypothesis-led inferences about interactions among regions, and DF providing a predictive model of both brain and behavior.

The next question was how to apply the GLM-based approach to the brain. One option is an exploratory whole-brain approach. We opted, however, for a more constrained approach using a recent meta-analysis of the VWM literature (Wijeakumar et al., 2015). In particular, we extracted the BOLD response from 23 regions of interest (ROIs) implicated in fMRI studies of VWM. Twenty-one of these ROIs were from Wijeakumar et al. (2015); we added two ROIs so all bilateral entries were present with the exception of lSFG which was centrally located.

Consider what this GLM-based approach might reveal. It could be that specific model

components such as the WM field capture variance in just 1 or 2 ROIs. This would constitute evidence that the WM function was implemented in those cortical areas. It is also possible, however, that multiple components capture activation in the same ROI. In this case, we can conclude that multiple functionalities are evident in this ROI and the model does not unpack the specificity of the function. For instance, the CF and WM fields work together during the initial encoding and consolidation of the colors, while CF and the 'different' node conspire during comparison. In the brain, these functionalities might be handled by separate but coupled cortical fields. Indeed, we know this is the case already and have proposed a more complex DF architecture to pull functions like encoding and consolidation apart (see Schoner et al., 2015). Unfortunately, this new model is more complex, harder to fit to behavior, and has not been tested as fully as the model used here. We acknowledge up front, then, that there might be some lack of specificity in the mapping of model components to ROIs that suggests more work needs to be done to articulate what these brain regions are doing. Our hope is that the work we present here gives us a theoretical tool to use as we search for this more articulated understanding of VWM.

To determine whether the model statistically outperforms the standard task-based GLM approach and makes accurate predictions about activation in specific cortical regions, we used a Bayesian Multilevel Model (MLM) approach using equation 8 with d ROIs, N time points, and p regressors where Y is an N by d data matrix, X is an N by p design matrix, W is a p by d matrix of regression coefficients, and E is an N by d matrix of errors (using functions provided by SPM12). The errors, E, have a zero-mean Normal distribution with [d x d] precision matrix Λ.

$$Y = XW + E \tag{8}$$

A specific MLM can then be specified by the choice of the design matrix. In the following analyses, we use regressors derived from the DF model or sets of regressors capturing

the factorial design of the experiment (e.g., main effects of set-size, accuracy, same/different, or interactions thereof). A Variational Bayes algorithm (Roberts & Penny, 2002) was then used to estimate the model evidence for each MLM, $p(Y|m)$, and the posterior distributions over the regression coefficients $p(W|Y,m)$ and noise precision $p(\Lambda|Y,m)$. The model evidence takes into account model fit but also penalizes models for their complexity (Bishop, 2006; Penny et al., 2004). It can be used in the context of random effects model selection to find the best model over a group.

Methods

To assess the quality of fit between the predicted hemodynamic responses from the model components and the BOLD data obtained from participants, we first ran the model through the fMRI paradigm 10 times, calculating the average LFP timecourse for each model component ('different' node, 'same' node, CF, WM) on each trial type (same correct, same incorrect, different correct, different incorrect) for each set-size (2, 4, and 6). Figure 9 shows the full set of hemodynamic predictions for all trial types and components calculated from these LFPs (showing *M* HDR signal change for simplicity). To the extent that the model captures what is happening in the brain during change detection, we should see these same patterns reflected in participants' fMRI data. Note that these predictions are quite specific. For instance, as noted previously, the 'same' node shows a stronger hemodynamic response on hits than on correct rejections. This holds across memory loads. By contrast, the 'different' node shows a stronger hemodynamic response on hit and miss trials, except at the highest memory load where the strongest hemodynamic response is on false alarms. This reflects the strong 'different' signal on false alarm trials at high memory loads when a WM peak fails to consolidate. The other two layers in the model – CF and WM – show strong effects of the memory load, with an increase in

activation as the set size increases. Interestingly, differences across trial types emerge in WM as the memory load increases, with higher activation for miss and correct rejection trials, that is, when the model responds ‘same’.

Next, the LFP timecourses were turned into subject-specific time courses. These time courses were created by setting the time windows corresponding to each trial equal to the average LFP timecourse based on the timing and type of each trial for each participant. For each participant, four separate time courses were created corresponding to the LFPs from the ‘different’, ‘same’, CF, and WM model components. The variations in timing in the time courses for a participant reflect the random jitter between trials from the fMRI experiment, while the variations in the trial types reflect both the trial-by-trial randomization in trial types as well as participant’s performance—whether each trial was, for instance, a set size 2 ‘correct’ trial, a set size 4 ‘incorrect’ trial, and so on. The LFP time courses were then convolved with an impulse response function and down-sampled at 2 TR to match the fMRI experiment. Individual-level GLMs were first fit to each participant’s fMRI data. These results were then evaluated at the group level using Bayesian MLM.

Results

Categorical versus DF Model

In a first analysis, we generated standard task-based regressors that include the stimulus timing for each trial type. For example, a standard task-based analysis of the change detection task would model hemodynamic activation across voxels with regressors for correct-same trials, correct-change trials, incorrect-same trials, and incorrect-change trials at each set-size – 12 categorical regressors in total (4 trial types * 3 set sizes). To explore the full range of task-based models, we specified eight models based on combinations of task-based regressors: 1) a model

with three factors that categorize trials based on set-size, change, and accuracy (12 total task-based regressors), 2) a model with two factors that categorize trials based on set-size and change (6 total task-based regressors), 3) a model with two factors that categorize trials based on set-size and accuracy (6 total task-based regressors), 4) a model with two factors that categorize trials based on change and accuracy (4 total task-based regressors), 5) a model with one factor that categorizes trials based on set-size (3 total task-based regressors), 6) a model with one factor that categorizes trials based on change (2 total task-based regressors), 7) a model with one factor that categorizes trials based on accuracy (2 total task-based regressors), and 8) a null model (1 constant regressor). For all of these models, the hemodynamic response at each trial was modeled based on the GAM function in AFNI.

Second, we generated regressors from the four components of the DF model as described above. Note that all nine models were individualized based on the specific sequence of trials for each participant. Additionally, all models included 6 regressors based on motion (roll, pitch, yaw, translations right-left, translations inferior-superior, and translations anterior-posterior), 6 regressors based on the motion regressors with a time lag of 1 TR, and 25 baseline parameters reflecting a 4 degree polynomial model for the baseline of each of the five blocks. Lastly, all models were normalized to have zero-mean unit variance among columns prior to model estimation (for each column, the mean was subtracted and then divided by the standard deviation).

Random Effects Bayesian Model Comparison (Rigoux et al., 2014; Stephan et al., 2009) was then implemented across all models and participants using the statistical function provided by SPM12. This method uses the concept of model frequencies, which are the relative prevalence of models in the population from which the sample subjects were drawn. For example, model

frequencies of 0.90 and 0.10 indicate a prevalence of 90 percent for model 1 and 10 percent for model 2. Random Effects Bayesian Model Comparison provides for statistical inferences over model frequencies and Stephan et al. (2009) describe an iterative algorithm for computing them. Initial inspection of the data revealed that frequencies were non-uniform (Bayes Omnibus Risk (BOR) = 4.78×10^{-5}). The DF model, accuracy categorical model, and change categorical model had the largest frequencies of 0.44, 0.20, and 0.12, respectively. The probability that the DF model had the highest model frequency (quantified using the "protected exceedance probability") is PXP = 0.9312. This value is a posterior probability so has no simple relation to a classical p-value. One can also express posterior probabilities as Bayes Factors, with the Log Bayes Factor being the log-odds of the marginal likelihoods. For example, for PXP=0.9312, the log Bayes Factor is log [0.9312/(1-0.9312)]=2.61 and the Bayes Factor is exp(2.61)=13.5, meaning there is 13.5 times the evidence for the statement than against it. Conventionally, a Bayes factor of 1 to 3 is considered "Weak" evidence, 3 to 20 as "Positive" evidence, and 20 to 150 as "Strong" evidence (Kass & Raftery, 1995). It is in this sense that the DF model "best" explains the fMRI data. In our group of 16 participants, the posterior model probabilities were highest for the DF model for 10 individuals, the accuracy categorical model for 4 individuals, and the change categorical model for 2 individuals. Table 2 shows the log Bayes Factors for the different models across participants. These results indicate that some individuals showed differences in activation across accuracy or change factors that were not effectively captured by the DF model.

Testing the Specificity of the DF Model

It is an open question to what extent the dynamics implemented by the model are important for its explanatory value in the MLM results. One of our key claims is that the neural

dynamics that are implemented in the model provide an explanation of what the brain is doing to give rise to same/different decisions in the change detection task, both on correct and incorrect trials. To probe this issue, we generated new sets of four randomized DF regressors and re-ran the MLM analysis. For each participant and for each trial, an LFP was selected from a randomly determined trial type and component. These LFPs were slotted in based on the timing of trials for each individual participant and then convolved to generate sets of 4 DF model regressors as described above. We refer to this as the Random Trial and Component DF Model (DF-RTC). If the structure of activation within each component for each trial type is important for the explanatory value of the model, then this model should do poorly compared to the categorical model.

Results from the MLM analysis showed that observed model frequencies were non-uniform (BOR = 1.20×10^{-5}). In contrast to the prior analysis, the DF model was not the most frequent; rather, the accuracy categorical model, change categorical model, and DF-RTC model had the largest frequencies of 0.47, 0.21, and 0.08, respectively. The probability that the accuracy categorical model had the highest model frequency is PXP = 0.9459. Thus, in this new analysis, the accuracy categorical model best explains the fMRI data. In our group of 16 participants, the posterior model probabilities were highest for the accuracy categorical model for 11 individuals, the change categorical model for 2 individuals, and the DF-RTC model for 1 individual. Importantly, these results show that the DF-RTC model regressors poorly explain the fMRI data when the trial and component structure is removed.

Next, we asked whether preserving the component structure but disrupting the trial structure would impact the explanatory power of the DF model. To accomplish this, we generated new sets of four DF regressors for each participant. In particular, an LFP was selected

from a randomly determined trial type on each trial, but each regressor was sampled from a single component to maintain the ingrity of the component-level predictions. As before, these LFPs were inserted into the predicted time series based on the timing of each individual trial for each participant, the individual-level GLMs were re-estimated, and the MLM analysis was repeated at the group level. We refer to this as the Random-Trial DF model (DF-RT). If the specific structure of activation pattern across trials within each component is important for the explanatory value of the model, then this model should do poorly compared to the categorical model. If, however, this model still captures the data well, then this would suggest that the relative differences in activation dynamics across components are an important contributor to the model's explanation of the data.

In this new analysis, we observed that frequencies were non-uniform (BOR = 4.78×10^{-5}). The DF-RT model, accuracy categorical model and change categorical model had the largest frequencies of 0.44, 0.20, and 0.12, respectively. The probability that the DF-RT model has higher model frequency than any other model is PXP = 0.9312. Thus, the DF-RT model still "best" explains the fMRI data. In our group of 16 participants, the posterior model probabilities were highest for the DF-RT model for 12 individuals, the accuracy categorical model for 2 individuals, and the change categorical model for 2 individuals.

We then asked whether the "standard" DF model provides a better fit to the data than the DF-RT model. In this comparison, we observed that frequencies were non-uniform (BOR = 0.0396). The standard DF model had a frequency of 0.83 which was higher than the DF-RT model frequency of 0.17 (PXP = 0.9789). Thus, the standard DF model better explains the fMRI data compared to the random-trial DF model. In our group of 16 participants, the posterior model probabilities were highest for the standard DF model for 15 individuals. Thus, the detailed

predictions of the DF model regarding how brain activity varies over trial types is, in fact, important in capturing the fMRI data from the present study.

Are all DF model components necessary?

The correlation among the DF regressors was very high, most likely reflecting the strong reciprocal connections between model components. Averaged over the group, the maximum correlation was between CF and WM (r^2 = .93) and the minimum was between the 'same' node and WM (r^2= .52). Thus, it is important to assess whether all model components are adding explanatory value.

We compared the model evidence of the full model with all four regressors to four other models that eliminated one regressor. These results indicated that removing the 'different' node regressor yielded a better model. Specifically, the frequency of this model was higher than the other four models that were compared (PXP = .9998). The model frequencies were non-uniform (Bayes Omnibus Risk (BOR) = 5.72×10^{-7}) indicating a very low probability that the model frequencies are equal (this is a posterior probability and can also be converted into a Bayes Factor as above). We examined whether further reducing the model would yield a better model. We compared the model with the 'different' node regressor removed to three other models with one of the remaining three regressors removed. Results from this comparison indicated that the model with three regressors had the highest model frequency, PXP = 1.0000, and the model frequencies were significantly non-uniform (BOR = 2.26×10^{-7}). From this we concluded that the best variant of the DF model across participants was a 3-regressor model with CF, WM, and 'same' regressors included.

In an additional MLM analysis, we examined how the reduced model compared to the set of categorical models described above. We observed that frequencies were non-uniform (BOR =

4.34 x 10^{-6}). The DF model, accuracy categorical model, and change categorical model had the largest frequencies of 0.52, 0.12, and 0.12, respectively. The probability that the DF model has the highest model frequency is PXP = 0.9922. Thus, the reduced DF model still "best" explains the fMRI data. In our group of 16 participants, the posterior model probabilities were highest for the DF model for 12 individuals, the accuracy categorical model for 2 individuals, and the change categorical model for 2 individuals (see Table 2).

To explore why the 'different' node regressor failed to contribute much to model performance, we explored the multicollinearity of the four DF regressors using Belsley collinearity diagnostics (Belsley, 1991). This revealed that the three remaining regressors were multicollinear (variance decompositions larger than .5), and that the 'different' node was independent of this collinearity (condIdx=56.97; 'different'= 0.3155, 'same'= 0.8212, CF= 0.9945, WM= 0.9811). Interestingly, when we examined the connection weights between the 'different' node and the regions of interest, all of the regions with relatively large 'different' weightings had negative weights. Thus, the 'different' hemodynamics in the model appear to be relatively distinct and inversely mapped to brain hemodynamics. This may indicate that difference detection in the model is too simplistic. For instance, evidence suggests that people typically both detect changes in the test array and shift attention to the changed location (Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009); this second operation is not captured by our model.

Mapping model components to cortical regions

The analyses thus far indicate that the DF model provides a better account of the fMRI data than 8 standard categorical models, the trial type and component structure of the DF model regressors *both* matter to the quality of the data fits, and a streamlined 3-regressor DF model

provides the most parsimonious account of the data. Our next goal was to understand how the model maps onto specific brain regions and which aspects of the fMRI data the model captures. In this context, it is important to emphasize that the beta weights for each component of the model are estimated together along with the other components that are being considered. That is, neural activation in a ROI is the dependent variable and the predicted neural activation from the model components are the independent variables. Since the model components are entered into the model together, the beta weight estimated for each component controls for the other predictors. At the group level, described below, the statistical comparison was performed individually on each component using a t-test. Here, the question is whether each component contributes significantly to prediction at the group level, allowing for inferences to be made about the partial correlation between each model component and each region of interest.

We performed group-level t-tests (Bonferroni corrected) to examine which of the three components from the reduced DF model explained activation in different cortical regions across our group of participants. We focused on connections that were positive. Note that negative connections were observed (i.e., 'same' node: lIFG, lIPS, lOCC, lSFG, lsIPS, rIFG, rMFG, rOCC, rsIPS; CF: lTPJ, rTPJ; WM: alIPS, lIFG, lIPS, lsIPS, rIFG, rsIPS). In all but one case (rMFG), a negative connection was paired with a positive connection with another component. Thus, negative connections could be explained by the inverse nature of different components involved in "same" and "different" decisions, in which case it is easier to interpret the positive connection weights. The CF component explained significant activation in 9 regions (alIPS: $t(14)=4.85$, $p< .001$; lIFG: $t(14)= 5.65$, $p< .001$; lIPS: $t(14)= 5.60$, $p< .001$; lOCC: $t(14)= 4.41$, $p< .001$; lSFG: $t(14)= 4.67$, $p< .001$; lsIPS: $t(14)= 4.94$, $p< .001$; rIFG: $t(14)= 5.56$, $p< .001$; rOCC: $t(14)= 4.32$, $p< .001$; and rsIPS: $t(14)= 6.79$, $p< .001$). Additionally, WM and the 'same'

node explained activation in lTPJ ($t(14)= 8.25$, $p< .001$; $t(14)= 7.83$, $p< .001$) and rTPJ ($t(14)= 9.59$, $p< .001$; $t(14)= 7.89$, $p< .001$). Figure 10A shows the mapping of model components to cortical regions. A first observation from this pattern of results is that bilateral IPS is once again mapped to the contrast layer, consistent with our first simulation experiment. In addition to IPS, the CF regressor also captured significant variance in other regions associated with the dorsal frontoparietal network including bilateral OCC and IFG, as well as another brain region commonly linked to an aspect of the ventral right frontoparietal network –SFG (Corbetta & Shulman, 2002).

Given the striking presence of bilateral activation in the t-test results, we tested if the regression vectors were significantly different between paired regions across hemispheres. In our sample of ROIs there were 11 such regions (e.g., left/right IPS, left/right IFG, etc). We tested for differences within-subject using the Savage-Dickey (Rosa, Friston, & Penny, 2012) approximation of model evidence and then examined consistency over the group (using random effects model comparison). For all pairs, no log Bayes Factors were decisively negative. This indicates that the regression vectors were different. Thus, although both hemispheres may be engaged in the same type of function (e.g., contrasting items with the content of VWM), activation profiles between hemispheres differ. This is consistent with data suggesting, for instance, that IPS might be most sensitive to visual information in the contralateral visual field (Gao et al., 2011; Robitaille, Grimault, & Jolicœur, 2009).

To asses the quality of the data fits between the DF regressors and activation in these brain regions, we plotted the predicted data from the model against the fMRI timecourses. We selected three regions of interest – rTPJ which was mapped to the WM+Same component across the group (Figure 10B), lIPS which was mapped to the CF component across the group (Figure

10C), and a contrast area -- lMFG -- which was not robustly mapped to any component (Figure 10D). In each panel, we show an example plot from one individual who 'preferred' the DF model based on our MLM analysis along with data from one individual who 'preferred' the accuracy categorical model. The annotation in each figure (see green ovals) show time epochs where the preferred model showed a better fit to the empirical data. For instance, in the top panel of Figure 10B, there are several epochs where the DF model fit the empirical data better; by contrast, the categorical model generally shows a negative undershoot relative to the data. In the lower panel, however, there is a run of trials where the categorical model provides a better fit. Figure 10C shows comparable results, with clear time epochs where the DF model (top panel) or categorical model (lower panel) provides a better data fit. Finally, in Figure 10D, one can see two examples where neither model fits the BOLD data particularly well.

To explore individual differences in further detail, we examined whether the connection values (i.e., β weights) between model components and cortical regions were correlated (Spearman's correlation) with an individual's WM performance as indexed by the maximum value of Pashler's *K*. Note that our sample size of 16 may not be large enough to provide strong evidence of brain-behavior relationships. Further, we tested only positive β weights between regions and model components (13 total comparisons). Using the Benjamin-Hochberg (Benjamini & Hochberg, 1995) correction procedure and a false-discovery rate of .1 (given the exploratory nature of these comparisons), we found that capacity was significantly correlated with the connection weight between WM and lTPJ ($r=-0.66$, $p=.0055$; Figure 10E). As is evident in the scatter plot, higher capacity individuals show weaker β weights for the WM component in lTPJ. Recall from the behavioral data in Figure 8C that performance drops over set sizes, particularly in the 'different' condition; thus, higher capacity individuals (who had the highest

percent correct) show less of a 'same' bias and more selective responding on 'different' trials. This is consistent with the correlation in TPJ: higher capacity individuals show a weaker 'same' bias in TPJ (negative correlations between brain activation and the WM regressor).

Assessing the quality of the mapping of model components to cortical regions

One way to evaluate the mapping of model components to cortical regions was shown in Figure 10, where we highlighted both group-level data as well as data from individual participants. While this is helpful in evaluating model fits, in this final section we use a quantitative metric to help understand what details of the data the DF and categorical models are explaining.

To quantitatively assess the quality of the fit for the DF model relative to the categorical models, we examined the precision of the different models. Precision was derived from the inverse covariance matrix for each model. Specificially, given the linear model $Y = XW + E$ where the errors have covariance matrix C, the corresponding precision matrix is Λ (the inverse of C). The precision metric reflects the partial correlation between variables independent of covariation with other variables (Varoquaux & Craddock, 2013). We defined a diagonal version of the precision matrix to get region by region precisions: $\lambda = \text{diag}(\Lambda)$ such that $\lambda(r)$ is high if the model fit is good in region r, that is, if a lot of unique variance is captured in this region. Improvements in model precision were calculated as the relative percent improvement in precision for the DF model relative to the different categorical models. For instance, we can calculate the precision of the DF model for subject 1 in left IPS, the precision of a categorical model for subject 1 in left IPS, and then compute the relative percent increase (or decrease) in precision for the DF model.

Figure 11A shows the average improvement in precision over subjects for the 23 brain

regions. The arrows below specific regions highlight the mapping of model components to regions shown in Figure 10A (yellow arrows = CF, red arrows = WM+Same). As can be seen in the figure, regions mapped to specific DF regressors in the group-level comparisons generally showed a large relative increase in precision for the DF model (positive values). Interestingly, some regions such as rFEF showed a large average change in precision even though this region was not mapped to a particular DF component in the group-level t-tests. Figure 11B shows the average improvement in precision over regions split by participants. The arrows below specific participants indicate the participants that 'preferred' a categorical model in the MLM analysis. These participants all have small changes in relative precision, indicating that the precision of the DF model was only slightly higher than the precision of the categorical model. Considered together, then, the data in Figure 11 largely mirror the group-level results that mapped DF components to ROIs as well as the MLM results showing which models were preferred by which subjects.

Critically, the precision for some regions for the categorical-preferring participants showed higher precision for the categorical model of interest. This can give us a sense of what the DF model is failing to capture. Figure 12 shows two exemplary participants. Figure 12A shows data from subject 1 -- a DF 'preferring' participant with high relative precision in rIPS (relative precision = 1.7527), while Figure 12B shows data from subject 8 – a categorical 'preferring' participant with a negative relative precision in this same ROI, that is, higher precision for the change categorical model (relative precision = -1.9862). Each panel shows the BOLD data, the DF time series predictions, and the categorical time series predictions with the data split by trial types. All time traces were constructed by averaging the time series data from trial onset (0s) through 10 s post-trial onset, where data were baselined at 0s.

As can be seen in Figure 12A, the DF-preferring participant showed a large hemodynamic response in the SS2-correct conditions as well as a large hemodynamic response on all SS6 trial types (bottom row). This highlights how brain activity is modulated by the memory load. Note that the SS4 condition had the most trials; this appears to have reduced the magnitude of the response (note the scale difference in the middle row). Comparing the DF time series data with the categorical time series data, the DF model data are closer to the empirical values for all SS2 trials, for SS4-same-correct trials, and SS6-same trials (both correct and incorrect), with mixed results in the other conditions. Thus, in this region, the DF model is doing relatively well, with weaker performance on high set size change trials. Note that the amplitude of the model predictions are low in all cases reflecting the limited degrees of freedom in the overall model (only 3 predictors).

In Figure 12B, we see a similar modulation in the HDR over SS, although this participant shows a robust HDR across all SS2 conditions (top row). Looking at the relative accuracy of the DF and categorical time series data, the top row shows mixed results with one exception -- the DF model is closer to the data on the SS2-different-incorrect trials. The categorical model generally fares better on the SS4 trials (middle row). SS6 is again mixed with the categorical model closer to the BOLD data, particularly early in the trial. Interestingly, even though this region showed high precision for the categorical model, this improved fit is subtle. We conclude, therefore, that the DF model is generally doing reasonably well -- even with categorical-preferring participants – and is not overtly failing on a small subset of conditions.

Finally, we examined the differences between the observed BOLD data measured from rIPS relative to the DF and categorical models. Here, we focused on the accuracy and change categorical models since these were the only categorical models ‘preferred’ by any participants.

First, we computed the average absolute difference between the DF model and the BOLD signal and the categorical models and the BOLD signal to determine how much these models deviated from the observed BOLD signal for each trial type. Plotted in Figure 13 is the difference in deviation between the two categorical models and the DF model averaged across participants. This visualization provides a sense of which trial types the DF model did well (where ther are large positive values in Figure 13) and where the DF model did poorer (where there are negative values in Figure 13). As can be seen, the DF model does very well relative to these categorical models on incorrect trials at SS2 and across trial types for SS6. Most notably, the DF model does the worst on correct change trials at SS2. Note, however, the degree of difference for this trial type is small relative to the degree of difference on other trial types in which the DF model does better.

General Discussion

The central goal of the current paper was to test whether a neural dynamic model of visual working memory could directly bridge between brain and behavior. We initially fit a model that simulates behavioral and hemodynamic data simultaneously to data from two fMRI studies that reported seemingly contradictory findings. The model simulated results from both studies. Interestingly, simulated results from the model's contrast layer most closely mirrored fMRI data from IPS, suggesting that IPS plays more of a role in comparison and change detection than in the maintenance of items in VWM. Moreover, the model explained why IPS fails to show an asymptote in a long-delay paradigm – the longer-delay allows for more subtle variations in the neural dynamics of the contrast layer to be reflected in the hemodynamic response.

We then used a Bayesian MLM approach to test model predictions against BOLD data

from a set of ROIs to assess the fit of the model's predicted patterns of hemodynamic activation. This method was used to shed light on the mechanisms that underlie VWM and change detection performance with a special emphasis on the neural processes that underlie errors in change detection. Results showed that the model-based regressors explained more variance in the BOLD data than standard task-based categorical regressors. Additional analyses showed that both the component structure of the model and the details of neural activation on each trial type mattered to the quality of the data fits. Evidence that the trial types matter is important because the DF model offers a novel account of why people make errors in change detection. In particular, the model predicts a false alarm when an item is not maintained in WM and a miss due to decision errors caused by widespread suppression of the contrast layer. By contrast, previous cognitive accounts hypothesized that *misses* occur when items are not maintained in WM and *false alarms* reflect decision errors / guessing (Cowan, 2001; Pashler, 1988). The fMRI data support the DF account.

The model-based fMRI approach not only provided robust fits to the BOLD data in specific ROIs, this approach also conferred new understanding of the neural bases of VWM. In particular, group-level analyses mapped model components to patterns of activation in specific regions of the brain and *this mapping offers an explanation of the functional significance of this brain activity.* Although our results here are still correlational in nature, future work could use methods such as TMS to more directly probe model predictions that can push this explanation to the causal level. Notably, once again, the contrast layer provided the best account of data from IPS. This helps resolve on-going debates in the literature. Previous work has suggested IPS is a critical site for VWM because this area shows an asymptote in the BOLD signal at higher set sizes (Todd & Marois, 2004) while other work suggests IPS plays an attentional role

(Szczepanski, Pinsk, Douglas, Kastner, & Saalmann, 2013). Our results provide a new account of these data suggesting that IPS is critically involved in the comparison operation. This highlights how a model-based fMRI approach can lead to an integrated account when current experimental results have yielded contradictory findings.

More generally, the contrast field provided a robust account of neural activation across 10 regions linked to a dorsal frontoparietal network as well as key regions in a ventral right frontoparietal network (Corbetta & Shulman, 2002). One critique of the model is that it failed to make functional distinctions across these 10 ROIs. We suspect this reflects the simplicity of the model tested here. The model only had four components. While results show that these components were sufficient to capture key aspects of the behavioral and neural dynamics in the task, the model does not specify all the processes that underlie participants' performance. For instance, in the current model, encoding and comparison both happen in the CF layer. In a more recent model of VWM and change detection (Schneegans, 2016; Schneegans, Spencer, Schoner, Hwang, & Hollingworth, 2014), we have unpacked these functions by including new cortical fields that implement encoding within lower-level visual fields as well as attentional fields that capture known shifts of attention that occur in change detection. If we were to test this more articulated VWM model using the tools developed here, it is possible that some of the CF ROIs like OCC would now show an encoding function while other ROIs like SFG would be mapped to an attentional function. Future work will be needed to explore these possibilities. Importantly, this work can directly use all of the tools developed here.

Another key result in the present paper was the mapping of the WM and 'same' functionalities to brain activation in bilateral TPJ. The link between WM and TPJ is consistent with previous fMRI studies (Todd et al., 2005). Moreover, we found significant correlations

between the WM and 'same' β weights in rTPJ and individual differences in WM capacity. Although this suggests TPJ is a central hub for VWM, one could once again critique the specificity of the model predictions: shouldn't the model reveal a neural site for VWM that is distinct from activation predicted by the 'same' node? We suspect there are two key limitations on this front. First, as noted above, the model is relatively simple. In our recent model of VWM, for instance, we tackle how working memories for features are bound to spatial positions to create an integrated working memory for objects in a scene that is distributed across multiple cortical fields. Moreover, working memory peaks in this new model build sequentially as attention is shifted from item to item. This leads to differences in the neural dynamics of working memory through time that are not captured by the model used here (which builds peaks in parallel). It is possible that this more articulated model of VWM would help pull part the details of neural processing in TPJ, potentially capturing data in other brain regions as well that the current model failed to detect.

A second limitation of the present work was hinted at by our simulations of data from Magen et al. (2009). Those simulations show that short-delay change detection paradigms may provide only limited information about the neural dynamics that underlie VWM because subtle variations in the dynamics are not detected in the slow hemodynamic response. We suspect this contributed to the high collinearity of our model regressors which ultimately contributed to the removal of the 'different' regressor in our final model. That is, the design of the task may not have been optimized to elicit distinguishing patterns of activation from the model components. One way to reduce collinearity in future model-based fMRI would could be to vary the task. If, for instance, the model was put in a variety of task settings, including both short-delay and long-delay trials as well as variations in the memory load, the collinearity would likely reduce. Indeed,

one advantage of having a neural process model is that the properties of the design matrix could be optimized in advance by simulating the model directly. To explore the relationship between model dynamics and hemodynamics in more detail, we ran additional simulations in which we varied the timing parameters of the canonical HRF function used to generate hemodynamics from the simulated LFP (see supplemental figure). This illustrates how future work can use an iterative process to not only inform interpretation of neural data but to influence the parameters used in the model.

In summary, although there are limitations to our findings, the integrative cognitive neuroscience approach used here opens up a new way to assess how well a particular class of neural process theories explain and predict functional brain data *and* behavioral data. In this regard, the DF model presents a bridge between cognitive and neural concepts that can shed new light on the functional aspects of brain activation.

Relations between the DF model and other theoretical accounts

DFT provides a rich computational framework that generates novel predictions not explained by other accounts focusing on slots and resources (Bays et al., 2009; Bays & Husain, 2009; Brady & Tenenbaum, 2013; Donkin et al., 2013; Kary et al., 2016; Rouder et al., 2008b; Sims et al., 2012; Wilken & Ma, 2004). One novel prediction previously reported using a DF model of VWM demonstrated enhanced change detection performance for items in memory that are metrically similar (Johnson, Spencer, Luck, et al., 2009). Other more recent models have also addressed metric effects. For example, Sim, Jacobs, and Knill (2012) explain such effects in terms of informational bits contained in the memory array. Items that are more similar to one another contain fewer informational bits, leading to items being encoded more precisely and change detection performance is improved. The model reported by Oberauer and Lin (2017)

implement neural processes that explain the benefits of have similarity between items in VWM. In this case, the benefit arises from the partial overlap of representations in VWM which mutually support one another. This contrasts with the explanation offered by the DNF model which suggests that benefits in performance arising from item similarity are due to the sharpening of representations through shared lateral inhibition (Johnson, Spencer, Luck, et al., 2009).

Here, we extended the DF model to also generate novel neural predictions which no other behaviorally-grounded model of VWM has achieved. Beyond the capability to generate both behavioral and neural predictions, the DF model of change detection is also the only model that specifies the neural processes that underlie comparison (Johnson et al., 2014). Swan and Wyble (2014) implement a comparison process in their model that calculates the difference between items held in VWM and items displayed in the test array. This calculation results in a vector whose angle is the degree of difference between a memory item and test display item and whose length is the confidence that the model has about the accuracy of that difference calculation. To make a 'change' decision, the vector must be sufficiently different and sufficiently confident. The response that the model generates is determined by an algorithm that sets thresholds on these two values which linearly scale with SS. Swan and Wyble (2014) also demonstrated how this same process could generate color reproduction responses, suggesting this a general process that can be used to both recollect items from memory and compare the recollected value with an available perceptual input. It should be noted that the DF model engages in a similar comparison process, but generates active neural responses based on non-linear neural dynamics without the need for a separate comparison algorithm.

More recent debates about whether VWM is best explained via slots or resources have

examined color reproduction responses. Other variations of neural models discussed above have simulated these type of data using neural units that bind features and spatial locations (Oberauer & Lin, 2017; Swan & Wyble, 2014). In these models, the spatial or featural cue in the task is used to recollect a color or line orientation value from memory. Although the model we presented here has not been used to generate color reproduction responses, the model can be adapted in this direction (Johnson et al., 2014). For example, Johnson et al. (submitted) tested a novel prediction of the DF model that similar items in VWM should be repelled from one another during short-term delays and this should be reflected in color reproduction estimates. Recent extensions of the DF model have also been used to explain how object features are bound into integrated object representations (Gregor Schoner et al., 2016).

Although there are ways in which DFT is unique, it also shares considerable overlap with other theories. The neural mechanism of self-sustaining activation is similar to the mechanism used in models proposed by Edin and colleagues (Edin et al., 2009, 2007) and Wang and colleagues (Compte et al., 2000). Additionally, capacity limitations in the DF model arise from competitive dynamics instantiated through inhibition among active representations, similar to the neural model reported by Swan and Wyble (2014). The model also overlaps with concepts from the slots and resources frameworks. Specifically, the non-linear nature of peak formation bears similarity to the qualitative nature of slots. Relatedly, the width of peaks and their shifting over time leads to spread of variance that is consistent with resource accounts. Moreover, the gradual rise in activation for each peak is consistent with the idea of the gradual accumulation of information over time in resource models. It is notable that there are inconsistencies regarding whether a slots or a resources account fit different datasets (Sim, Jacobs, & Knill, 2012; Rouder et al., 2008; Donkin et al., 2013; van den Berg & Ma, 2017). Since the DF model has aspects

consistent with both approaches, the model may have the flexibility needed to bridge these disparate findings in the literature (for discussion, see Johnson et al., 2014).

The DNF model presented here is relatively simple, but has been extensively used to examine VWM from childhood to older adults (Costello & Buss, 2018; Johnson, Spencer, Luck, et al., 2009; Simmering, 2016). Other applications have implemented a more elaborated model that captures aspects of visual attention, saccade planning, and spatial-transformations (Ross-Sheehy, Schneegans, & Spencer, 2015; Schneegans, 2016; Schneegans et al., 2014). These models incorporate a similar network to the model we presented here, but embedded it within a broader architecture that binds visual features to multiple different spatial frames of reference and performs spatial transformation across these reference frames (Schneegans, 2016). For example, this DF model architecture has been used to explain how VWM is updated across how eye movements (Schneegans et al., 2014) and how spontaneous exploration of an array of visual stimuli can build a representation of a scene (Grieben et al., 2020). Other applications have explained how change detection can occur if the same color occupies multiple spatial locations and how changes can be detected if two colors swap locations as well as differences in performance across these scenarios (Schneegans et al., 2016). Future work using the model-based fMRI methods we describe here can explore how this fuller architecture accounts for patterns of cortical activation.

Limitations and future directions

It is important to highlight several limitations of the integrative cognitive neuroscience approach used here as well as future directions. One issue that will need to be addressed by future work is the strong collinearity between regressors generated from the model. The DF model is dominated by recurrent interactions meaning that many properties of the pattern of

activation, such as the timing or duration, are likely to be shared across components. The approach used here could be strengthened by designing tasks that are not only optimized for fMRI but also optimized from the perspective of the theory to be tested so that the regressors from a model are as independent as possible.

Beyond such challenges, this work presents an important step forward in understanding brain-behavior relationships that opens up new avenues of future research. In particular, we are currently using this method to determine if model-based fMRI can adjudicate between competing neural process models to determine which provides a better explanation of brain data. If different models use different neural mechanisms or processes to produce the same pattern of behavior, can the Bayesian MLM and model-based fMRI methods be used to determine which model provides a better explanation of the functional brain data?

We are also exploring the transferability of models. One way to achieve this might be to use one model to simulate two different tasks. If cortical fields implemented by the model correspond directly to processing in specific ROIs, we would expect the same field to map onto the same ROI across tasks. However, it is also possible that the function of specific cortical fields might be softly-assembled from interactions among different ROIs in the brain. In this case, the function implemented by a cortical field might correspond to different ROIs across different tasks. This exploration can determine whether the architecture of a model reflects the architecture of the brain, or if the functional mapping is more complex.

Future work can also explore the relationship between the DF model and the large body of work examining VWM processes with EEG and ERP. Such efforts would complement the work presented here by evaluating the fine-grained temporal predictions of the model. The model is implemented with distinct neural processes corresponding to excitatory and inhibitory

interactions; thus, the model is well-positioned to generate simulated voltage changes and previous reports have provided initial comparisons between DF model activation and electrophysiological measures (Spencer, Barich, Goldberg, & Perone, 2012).

Lastly, we are also exploring the brain-behavior relationship using other metrics of behavioral performance. In this project, we focused on accuracy as a measure of performance; however, reaction time can also be informative of the processes underlying VWM. Although the model's behavior unfolds in real-time and previous DF models have been used to simulate reaction time as a target beahvior, the current model was not optimized to fit patterns of reaction time, nor was the task optimized to reveal differences in reaction times across memory loads. Future work can use this behavioral metric to further constrain model parameters and potentially reveal novel aspects of the neural dynamics of VWM.

In conclusion, the DF account of VWM and change detection links behavioral and neuroimaging data in a new – and direct – way. We showed how a model that was initially constrained by behavioral data predicted patterns of fMRI data from a novel change detection paradigm, outperforming standard methods of analysis. The predicted and experimentally confirmed neural signatures of both correct and incorrect performance shed new light on the functional role of IPS, as well as lending support to the role of the TPJ in VWM maintenance. Critically, these functional neural signatures provide support for the neural dynamic account, contrasting with classic accounts of the origin of errors in change detection upon which more recent models are based. The model-based fMRI approach also raises new questions. For instance, how specific is the mapping between the different activation fields in the DF architecture and cortical sites in the brain? Integrating multiple different tasks within a single model and a single neural data set may be a way to address such questions about the mapping of

brain function to neural architecture.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87.

Ambrose, J. P., Wijeakumar, S., Buss, A. T., & Spencer, J. P. (2016). Feature-based change detection reveals inconsistent individual differences in visual working memory capacity. *Frontiers in Systems Neuroscience*, *10*(APR). https://doi.org/10.3389/fnsys.2016.00033

Anderson, J. R., Albert, M. V., & Fincham, J. M. (2005). Tracing Problem Solving in Real Time: fMRI Analysis of the Subject-paced Tower of Hanoi. *Journal of Cognitive Neuroscience*, *17*(8), 1261–1274. https://doi.org/10.1162/0898929055002427

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Anderson, J. R., Carter, C., Fincham, J., Qin, Y., Ravizza, S., & Rosenberg-Lee, M. (2008). Using fMRI to Test Models of Complex Cognition. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1323–1348. https://doi.org/10.1080/03640210802451588

Anderson, J. R., Qin, Y., Jung, K.-J., & Carter, C. S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology*, *54*(3), 185–217.

Anderson, J. R., Qin, Y., Sohn, M.-H., Stenger, V. A., & Carter, C. S. (2003). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic*

Bulletin & Review, *10*(2), 241–261. https://doi.org/10.3758/BF03196490

Ashby, F. G., & Waldschmidt, J. G. (2008). Fitting computational models to fMRI data. *Behavior Research Methods*, *40*(3), 713–721.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622–628. https://doi.org/10.1111/j.1467-9280.2007.01949.x

Bastian, A., Riehle, A., Erlhagen, W., & Schöner, G. (1998). Prior information preshapes the population representation of movement direction in motor cortex. *Neuroreport*, *9*(2), 315–319.

Bastian, A., Schoner, G., & Riehle, A. (2003). Preshaping and continuous evolution of motor cortical representations during movement preparation. *European Journal of Neuroscience*, *18*(7), 2047–2058.

Bays, P. M. (2018). Reassessing the Evidence for Capacity Limits in Neural Signals Related to Working Memory. *Cerebral Cortex*, *28*(4), 1432–1438. https://doi.org/10.1093/cercor/bhx351

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 1–11. https://doi.org/10.1167/9.10.7.Introduction

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851–854. https://doi.org/10.1126/science.1158023

Bays, P. M., & Husain, M. (2009). Response to Comment on "Dynamic Shifts of Limited

Working Memory Resources in Human Vision." *Science*, *323*(5916), 877d-877d. https://doi.org/10.1126/science.1166794

Belsley, D. A. (1991). A Guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, *4*(1), 33–50. https://doi.org/10.1007/bf00426854

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1628–1633. https://doi.org/10.1073/pnas.1221572110

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. R. (2015). Using Data-Driven Model-Brain Mappings to Constrain Formal Models of Cognition. *PLOS ONE*, *10*(3), e0119673. https://doi.org/10.1371/journal.pone.0119673

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Brunel, N., & Wang, X.-J. (2001). Effects of Neuromodulation in a Cortical Network Model of Object Working Memory Dominated by Recurrent Inhibition. *Journal of Computational Neuroscience*, *11*(1), 63–85. https://doi.org/10.1023/A:1011204814320

Buss, A. T., & Spencer, J. P. (2014). The emergent executive: A dynamic field theory of the development of executive function. *Monographs of the Society for Research in Child Development*, *79*(2). https://doi.org/10.1002/mono.12096

Buss, A. T., & Spencer, J. P. (2018). Changes in frontal and posterior cortical activity underlie the early emergence of executive function. *Developmental Science*. https://doi.org/10.1111/desc.12602

Buss, A. T., Wifall, T., Hazeltine, E., & Spencer, J. P. (2014). Integrating the behavioral and neural dynamics of reponse selection in a dual-task paradigm: A dynamic neural field model of Dux et al. (2009). *Journal of Cognitive Neuroscience*, *26*(2), 334–351.

Cohen, M. R., & Newsome, W. T. (2008). Context-dependent changes in functional circuitry in visual area MT. *Neuron*, *60*(1), 162–173. https://doi.org/10.1016/j.neuron.2008.08.007

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923.

Constantinidis, C., & Steinmetz, M. A. (1996). Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. *Journal of Neurophysiology*, *76*, 1352–1355.

Constantinidis, C., & Steinmetz, M. A. (2001). Neuronal Responses in Area 7a to Multiple-stimulus Displays: I. Neurons Encode the Location of the Salient Stimulus. *Cerebral Cortex*, *11*(7), 581–591. https://doi.org/10.1093/cercor/11.7.581

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Costello, M. C., & Buss, A. T. (2018). Age-related decline of visual working memory: Behavioral results simulated with a dynamic neural field model. *Journal of Cognitive Neuroscience*.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Daunizeau, J., Stephan, K. E., & Friston, K. J. (2012). Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? *NeuroImage*, *62*(1), 464–481. https://doi.org/10.1016/j.neuroimage.2012.04.061

Deco, G., Rolls, E. T., & Horwitz, B. (2004). "What" and "where" in visual working memory: A computational neurodynamical perspective for integrating FMRI and single-neuron data. *Journal of Cognitive Neuroscience*, *16*(4), 683–701.

Domijan, D. (2011). A computational model of fMRI activity in the intraparietal sulcus that supports visual working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(4), 573–599. https://doi.org/10.3758/s13415-011-0054-x

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*(4), 873–902. https://doi.org/10.1037/a0034247

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*, 1184–1191.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegner, J., & Compte, A. (2009). Mechanism for Top-down control of working memory capacity. *Procedings of the National Academy of Sciences*, *106*(16), 6802–6807.

Edin, F., Macoveanu, J., Olesen, P., Tegnér, J., & Klingberg, T. (2007). Stronger synaptic connectivity as a mechanism behind development of working memory-related brain activity during childhood. *Journal of Cognitive Neuroscience*, *19*(5), 750–760. https://doi.org/10.1162/jocn.2007.19.5.750

Engel, T. A., & Wang, X.-J. (2011). Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *The Journal of Neuroscience*, *31*(19), 6982–6996. https://doi.org/10.1523/JNEUROSCI.6150-10.2011

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., Schöner, G., Erlhange, W., … Schoner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods*, *94*(1), 53–66.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.109.3.545

Faugeras, O., Touboul, J., & Cessac, B. (2009). A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Frontiers in Computational Neuroscience*, *3*, 1. https://doi.org/10.3389/neuro.10.001.2009

Fincham, J. M., Carter, C. S., van Veen, V., Stenger, V. A., & Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(5), 3346–3351. https://doi.org/10.1073/pnas.052703399

Franconeri, S. L., Jonathan, S. V, & Scimeca, J. M. (2010). Tracking multiple objects is limited

only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*(7), 920–925.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.133.3469.2011

Gao, Z., Xu, X., Chen, Z., Yin, J., Shen, M., & Shui, R. (2011). Contralateral delay activity tracks object identity information in visual short term memory. *Brain Research*, *1406*, 30–42. https://doi.org/10.1016/J.BRAINRES.2011.06.049

Gerstner, W., Sprekeler, H., & Deco, G. (2012). Theory and simulation in neuroscience. *Science (New York, N.Y.)*, *338*(6103), 60–65. https://doi.org/10.1126/science.1227356

Grieben, R., Tekülve, J., Zibner, S. K. U., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search. *Attention, Perception, & Psychophysics*, 1–24. https://doi.org/10.3758/s13414-019-01898-y

Grossberg, S. (1982). Biological Competition: Decision Rules, Pattern Formation, and Oscillations. In *Studies of the Mind and Brain* (pp. 379–398). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-7758-7_9

Hock, H. S., Kelso, J. S., & Schöner, G. (1993). Bistability and hysteresis in the organization of apparent motion patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(1), 63–80. https://doi.org/10.1037/0096-1523.19.1.63

Hyun, J., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1140–1160.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schoner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *The Journal of Neuroscience*, *19*(20), 9016–9028.

Jilk, D., Lebiere, C., O'Reilly, R., & Anderson, J. R. (2008). SAL: an explicitly pluralistic cognitive architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 197–218. https://doi.org/10.1080/09528130802319128

Johnson, J. S., Ambrose, J. P., van Lamsweerde, A. E., Dineva, E., & Spencer, J. P. (n.d.). Neural interactions in working memory cause variable precision and similarity-based feature repulsion.

Johnson, J. S., Simmering, V. R., & Buss, A. T. (2014). Beyond slots and resources: grounding cognitive concepts in neural dynamics. *Attention, Perception & Psychophysics*, *76*(6), 1630–1654. https://doi.org/10.3758/s13414-013-0596-9

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*, 568–577.

Johnson, J. S., Spencer, J. P., & Schöner, G. (2009). A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. *Brain Research*, *1299*, 17–32.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, *72*, 210–219. https://doi.org/10.1016/j.jmp.2015.07.002

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kopecz, K., & Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, *73*(1), 49–60.

Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D.-S., … Deisseroth, K. (2010). Global and local fMRI signals driven by neurons defined optogenetically by type and wiring. *Nature*, *465*(7299), 788–792.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511.

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281. https://doi.org/10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Magen, H., Emmanouil, T.-A., McMains, S. A., Kastner, S., & Treisman, A. (2009). Attentional demands predict short-term memory load response in posterior parietal cortex. *Neuropsychologia*, *47*(8–9), 1790–1798. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2009.02.015

Markounikau, V., Igel, C., Grinvald, A., & Jancke, D. (2010). A dynamic neural field model of mesoscopic cortical activity captured with voltage-sensitive dye imaging. *PLoS*

Computational Biology, *6*(9).

Matsumora, T., Koida, K., & Komatsu, H. (2008). Relationship Between Color Discrimination and Neural Responses in the Inferior Temporal Cortex of the Monkey. *Journal of Neurophysiology*, *100*(6), 3361–3374. https://doi.org/10.1152/jn.90551.2008

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *Journal of Neuroscience*, *16*(16). Retrieved from http://www.jneurosci.org/content/16/16/5154.short

Mitchell, D. J., & Cusack, R. (2008). Flexible, Capacity-Limited Activity of Posterior Parietal Cortex in Perceptual as well as Visual Short-Term Memory Tasks. *Cerebral Cortex*, *18*(8), 1788–1798. https://doi.org/10.1093/cercor/bhm205

Moody, S. L., Wise, S. P., di Pellegrino, G., & Zipser, D. (1998). A Model That Accounts for Activity in Primate Frontal Cortex during a Delayed Matching-to-Sample Task. *Journal of Neuroscience*, *18*(1). Retrieved from http://www.jneurosci.org/content/18/1/399.short

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*(5796), 91–94. https://doi.org/10.1126/science.1127242

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Pashler, H. (1988). Familiarity and the detection of change in visual displays. *Perception and*

Psychophysics, *44*, 369–378.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157–1172. https://doi.org/10.1016/j.neuroimage.2004.03.026

Perone, S., Molitor, S. J., Buss, A. T., Spencer, J. P., & Samuelson, L. K. (2015). Enhancing the Executive Functions of 3-Year-Olds in the Dimensional Change Card Sort Task. *Child Development*, *86*(3). https://doi.org/10.1111/cdev.12330

Perone, S., Simmering, V. R., & Spencer, J. P. (2011). Stronger neural dynamics capture changes in infants' visual working memory capacity over development. *Developmental Science*, *14*(6), 1379–1392.

Pessoa, L., Gutierrez, E., Bandettini, P. A., & Ungerleider, L. G. (2002). Neural Correlates of Visual Working Memory. *Neuron*, *35*(5), 975–987. https://doi.org/10.1016/S0896-6273(02)00817-6

Pessoa, L., & Ungerleider, L. (2004). Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex*, *14*, 511–520.

Qin, Y., Sohn, M.-H., Anderson, J. R., Stenger, V. A., Fissell, K., Goode, A., & Carter, C. S. (2003). Predicting the Practice Effects on the Blood Oxygenation Level-Dependent (BOLD) Function of fMRI in a Symbolic Manipulation Task. *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences. https://doi.org/10.2307/3144047

Raffone, A., & Wolters, G. (2001). A Cortical Mechanism for Binding in Visual Working Memory. *Journal of Cognitive Neuroscience*, *13*(6), 766–785. https://doi.org/10.1162/08989290152541430

Rigoux, L., & Daunizeau, J. (2015). Dynamic causal modelling of brain-behaviour relationships. *NeuroImage*, *117*, 202–221. https://doi.org/10.1016/j.neuroimage.2015.05.041

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies — Revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/J.NEUROIMAGE.2013.08.065

Roberts, S. J., & Penny, W. D. (2002). Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, *50*(9), 2245–2257. https://doi.org/10.1109/TSP.2002.801921

Robitaille, N., Grimault, S., & Jolicœur, P. (2009). Bilateral parietal and contralateral responses during maintenance of unilaterally encoded objects in visual short-term memory: Evidence from magnetoencephalography. *Psychophysiology*, *46*(5), 1090–1099. https://doi.org/10.1111/j.1469-8986.2009.00837.x

Rosa, M. J., Friston, K., & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, *208*(1), 66–78. https://doi.org/10.1016/J.JNEUMETH.2012.04.013

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316). Retrieved from http://science.sciencemag.org/content/354/6316/1136

Ross-Sheehy, S., Schneegans, S., & Spencer, J. P. (2015). The Infant Orienting With Attention Task: Assessing the Neural Basis of Spatial Attention in Infancy. *Infancy*, *20*(5), 467–506. https://doi.org/10.1111/infa.12087

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008a). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979. https://doi.org/10.1073/pnas.0711295105

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008b). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979. https://doi.org/10.1073/pnas.0711295105

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, *18*(2), 324–330. https://doi.org/10.3758/s13423-011-0055-3

Schneegans, S. (2016). Sensori-motor transformations. In G. Schoner & J. P. Spencer (Eds.), *Dynamic Thinking--A Primer on Dynamic Field Theory*.

Schneegans, S., & Bays, P. M. (2017). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience*, *29*(12), 1977–1994. https://doi.org/10.1162/jocn_a_01180

Schneegans, S., Spencer, J. P., & Schöner, G. (2016). Integrating "What" and "Where": Visual Working Memory for Objects in a Scene. In G. Schöner, J. P. Spencer, & T. D. R. Group (Eds.), *Dynamic Thinking--A Primer on Dynamic Field Theory* (pp. 197–226). New York: Oxford University Press.

Schneegans, S., Spencer, J. P., Schoner, G., Hwang, S., & Hollingworth, A. (2014). Dynamic interactions between visual working memory and saccade target selection. *Journal of*

Vision, *14*(11), 9–9. https://doi.org/10.1167/14.11.9

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*(2), 273–299. https://doi.org/10.1037/0033-295X.113.2.273

Schoner, Gregor, Spencer, J. P., & Group, T. D. R. (2016). *Dynamic Thinking: A primer on Dynamic Field Theory.* New York: Oxford University Press.

Schutte, A. R., & Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: capturing a qualitative developmental transition in spatial working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1698–1725. https://doi.org/10.1037/a0015794

Schutte, A. R., Spencer, J. P., & Schoner, G. (2003). Testing the Dynamic Field Theory: Working Memory for Locations Becomes More Spatially Precise Over Development. *Child Development*, *74*(5), 1393–1417.

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2016). Object selection costs in visual working memory: A diffusion model analysis of the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(11), 1673–1693. https://doi.org/10.1037/a0040213

Sheremata, S. L., Bettencourt, K. C., & Somers, D. C. (2010). Hemispheric asymmetry in visuotopic posterior parietal cortex emerges with visual short-term memory load. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(38), 12581–12588. https://doi.org/10.1523/JNEUROSCI.2689-10.2010

Simmering, V. R. (2016). Working memory in context: Modeling dynamic processes of behavior, memory, and development. *Monographs of the Society for Research in Child*

Development, *81*(3), 1–166. https://doi.org/10.1111/mono.12249

Simmering, V. R., & Spencer, J. P. (2008). Generality with specificity: the dynamic field theory generalizes across tasks and time scales. *Developmental Science*, *11*(4), 541–555. https://doi.org/10.1111/j.1467-7687.2008.00700.x

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830. https://doi.org/10.1037/a0029856

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., … Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Spencer, Barich, K., Goldberg, J., & Perone, S. (2012). Behavioral dynamics and neural grounding of a dynamic field theory of multi-object tracking. *Journal of Integrative Neuroscience*, *11*(3), 339–362.

Spencer, J. P., Perone, S., & Johnson, J. S. (2009). The Dynamic Field Theory and Embodied Cognitive Dynamics. In J. P. Spencer, M. S. Thomas, & J. L. McClelland (Eds.), *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered* (pp. 86–118). New York, NY: Oxford University Press.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

https://doi.org/10.1016/J.NEUROIMAGE.2009.03.025

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2136–2157. https://doi.org/10.3758/s13414-014-0633-3

Szczepanski, S. M., Pinsk, M. A., Douglas, M. M., Kastner, S., & Saalmann, Y. B. (2013). Functional and structural architecture of the human dorsal frontoparietal attention network. *Proceedings of the National Academy of Sciences*, *110*(39), 15806–15811. https://doi.org/10.1073/pnas.1313903110

Tegnér, J., Compte, A., & Wang, X.-J. (2002). The dynamical stability of reverberatory neural circuits. *Biological Cybernetics*, *87*(5–6), 471–481. https://doi.org/10.1007/s00422-002-0363-9

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. *The Behavioral and Brain Sciences*, *24*(1), 1–34; discussion 34-86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11515285

Todd, J. J., Fougnie, D., & Marois, R. (2005). Visual Short-term memory load suppresses temporo-pariety junction activity and induces inattentional blindness. *Psychological Science*, *16*(12), 965–972.

Todd, J. J., Han, S. W., Harrison, S., & Marois, R. (2011). The neural correlates of visual working memory encoding: A time-resolved fMRI study. *Neuropsychologia*, *49*(6), 1527–1536. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2011.01.040

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *166*(2003), 751–754.

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(2), 144–155.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2016). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2016.01.001

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*(2), 197–214. https://doi.org/10.1037/rev0000060

Varoquaux, G., & Craddock, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, *80*, 405–415. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.007

Veksler, B. Z., Boyd, R., Myers, C. W., Gunzelmann, G., Neth, H., & Gray, W. D. (2017). Visual Working Memory Resources Are Best Characterized as Dynamic, Quantifiable Mnemonic Traces. *Topics in Cognitive Science*, *9*(1), 83–101. https://doi.org/10.1111/tops.12248

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of Color Stimuli in Awake Macaque Primary Visual Cortex. *Neuron*, *37*(4), 681–691. https://doi.org/10.1016/S0896-6273(03)00035-7

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *The Journal of Neuroscience*, *32*(33), 11228–11240. https://doi.org/10.1523/JNEUROSCI.0735-12.2012

Wijeakumar, S., Ambrose, J. P., Spencer, J. P., & Curtu, R. (2016). Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2016.11.002

Wijeakumar, S., Magnotta, V. A., & Spencer, J. P. (2017). Modulating perceptual complexity and load reveals degradation of the visual working memory network in ageing. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.06.019

Wijeakumar, S., Spencer, J. P., Bohache, K., Boas, D. A., & Magnotta, V. A. (2015). Validating a new methodology for optical probe design and image registration in fNIRS studies. *NeuroImage*, *106*, 86–100. https://doi.org/10.1016/j.neuroimage.2014.11.022

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 1120–1135.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*(1), 1–24.

Xiao, Y., Wang, Y., & Felleman, D. J. (2003). A spatially organized representation of colour in macaque cortical area V2. *Nature*, *421*(6922), 535–539.

Xu, Y. (2007). The role of the superior intraparietal sulcus in supporting visual short-term memory for multifeature objects. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(43), 11676–11686. https://doi.org/10.1523/JNEUROSCI.3545-07.2007

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.

Acknowledgements This work was supported by National Science Foundation BCS-1029082 awarded to JPS. We thank Rodica Curtu, Eliot Hazeltine, and Larissa Samuelson for helpful comments on this work.

Author Contributions J.P.S. and A.T.B. designed the experiment. G.S. and J.P.S. provided input on hemodynamic and behavioral simulations. T.H. provided input on the statistical methods to compare GLMs. W.P. developed the statistical methods for comparing GLMs. A.T.B. ran the simulations and analyzed the behavioral data. A.T.B. and V.M. analyzed the fMRI data. A.T.B. prepared the figures. A.T.B. and J.P.S. wrote the paper. J.P.S. supervised all aspects of the work.

Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.T.B. (abuss@utk.edu) or J.P.S. (j.spencer@uea.ac.uk).

Tables

Table 1: Summary of variations in the CD task that have been simulated by the DF model.

	N Subjects	SSs	Array 1 Duration	Delay Duration	Test Array Type
Johnson et al. (2009a); Experiment 1a	10	3	500 ms	1,000 ms	Single-item
Todd & Marois (2004); Experiment 1	17	1-4, 6, 8	100 ms	1,200 ms	Single-item
Magen et al (2009); Experiment 3	12	1, 3, 5, 7	500 ms	6,000 ms	Single-item
Costello & Buss (2017); Experiment 1	26	1, 3, 5	500 ms	1,200 ms	Whole-array
Current report	16	2, 4, 6	500 ms	1,200 ms	Whole-array

Table 2: Log Bayes factors across models for all subjects.

Participant	Null	Set size	Accuracy	Change	SS*Acc	SS*Ch	SS*Ch*Acc	DF
1	813.1	447.9	402.3	388.2	526.7	530.7	464.7	**638***
2	986.2	421.9	357.7	348.2	514.4	510.3	410.7	**635.9***
3	1002	158.1	67.1	76.3	267.7	271.4	120.9	**454.6***
4	583.7	18.5	**-47.8***	-42	103.2	115.1	19.8	245.6
5	529	167.2	94.3	127.8	214.3	252.3	135.1	**302.1***
6	604.8	180.7	102.8	122.9	251.6	293.5	177.1	**414.5***
7	844.8	-39.3	-91	**-106.7***	91.5	63.3	-34	251.5
8	230.9	-80.8	-131.8	**-141.5***	-9.7	-6.4	-90.5	88.7
9	460.6	-15.1	**-86***	-79.4	56.6	69.5	-42.2	170.8
10	286.1	284.9	278.9	281.2	285.7	286.8	281	**286.5***
11	1381	457	383.2	407.9	580.7	603.3	487.5	**804.8***
12	1052	298.4	243.9	260.1	400.1	419	333.7	**596.9***
13	261.2	115.1	58.4	58.5	157.7	178.9	102.9	**202***
14	1207	317	255.6	286.2	471.2	496.1	384.4	**755.8***
15±	420.9	54.6	**-12.1***	-1.4	123.9	128.2	39.8	231.7
16±	691.1	68.5	**-19.6***	21	177	215	77.2	392.7

Preferred model is indicated by *. Negative values indicate cases in which a categorical model outperformed the DF model. The reduced DF model with three components was preferred by participants denoted with ±.

Figures

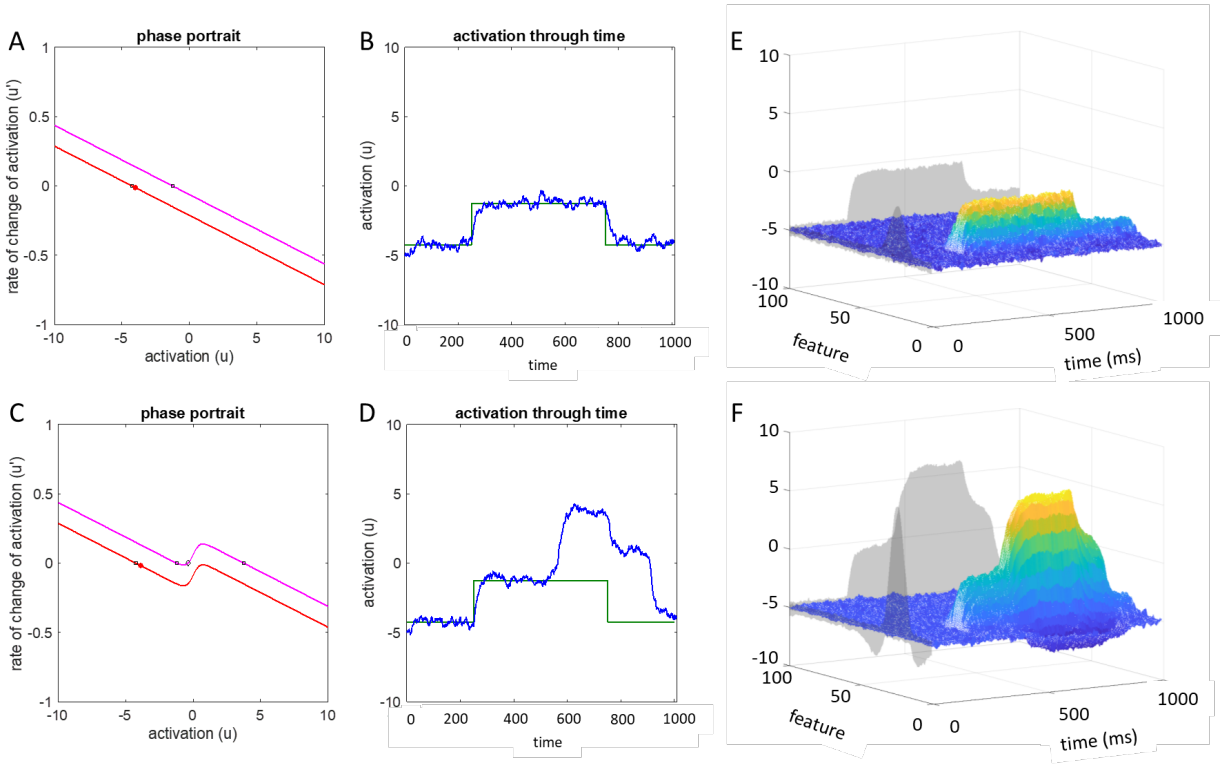

Figure 1 | **Illustration of activation dynamics. A-B,** the phase-space and activation over time of a neuron with linear dynamics. The purple line in panel A corresponds to the period of time in panel B during which activation is boosted by an input, the red line in panel A corresponds to the other time points. **C-D,** the phase-space and activation over time of a neuron with non-linear dynamics created through the addition of self-excitation (note the curves in phase-space around the activation value of 0). When the neuron is boosted by an input in panel D, self-excitation creates a non-linearity which pulls activation fluctuations push activation back below 0 and self-exitation is disengaged. **E-F,** corresponding activation profiles for these two different systems in a field of interactive neurons. Note the correspondence in profiles between B-E and D-F.

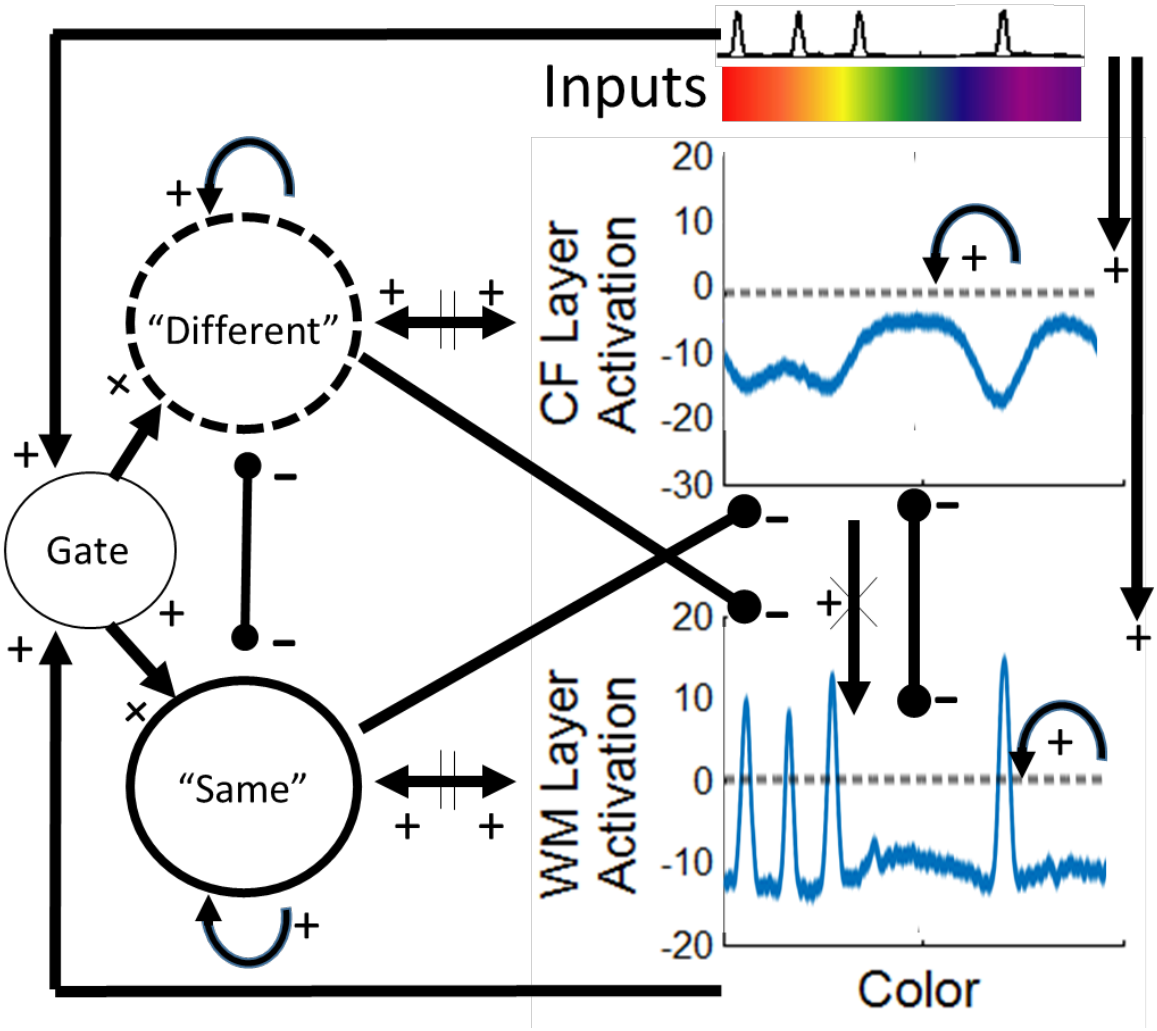

Figure 2 | **Model architecture.** Excitatory connections are indicated by lines with pointed end and inhibitory connections are illustrated with lines with balled end. Connections with parallel lines (i.e., between "Different" and CF and between "Same" and WM) are engaged when the Gate node is activated. Connections with perpendicular lines (i.e., from CF to WM) are turned off when the Gate node is activated.

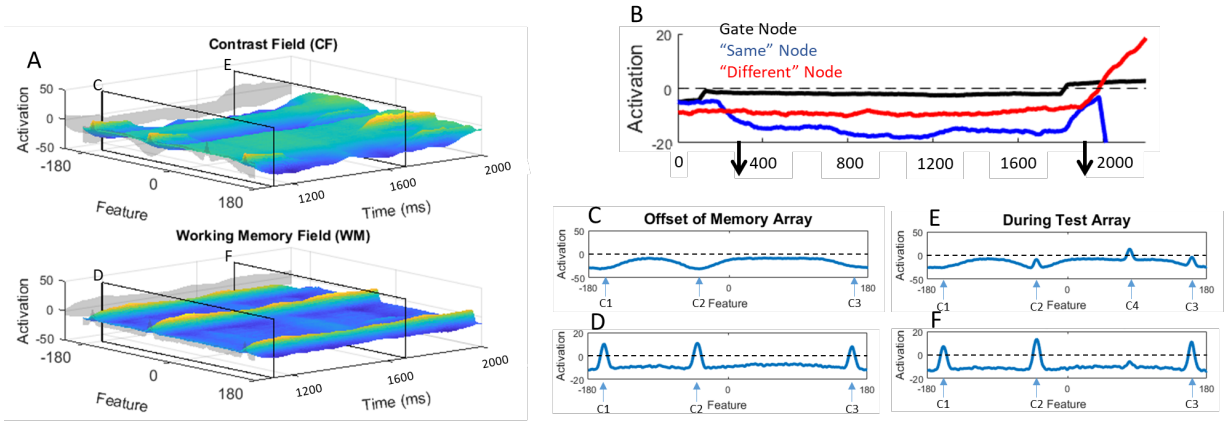

Figure 3 | **Model dynamics. A,** activation of the model architecture on a set-size 3 trial. **B,** activation of the decision nodes over the course the trial. **C-D,** time-slices from CF and WM at the offset of the memory array (note the corresponding boxes in Panel A). **E-F,** time-slices from CF and WM during the presentation of the test array (note the corresponding boxes in Panel A). In this trial, a different color value is presented during the test array (note the above-threshold activation in Panel E) and the model responds "different" (note the activation profile of the decision nodes in panel B).

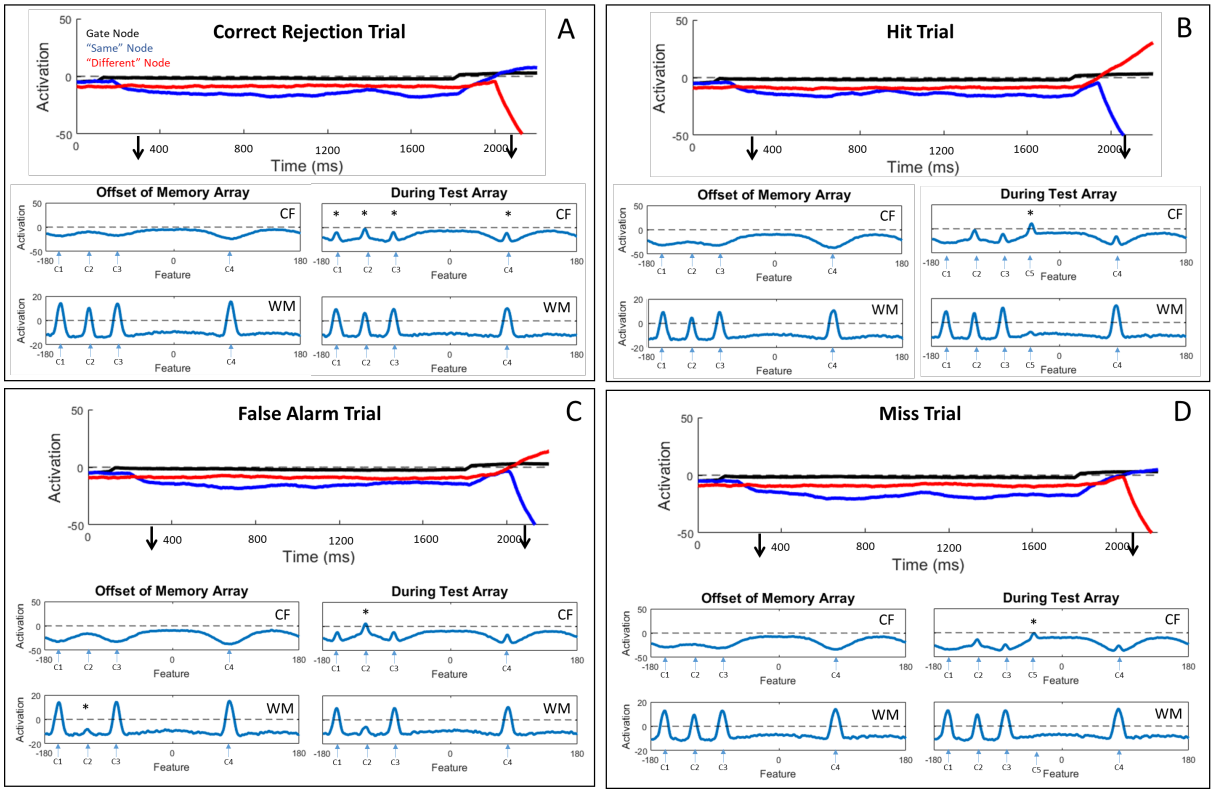

Figure 4 | **Model performing different trial types. A,** the model correctly performing a "same" trial. At the offset of the memory array, the WM field has built peaks corresponding to the four items in the memory array. During test when the same item are presented, activation in CF stays below threshold (note the asterisks above CF). Here, the model responds "same" (note the activation of the decision nodes). **B**, the model correctly performing a "different" trial. Now, during the test array, a new item is presented which goes above threshold in CF (note the asterisk above CF). **C**, the model performing a "same" trial but generating an incorrect response. At the offset of the memory array, the WM field has failed to consolidate one of the items into memory (note the asterisk above WM). Subsequently, during the presentation of the same items during the test array, the corresponding stimulus goes above threshold in CF (note the asterisk above CF) and the model generates a "different" response. **D**, the model performing a "different" trial but generating an incorrect response. In this example, the model has overly robust activation with

the WM field which leads to stronger inhibition within CF and a failure of the new item to go above threshold in CF (note the asterisk above CF).

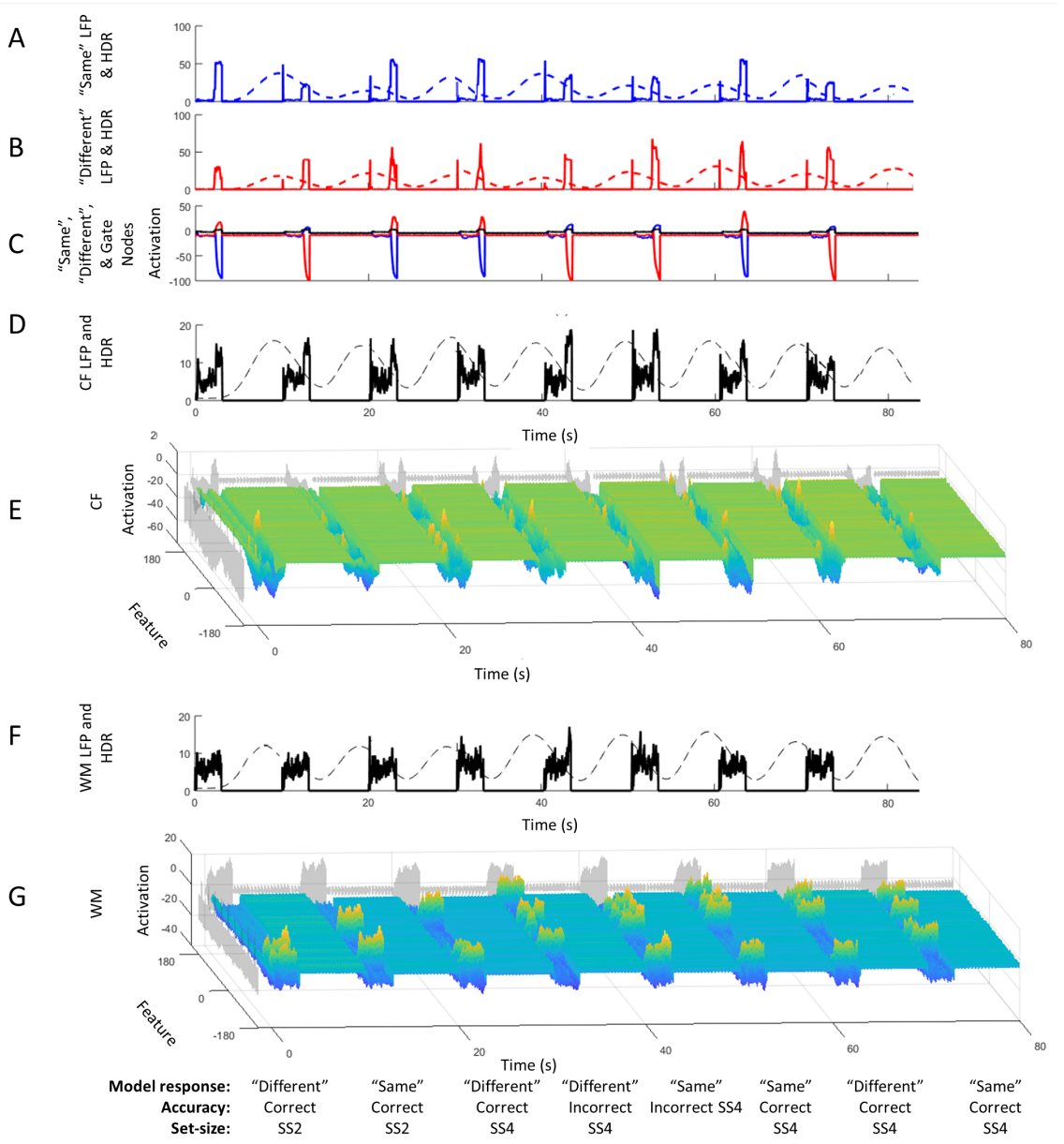

Figure 5 | Illustration model activation dynamics and hemodynamics. A, B, D, and F, stimulated local field potential (solid lines) and corresponding hemodynamic responses (dashed lines) from the "same" node (A), "different" node (B), CF (D) and WM (F). **C, E, and G,** activation of model components over a series of 8 trials (note the labels at the bottom which categorize each trial type) for the decision nodes (C), CF (E), and WM (G).

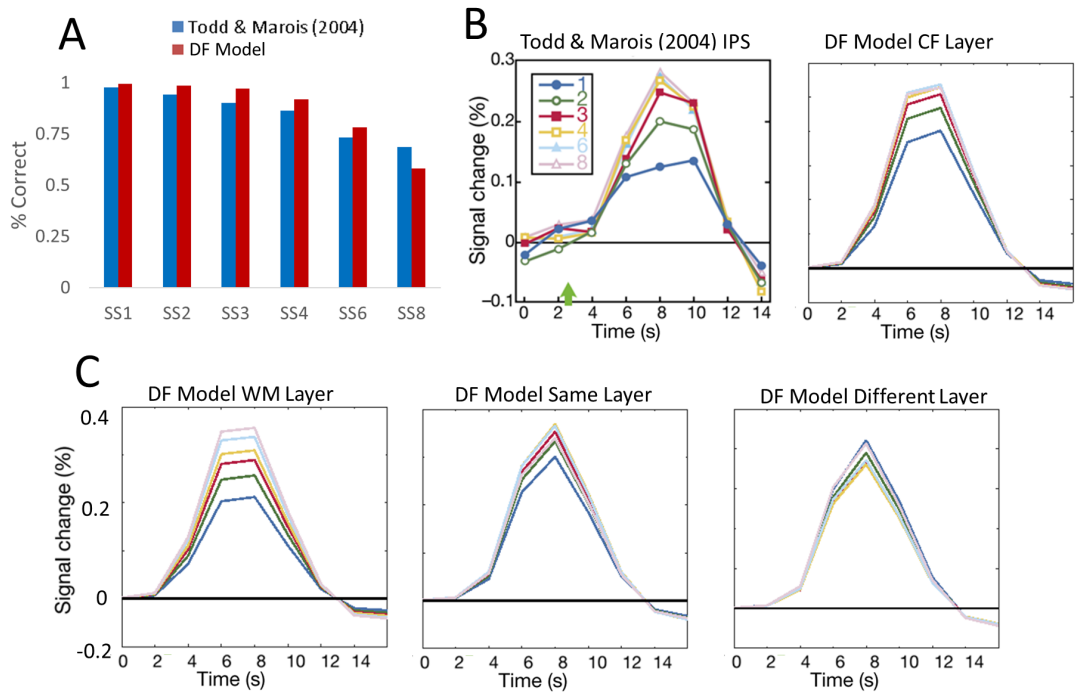

Figure 6 | Simulations of Todd & Marois (2004). A, Behavioral performance and model simulations. **B,** BOLD response from IPS across memory loads of 1, 2, 3, 4, 6, and 8 items (left) and simulated hemodynamic response from CF layer (right). **C,** Simulated hemodynamic response from the other 3 components of the model.

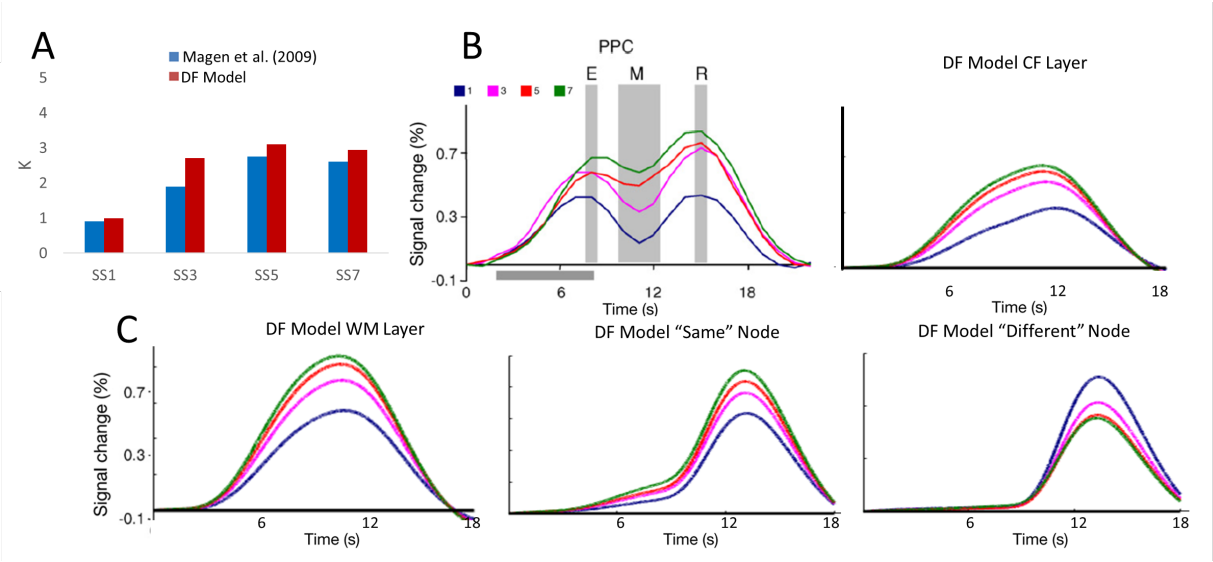

Figure 7 | Simulations of Magen et al. (2009). A, Behavioral performance and model simulations. **B,** BOLD response from PPC across memory loads of 1, 3, 5, and 7 items (left) and simulated hemodynamic response from CF layer (right). **C,** Simulated hemodynamic response from the other 3 components of the model.

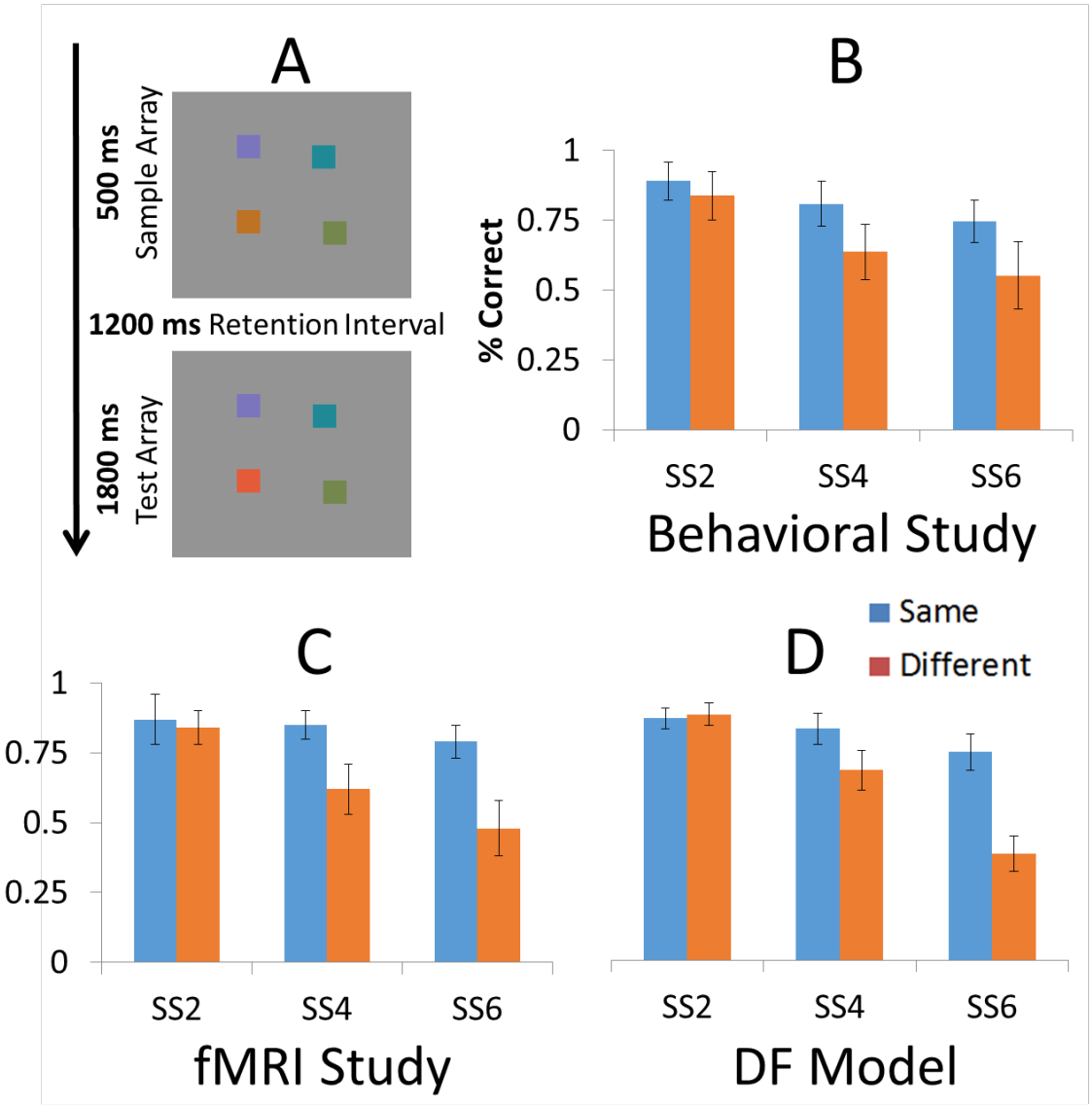

Figure 8 | Task design and behavioral / simulation data. A, A trial began with a sample array consisting of 2, 4, or 6 colored items. Next came a retention interval and presentation of a test array. On change trials (50% of trials), one randomly-selected item was shifted 36° in color space. **B,** Percent correct from behavioral study. **C,** Percent correct from fMRI study. In both studies, there were many errors at set-size four, but performance was above chance ($t(27)=23.5$, $p<0.001$). **D,** Simulations reproduced the behavioral pattern. Error bars show ± 1 *SD*.

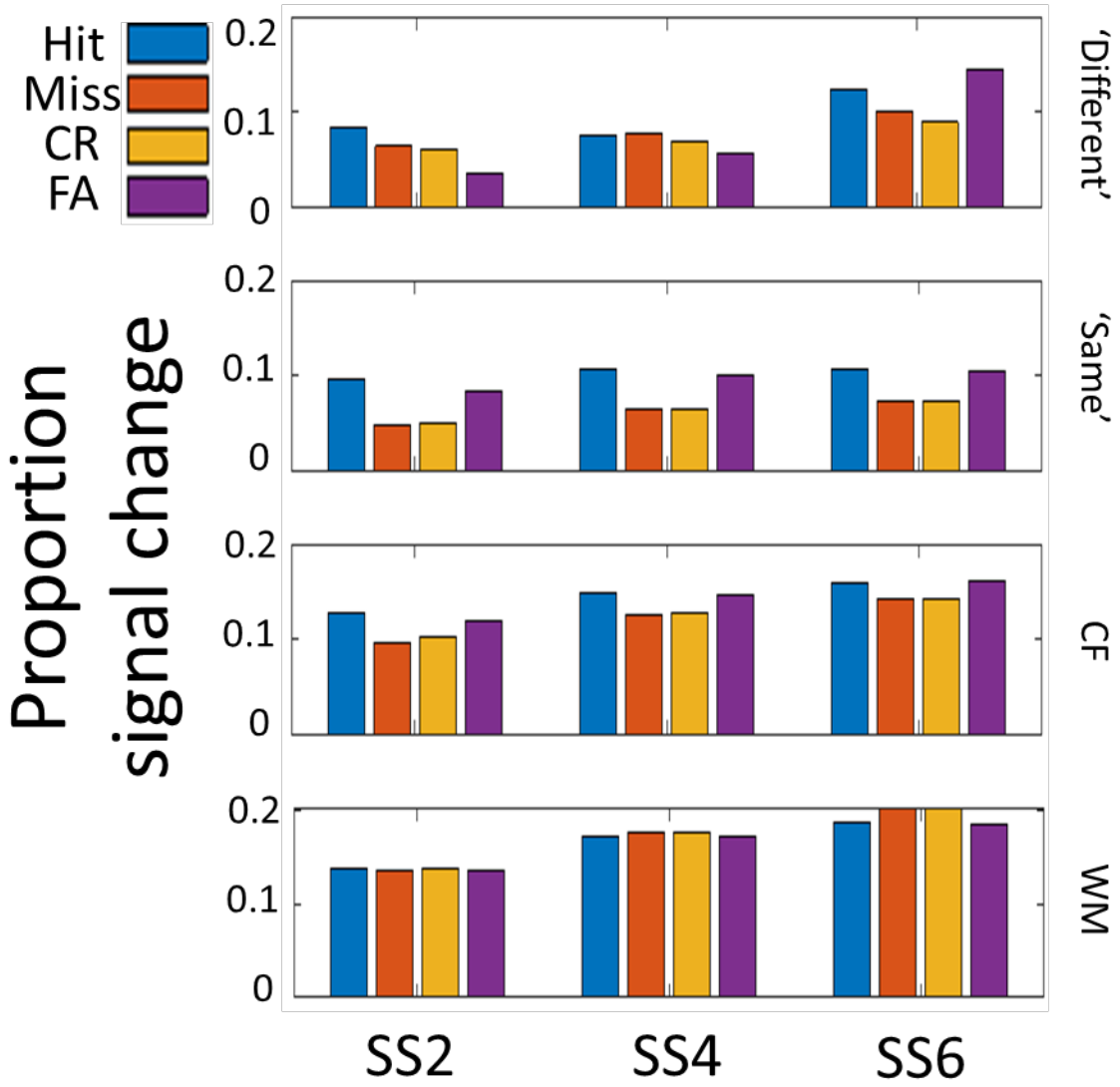

Figure 9 | Average amplitude of hemodynamic response across model components and trial types. This figure shows the variations in the amplitude of the hemodynamic response when perofrming our version of the change detection task (correct change trial = hit; correct same trial = correct-rejection (CR); incorrect change trial = Miss; incorrect same trial = false alarm (FA)).

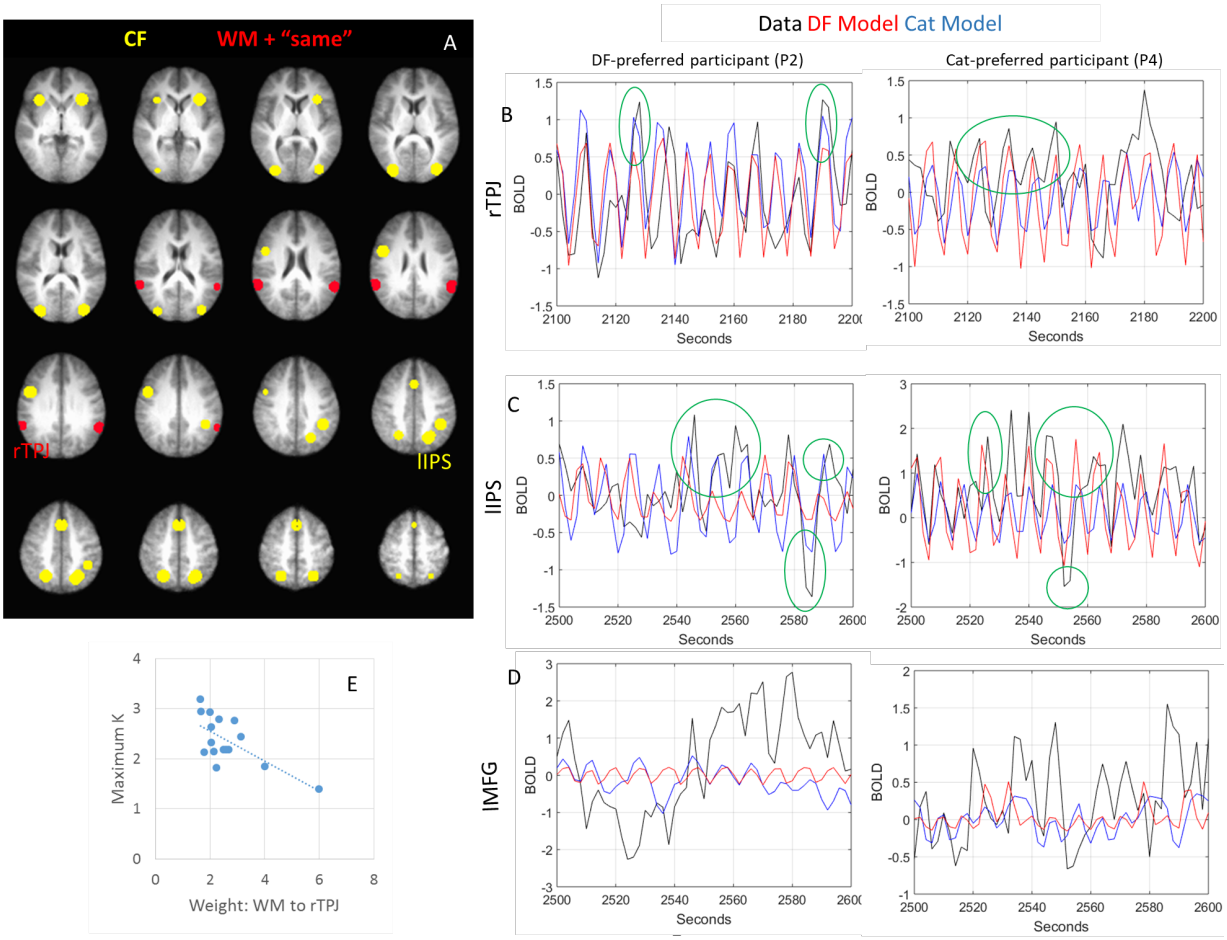

Figure 10 | **Mapping of model components to ROIs. A,** Yellow spheres show ROIs which corresponded to CF and red spheres show ROIs which corresponded to WM+"same". **B-C,** Time-course plots showing the BOLD response and predicted time-courses from the DF model and from the accuracy categorical model within regions that were mapped by DF components. A participant is shown that preferred the DF model (P1) and a participant that preferred the accuracy categorical model (P8). **D,** The same time-courses and participants are shown within a region that was not mapped by a DF component. **E,** Scatter plot showing the correlation between participant-specific weights of the WM component from the DF model to rTPJ activation and individual capacity.

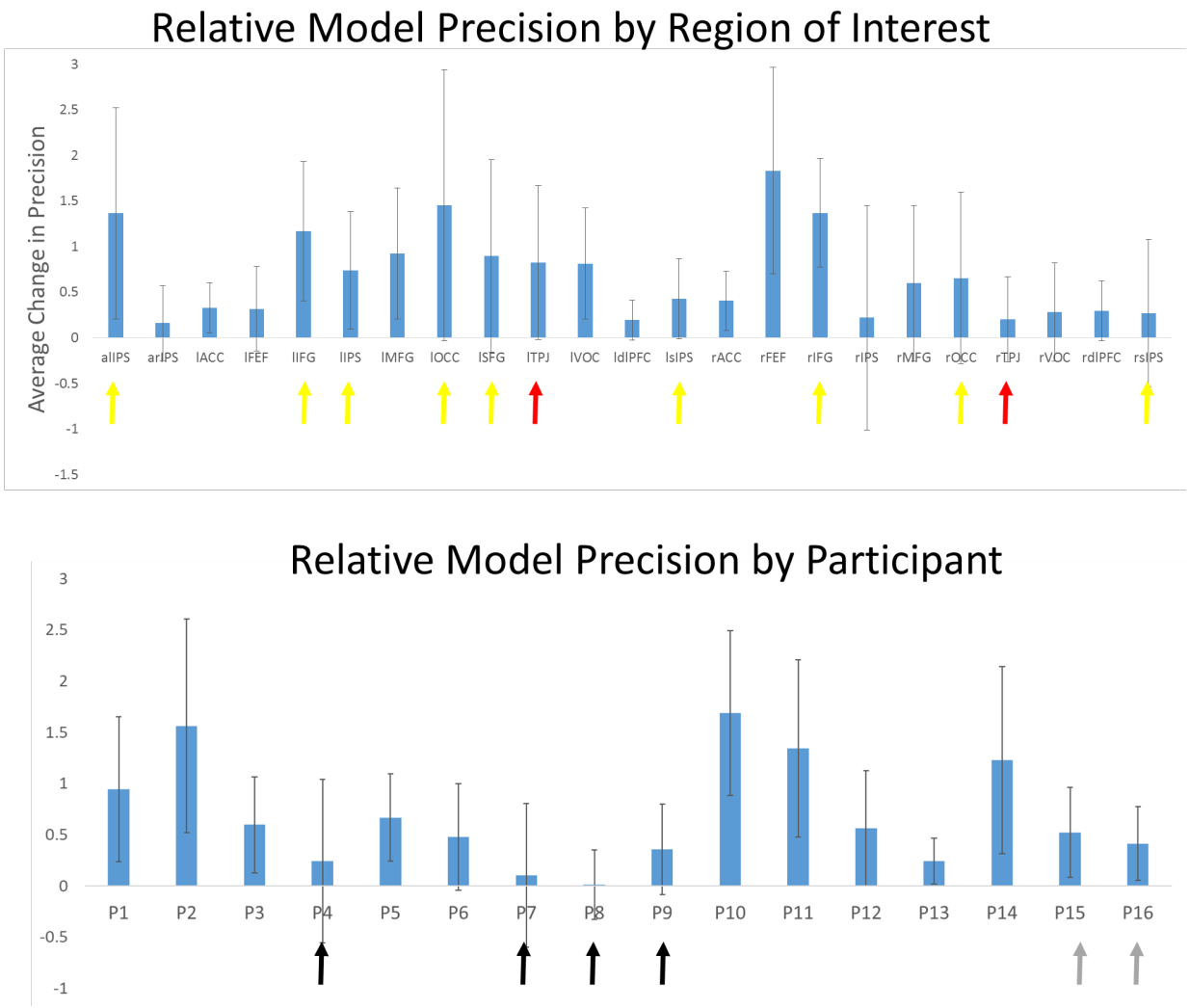

Figure 11 | Relative model precision. Average improvement in model precision for the DF model relative to the array of categorical models. Top panel shows relative improvement in model precision within the 23 ROIs. Yellow (CF) and red (WM+'same') arrows mark regions that were mapped to components of the DF model. Bottom panel shows relative improvement in model precision by participant. Arrows indicate participants that preferred a categorical model over the DF model with four components. Grey arrows indicate participants that switched to prefer the DF model when only three components of the DF model were included.

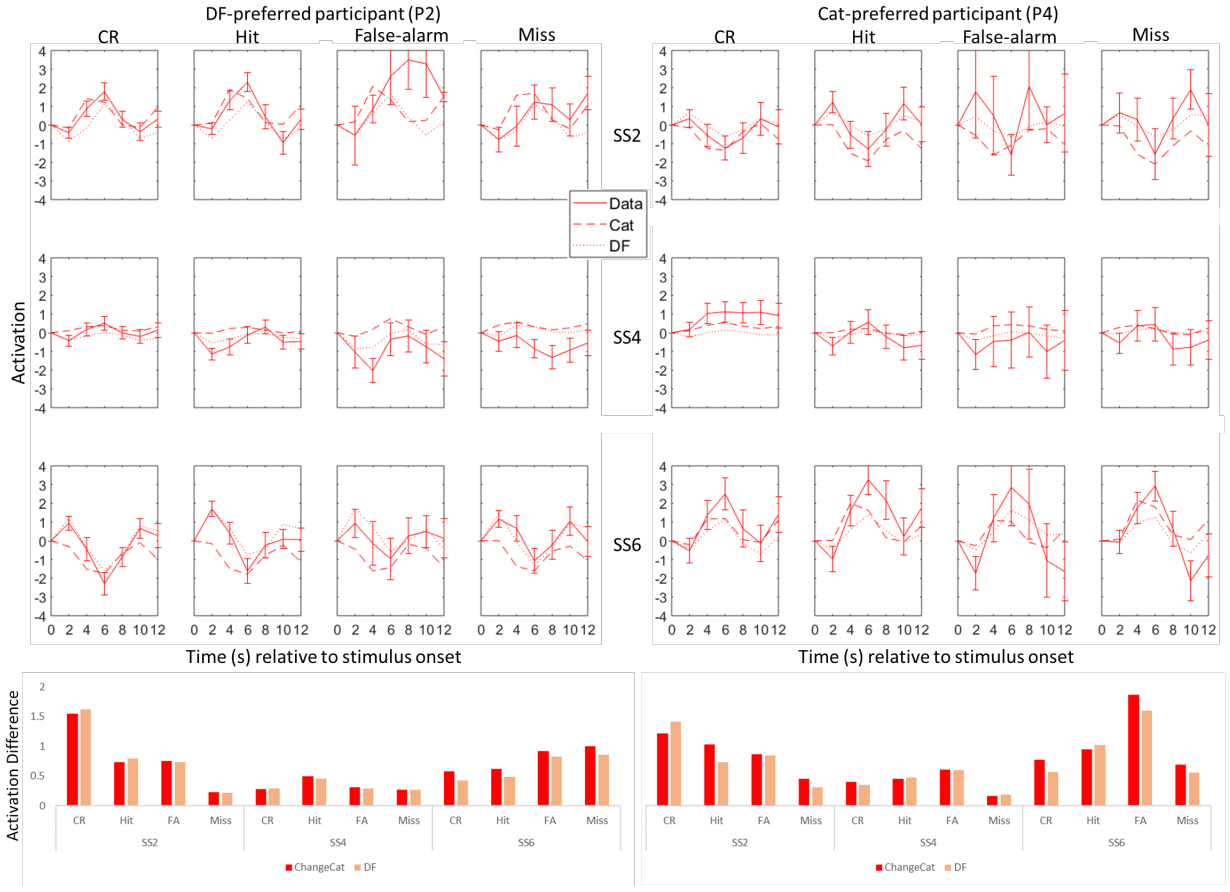

Figure 12 | Activation and model prediction across trial-types within lIPS. Activation (solid) and model predictions for the DF (dotted) and change categorical (dashed) models is plotted across trial-types and different set sizes. Left graphs represent activation for a participant that preferred the DF model. Right graphs represent activation for a participant that preferred the change categorical model. The bar graphs show the average absolute difference between activation and model predictions within the 10 second time window.

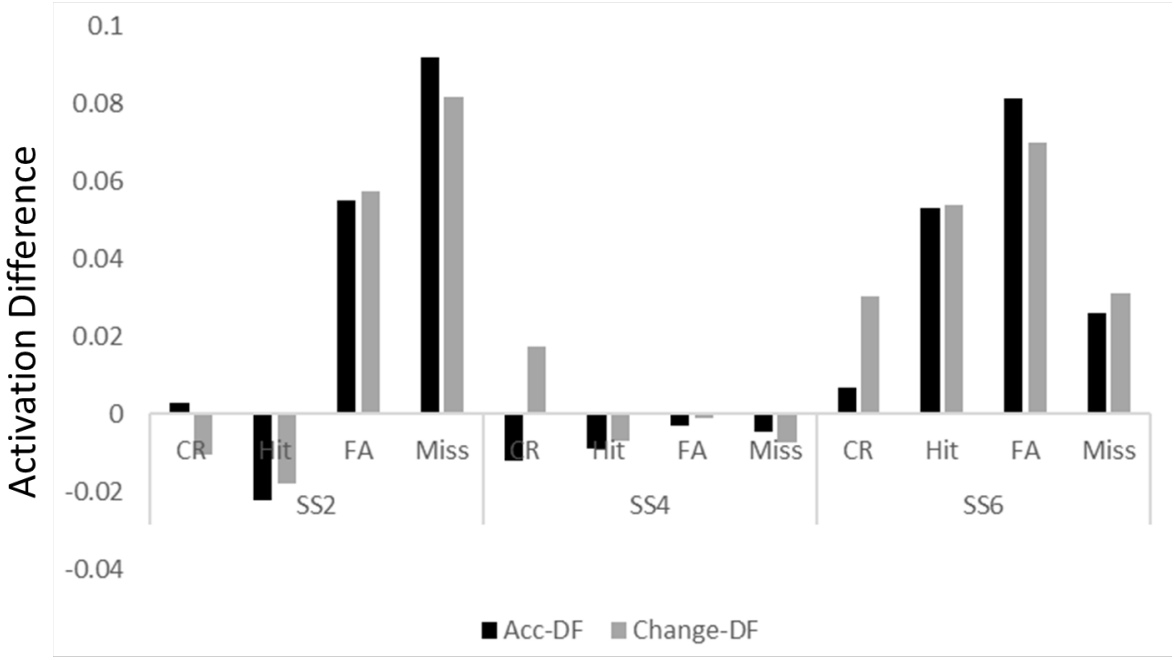

Figure 13 | Relative differences between activation in lIPS and model predictions by trial-type. We first calculated the absolute average difference between activation and model predictions for the DF model, accuracy categorical model, and change categorical model within a 10 second window for each trial type (as visualized in Figure 12). Next, the difference for the DF model was subtracted from the difference of each categorical model. Positive values, then, reflect instances where the categorical model deviated from observed activation more so than the DF model. Negative values indicate instances in which the DF model deviated from observed activation more so than the categorical model.