Articles in PresS. J Neurophysiol (July 15, 2015). doi:10.1152/jn.00362.2015

Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study

Abbreviated title: Neural circuit basis of working memory precision

Rita Almeida[1,2], João Barbosa[1], Albert Compte[1]

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

[2]Department of Neuroscience, Karolinska Institute, Stockholm, Sweden

Corresponding author:

Albert Compte
Institut d'Investigacions Biomèdiques August Pi I Sunyer (IDIBAPS)
C/ Rosselló 149
08036 Barcelona
Spain
Fax: +34 93 3129408
Tel: +34 93 2275400 ext. 4151
E-mail: acompte@clinic.ub.es

Abstract

The amount of information that can be retained in working-memory (WM) is limited. Limitations of WM capacity have been the subject of intense research, especially trying to specify algorithmic models for WM. Comparatively, neural circuit perspectives have barely been used to test WM limitations in behavioral experiments. Here, we used a neuronal microcircuit model for visuo-spatial WM (vsWM) to investigate memory of several items. The model assumes that there is a topographic organization of the circuit responsible for spatial memory retention. This assumption leads to specific predictions, which we tested in behavioral experiments. According to the model, nearby locations should be recalled with a bias, as if the two memory traces showed attraction or repulsion during the delay period depending on distance. Another prediction is that the previously reported loss of memory precision for increasing number of memory items (memory load) should vanish when the distances between items are controlled for. Both predictions were confirmed experimentally. Taken together, our findings provide support for a topographic neural-circuit organization of vsWM, they suggest that interference between similar memories underlies some WM limitations, and they put forward a circuit-based explanation that reconciles previous conflicting results on the dependence of WM precision with load.

Keywords: short-term memory, working-memory, precision, capacity, attractor model

Working-memory (WM) refers to the ability of actively retaining stimulus information over a short period of time and it is thought to be a core component of cognitive functions (Baddeley 1986, Conway et al. 2003). A hallmark of WM is that the information retained is limited. Currently, a significant effort is being devoted to characterizing the nature of WM capacity limitations, but their bases remain controversial (Luck and Vogel 2013, Ma et al. 2014). Important points of discordance have been whether or not the number of items in WM can be increased at a cost in precision (Bays and Husain 2008, Zhang and Luck 2008) and whether the similarity of the items to memorize improves (Johnson et al. 2009, Lin and Luck 2009) or degrades (Elmore et al. 2011) WM performance.

Recently, a neuronal circuit perspective is entering these debates: electrophysiological experiments have started to investigate the neural basis of multiple item WM (Buschman et al. 2011, Warden and Miller 2007, Lara and Wallis 2014), and neural-circuit modeling has been used to link cellular and network mechanisms with behavior to understand WM capacity limitations (Macoveanu et al. 2006, 2007, Edin et al. 2009, Wei et al. 2012, Wimmer et al. 2014, Bays 2014, Papadimitriou et al. 2015). Most of these models are variations of a model (Compte et al. 2000) developed to be consistent with neurophysiological data from behaving monkeys (Funahashi et al. 1989). They rely on the assumption that there is a topographic structure in the circuits supporting vsWM, which implements a continuous attractor mechanism responsible for the retention of spatial memory. Some evidence from fMRI (Schluppeck et al. 2006, Kastner et al. 2007) and electrophysiology studies (Constantinidis et al. 2001, Inoue and Funahashi 2002) supports a coarse degree of spatial WM maps in parietal and prefrontal cortex. Recently, neural evidence for attractor dynamics on a fine vsWM spatial map in prefrontal cortex has also been found (Wimmer et al. 2014). However, additional implications of such a spatial memory map for the relation between vsWM precision, capacity and stimulus similarity remain untested. We aimed here to advance our understanding of the neuronal underpinnings of vsWM by explicitly testing the assumption of a topographic structure of the vsWM buffer. One implication of this structure is that the efficiency with which different items are memorized should depend on their relative locations, since stronger interference of memory traces would be expected for nearby items. Using simulations we predicted an attractive bias when remembering locations of two nearby items, for very short inter-item distances. This prediction was validated in behavioral experiments in humans. We then sought to address how these interferences affected the relationship between memory load and precision. In our model, the effect of load on memory precision was largely accounted for by changes in inter-item distance with load. Behavioral data confirmed this prediction. We finally tested in an additional experiment that behavioral data was better explained by memory attraction than by memory swapping (Bays et al. 2009), and we also confirmed that intermediate distances between memorized items were characterized by a repulsive memory bias. The importance of our work is three-folded. First, we provide new experimental evidence concerning interference in vsWM. Second, we test a critical assumption of an important class of models of vsWM. Third, we put forward a plausible explanation reconciling previous results concerning the dependence of memory precision on load and concerning similarity effects on performance.

Materials and Methods

Model

We used a previously proposed computational model (Compte et al. 2000, Edin et al. 2009) to study the precision of vsWM of multiple items. The model (Compte et al. 2000) was originally developed to account for a candidate neuronal mechanism for vsWM, namely the selective sustained elevated neuronal firing of the prefrontal cortical neurons of monkeys performing a vsWM task (Funahashi et al. 1989). The model consists of a network of interconnected excitatory and inhibitory spiking neurons. The neurons encode the spatial location of fixed-eccentricity visual stimuli in angle θ. That is, they encode positions (in angle) on a circle. Presentation of a stimulus at location θ is simulated by increasing the external input to the corresponding excitatory neurons. The selective response of the neurons in the network is maintained due to the structured connectivity of the network. Excitatory neurons encoding for nearby angles have stronger than average connections, which is essential for a selective group of neurons to sustain elevated spiking after stimulus cessation (Compte et al. 2000)

The parameter values used were as in the IPS circuit described in Edin et al. 2009, for a network capacity of 2 items. The model had 1024 excitatory and 256 inhibitory leaky integrate-and-fire neurons (Tuckwell 1988). The neuronal selectivity was imposed by external inputs, assumed to originate in upstream areas of the dorsal pathway. Specifically, the presence of a visual stimulus at an angle θ_{stim} was modeled by increasing the external input to excitatory neurons with preferred direction around θ_{stim}. The strength of the external input to a neuron encoding θ decayed with the distance to θ_{stim} according to $I_{stim}(\theta,\theta_{stim}) = \alpha \exp\left(\mu\left[\cos\left(2\pi/360(\theta-\theta_{stim})\right)-1\right]\right)$, where $\alpha = 0.025$ nA and $\mu = 39$.

The integrate-and-fire neuron model describes how the membrane voltage V_m integrates incoming inputs until a certain threshold value V_{th} is reached and an action potential or spike is fired. After reaching the threshold, V_m is reset to V_{res} for a refractory time period τ_{ref} before continuing to integrate inputs. The equation describing the sub-threshold changes in V_m is:

$$C_m \frac{dV_m}{dt} = -g_L(V_m - E_L) - I_{syn} - I_{ext}$$

Each cell is then characterized by the total membrane capacitance C_m , the total leak conductance g_L , the leak reversal potential E_L and by V_{th} , V_{res} , and τ_{ref} . For excitatory neurons the values used were: $C_m = 0.5$ nF, $g_L = 25$ nS, $E_L = -70$ mV, $V_{th} = -50$ mV, $V_{res}, = -60$ mV, $\tau_{ref} = 2$ ms; and for inhibitory neurons: $C_m = 0.2$ nF, $g_L = 20$ nS, $E_L = -70$ mV, $V_{th} = -50$ mV, $V_{res}, = -60$ mV, $\tau_{ref} = 1$ ms.

The network of neurons was organized according to a ring structure: excitatory and inhibitory neurons were spatially distributed on a ring so that nearby neurons encoded nearby spatial locations. An illustration of this structure is shown in Figure 1A. Connections between neurons were spatially tuned so that nearby neurons with similar preferred directions had stronger than average connections, while distant neurons had weaker connections. The distance dependent connection strength $g_{syn,ij}$ between cells i and j was described by $g_{syn,ij} = W(\theta_i - \theta_j)G_{syn}$, where

$$W(\theta_i - \theta_j) = J^- + (J^+ - J^-)e^{-(\theta_i - \theta_j)^2/2\sigma^2}$$

and J^- was set to satisfy a normalization condition (see Compte et al. 2000). The parameters used were: $\sigma_{E\rightarrow E} = 9.4$ deg, $\sigma_{E\rightarrow I} = \sigma_{I\rightarrow E} = 32.4$ deg, $J^+_{E\rightarrow E} = 5.7$, $J^+_{E\rightarrow I} = J^+_{I\rightarrow E} = 1.4$, $J^+_{I\rightarrow I} = 1.5$. Thus, the connectivity between excitatory and inhibitory neurons was wider and flatter than that between excitatory neurons. The connectivity between inhibitory neurons was not spatially tuned. The strengths of the connections were $G_{E\rightarrow E} = 0.7$ nS, $G_{E\rightarrow I} = 0.49$ nS, $G_{I\rightarrow E} = 0.935$ nS, $G_{I\rightarrow I} = 0.7413$ nS. Apart from stimulus selective inputs, all neurons received uncorrelated random background excitatory input. The times of incoming action potentials were modeled according to a Poisson process with rate 1,800 sp/s. The conductances of this input were $g_{ext\rightarrow E} = 6.5$ nS, $g_{ext\rightarrow I} = 5.8$ nS. The effect of incoming action potentials was modeled through conductance-based synapses. Thus, postsynaptic currents followed the equation:

$$I_{syn} = g_{syn}\, s\, (V_m - V_{syn}),$$

where g_{syn} is the synaptic conductance, s is the synaptic gating variable, and V_{syn} is the synaptic reversal potential ($V_{syn} = 0$ for excitatory synapses, $V_{syn} = -70$ mV for inhibitory synapses). Recurrent excitatory connections were modeled to follow the dynamics of NMDAR mediated transmission, external excitatory inputs to follow AMPAR mediated transmission and inhibitory inputs to follow GABAAR transmission. The dynamics of the AMPAR and GABAAR synaptic gating variables were modeled as an instantaneous jump of magnitude 1 when a presynaptic action potential occurred, followed by an exponential decay with time constant 2 ms for AMPA and 10 ms for GABAA. The NMDAR conductance was voltage-dependent and this was modeled by multiplying g_{syn} by $1/(1+[Mg^{2+}]\exp(-0.062V_m)/3.57)$, with $[Mg^{2+}]=1.0$ mM. The dynamics of the NMDAR synaptic gating were modeled by:

$$\frac{ds}{dt}=\frac{-s}{\tau_s}+\alpha_s x\left(1-s\right),\quad \frac{dx}{dt}=\frac{-x}{\tau_x}+\sum\delta\left(t-t_i\right)$$

where s is the gating variable, x is a synaptic variable proportional to the neurotransmitter concentration in the synapse, t_i are the presynaptic action potential times, $\tau_s=100$ ms is the decay time, $\tau_x=2$ ms controls the rise time, and $\alpha_s=0.45$ kHz controls the saturation properties of NMDAR channels.

Predictions from the model were derived from simulation results. Each simulation started with 100 ms of baseline activity, followed by stimulus specific stimulation during 500 ms and ended with a 500 ms delay period (Figure 1B,D). The locations of the memories for each item were read out using Bayesian or maximum *a posteriori* decoding assuming an extended Poisson model as described by Zemel et al. (1998). This encoding-decoding framework was developed to handle situations where more than a single value (for example several locations) should be encoded and decoded from the neural activity of a population of neurons. Using this method, from the neuronal activity one determines a whole probability distribution over possible locations instead of a single most likely location. This allows for the encoding and decoding of different locations. The decoding distribution of items, that is the probability distribution of angular locations ϕ_j, was estimated given the activity of the excitatory neurons in the last 100 ms of the delay period. For this, we used the function *sqp* from the software package GNU Octave (Eaton et al., 2009) to maximize an approximation of the logarithm of the probability distribution of angular locations ϕ_j (equation 17 of Zemel et al. 1998):

$$AP(\phi_j)=\sum_i r_i\log\left[\sum_j\phi_j f\left(x_{ij}\right)\right]-\varepsilon(\phi_j-\phi_{j+1})^2$$

where r_i is the activity of neuron i, x_{ij} is the difference between the preferred angles of neurons i and j, $f(x_{ij})$ is a neuronal tuning function assumed to be Gaussian with standard deviation 10 deg, set to match the dispersion of the network response to one item (the tuning), and $\varepsilon = 10^{-7}$ is a weighting coefficient of the smoothness prior $\sum(\phi_j - \phi_{j+1})^2$, which imposes smoothness across angular locations ϕ_j. Single values for the estimated locations of memorized items were found by determining the locations ϕ_j corresponding to the local maxima of $AP(\phi_j)$. Before estimation, the spiking activity was resampled to a resolution of 360 for efficiency. Memory imprecision for each stimulus item was quantified as the distance in angle between that item location and the closest local maximum of the posterior probability of item locations, with the restriction that the distance had to be smaller than 35 deg. This restriction assured that in cases where the memory trace vanished during the delay period the particular item was not attributed to a memory trace and instead it was counted as forgotten. In these cases the read-out was taken to be a random location on the circle to mimic a subject guessing a forgotten spatial location. In cases where memory traces merged, the items were attributed to the same local maximum of the posterior probability. To study the effect of the distance between two simultaneously presented items on WM performance, we ran 100 simulations for different angular distance $\Delta\theta$ between the two items (Figure 2A, $\Delta\theta$ from 45 to 90 deg). From these simulations we calculated the angular distance between remembered locations and corresponding item locations. This angular distance is a measure of error or bias in remembered location. If this bias was in the direction of the location of other memorized items (Figure 1B) we defined it as a positive memory bias, corresponding to the attraction of memory traces. If the bias was in the direction opposed to close-by memorized items we defined it as a negative memory bias, corresponding to the repulsion of memory traces. To study the relation between precision and load for different positions of the items we ran 300 simulations for each load and for each stimulus distribution (far or random cases, Figure 2B). For trials labeled random, items were simulated at random around a circle, with the restriction that they could not be closer than 33 deg. In trials labeled far, we applied the additional condition that at least one item per simulation (*far item*) was more than 80 deg apart from all other items. The results were then calculated probing these far items. In particular, we computed standard deviations of the angular distances between remembered locations and corresponding item locations. We also calculated psychometric curves for each load and stimulus distribution. To this end, we counted for all simulations and for a given probed angular distance how many memory traces were counter-clockwise in relation to the probed distance. The results are presented as proportion of memories counter-clockwise to the probed location, as a function of angular distance between the probe and item. We fitted these proportions using probit models with angular distance as independent variable. The probit models were estimated using the Statistics Toolbox of Matlab.

The integration of the model equations was done using a second order Runge-Kutta algorithm. The simulations were performed with code implemented in C++.

Behavioral experiments

We used a vsWM task where the subjects were presented with a set of dots and had to judge after a blank delay period whether a re-appearing dot had been displaced clockwise or counter-clockwise. The experimental paradigm is schematically illustrated in Figure 3A. The stimuli were displayed on a computer screen, on gray background. Participants sat ~60 cm from the screen and were asked to fixate the central black square present during the whole trial time. Participants were also asked to memorize each item *per se* and avoid remembering the dots as a pattern. To limit the efficacy of pattern encoding strategies, we introduced specific constraints for the location of the items in each trial so that geometric symmetries or cardinal directions were avoided (see below).

Each trial started with the presentation of a central fixation cue for 1 s, followed by the presentation of the visual stimulus for 1 s. The stimulus consisted of a set of three or four colored dots (*items*) presented on an invisible circle centered on the fixation point and with a radius subtending a visual angle of 12.4 deg. The items were never presented on the horizontal and vertical diameters of the circle. The colors were attributed randomly to the different items for each trial. The stimulus was followed by 100 ms presentation of a mask consisting of an annulus (radii in visual angle 11.5 deg and 13.2 deg) of a pixelized noise pattern in a gray scale. The mask was followed by the presentation of a probe (in no-delay trials) or by a delay of 1 or 3 s (delay trials) followed by presentation of a probe. The probe stimulus consisted of one of the stimulus dots displaced clockwise or counterclockwise on the invisible circle relative to the original stimulus location. The task consisted in judging the direction of displacement and reporting it by pressing one of two possible keys in a keyboard. Participants were given 5 s to respond and always did it before this time had elapsed. The probe was displayed until the subjects responded. Participants were trained until they showed no problems in associating the directions with the respective keys. It always took less than 48 trials to automatize the association. The amount of displacement in visual angle of the probe item was 0.9, 1.3 or 1.7 deg (4, 6, and 8 deg along the circle), and the probe could not be in a different hemifield than the corresponding target item. In half of the trials the memory of an item that was far from all other items was probed. For these trials, half showed all items far from each other (minimal distance between items was 70 deg along the circle for load 3 and 50 deg for load 4). These trials are referred to as far trials in Figure 3 and as balanced trials in Figure 4. The other half of trials where an item far from all other was probed had two non-probed items close to each other (minimal distance along the circle from the probed item to another item was 90 deg for load 3, and 50 deg for load 4). These trials are referred to as unbalanced trials in Figure 4. Different restrictions on distances were imposed on trials with load 3 and 4 to ensure that a substantial part of the circle was spanned by the locations (in angle) of the items. With this we wanted to minimize possible effects of attention that could appear if subjects could focus on a small portion of the circle, and strategies to store items as geometric patterns. These restrictions resulted in trial types with balanced (invariant) and unbalanced (varying) distances across loads, which we used to demonstrate the *Prediction of conditional dependence of precision on load* (see *Results*). In half of the total number of trials the memory of an item located close to another item was probed (the distance between nearby items was between 10 and 20 deg along the circle, corresponding to a visual angle between 2.2 and 4.2 deg). In half of these trials the probe was displaced outwards or away from the nearby item and in the other half of trials the probe was displaced inwards or towards the nearby item. For each trial type, trials were balanced in relation to relative positions of the dots in the stimulus, the displacements of the probe, the number of items and the presence or absence of a delay period. The experiment was run in sessions of 48 trials, lasting around 5 minutes. Within each session the delay was fixed, and each participant ran 4 sessions for each of 3 possible delays (no-delay, 1 s or 3 s). Type of trial, direction and amount of displacement, color of dots and hemifield of the probed dot were randomized and balanced within each session. The order of the sessions was randomized across participants. 8 healthy participants (4 females) took part in the experiment, with ages between 23 and 37 years and normal or corrected-to-normal vision.

To check for evidence of errors due to misremembering the colors of the items (Bayes et al. 2009, Pertzov et al. 2012, Ma et al. 2014), we conducted a variant of this vsWM experiment. The experimental paradigm is schematically illustrated in Figure 5A. The experiment was exactly as the one described above, except for the response period. After the delay period the fixation dot changed from black to the color of one of the previously presented items. The subject was required to respond by indicating the remembered position of the item matching the color of the fixation mark. To indicate the remembered position, the subjects used a pressure-sensitive tablet and pen. The movement of the pen was reproduced in the visual display as a cursor so that the subjects saw the colored fixation dot moving from the fixation spot to the remembered position. The subject indicated the reported position by releasing the pen from the tablet. All trials had a delay of 3s and separation between nearby items ranged from 3.1 to 4.4 deg of visual angle (14 to 20 deg on the circle). Data was acquired from 4-8 sessions from each of 9 healthy participating subjects (4 females), ages between 21 and 27 years old and showing normal or corrected-to-normal vision. For each subject, sessions were typically acquired in different days. Some participants completed fewer sessions, because they were not available for more data collection. The trials where the probed item was near another item were classified into two trial types, according to the probed item being clockwise or counter-clockwise relative to the nearby item.

Participants for both experiments were recruited among a local community of researchers and students from the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). The experiments were conducted with the approval of the CEIC at the Hospital Clínic in Barcelona (Spain) and informed consent was obtained from all participants before the experiments took place.

Behavioral data analysis

Behavior from the first experiment was measured as the number of correct trials. The results were analyzed using generalized mixed probit models in R (R Development Core Team, 2013), MASS package (Venables and Ripley 2002), with participant as a random factor. For the first test of our first prediction (see *Results*), trial type, delay and the interaction between trial type and delay were used as independent variables or predictors. For the second test of this prediction, the amount of probe displacement was also included.

Since the interaction term was significant in both cases, the data was separated according to delay and a model was fitted using trial type as predictor for test 1 and trial type, amount of probe displacement and the interaction between these two variables as predictors for test 2. For the test of the second prediction (see *Results*), trial type, delay, load and amount of probe displacement were used as independent variables. The model also included interactions between these variables. Since an interaction between delay, trial type and displacement was found to be significant the data was separated according to the delay. A new model without the delay variable was fitted. Since for the delay trials we found an interaction between displacement, load and trial type, the data was further divided according to trial type. For these new data partitions, a model was fitted using amount of probe displacement, load and the interaction between these two variables as predictors.

Behavior in the second experiment was analyzed in three ways. For testing the prediction of attraction, the data was analyzed using a linear mixed model, with participant as a random factor and trial type as a predictor. To test the dependency of memory biases on inter-item distance (Figure 6) we fitted cumulative Gaussians to the cumulative fraction of error reports (Figure 5B), collapsing clockwise and sign-inverted counterclockwise errors, and we used the fitted mean as an estimate of the memory bias (Figure 6A). Positive biases thus reflected attraction and negative biases reflected repulsion of the two memories. In Figure 6B we assessed the significance of each participant's memory bias with a two-sample *t*-test on the error distributions of clockwise and counterclockwise trials. We used a multinomial regression model to test if the relative incidence of significant repulsion biases as compared with attraction biases increased with inter-item distance in our subject population (Figure 6B). The dependent variable could take 3 possible values: attraction, repulsion or no effect. For each subject, we got 3 measurements of the dependent variable, corresponding to 3 bins of distances between items (Figure 6). The model included an intercept and the inter-item distance (taking values 3, 3.75, 4.2) as predictors. The link function was a generalized logit function.

Finally, in order to test alternative statistical models, the data was fitted to three statistical models detailed below using a custom expectation maximization algorithm for the maximum likelihood estimation (Dempster et al. 1977) based on publicly available code (Bays et al. 2009, http://www.paulbays.com). Model comparison was done using Akaike information criterion (AIC) (Akaike 1974), which is a measure of the relative quality of a statistical model for a given data set. Information loss of one model relative to another is then calculated by the differences between AIC values (Burnham and Anderson 2004). The information loss ΔAIC of each model compared to the best (the one with the lowest AIC) was calculated for each subject and then averaged across subjects. The relative likelihood of model *i* relative to the best model was computed as $\exp(\Delta AIC_i/2)$.

Statistical models

A possible explanation for the errors in the task could be a wrong association (or binding) of color and location of the items (Bays et al. 2009, Pertzov et al. 2012, Ma et al. 2014). To access whether interference (attraction) between memory traces of item locations or misbinding best explains our experimental results we used three statistical models, hereby called *swap*, attraction and *attraction+swap* models. All the models assume that the experimental distribution $f_{EXP}(\Delta\theta)$ of errors in reported angle $\Delta\theta$ can be described as a mixture of von Mises components (Figure 5C), a circular analogue of the Gaussian distribution with dispersion parameter σ, defined as $\phi_\sigma(\Delta\theta) = \exp\left[cos(\Delta\theta)/\sigma^2\right]/(2\pi I_0(1/\sigma^2))$, with I_0 the modified Bessel function of order 0.

Swap model. This model is the one introduced by Bays et al. (2009), to account for performance on a recall task where both stimuli and responses are chosen from a circular parameter space. The model assumes that the experimental distribution can be described as a mixture of 3 components:

$$f_{EXP}(\Delta\theta) = p_t\,\phi_\sigma(\Delta\theta) + p_{nt}\frac{1}{n}\sum_i \phi_\sigma(\Delta\theta_i^*) + p_u\frac{1}{2\pi}$$

The first component, weighted by p_t , describes the responses to correctly remembered items, where the subject reports the remembered position with some uncertainty around the error to the actual location of the target item. This is modeled using the von Mises distribution centered around the error to the target $\Delta\theta$, with dispersion parameter σ. The second component, weighted by p_{nt}, describes the responses to nearby non-target items, i.e. responses indicating the remembered location of a non-target item (item with a color different from the probed color). Such responses reflect errors in the binding of color and location of an item (*swap errors*, Bays et al. 2009). This is also modeled using a von Mises distribution with dispersion parameter σ, but now centered on the error to the non-target location $\Delta\theta^* = \theta - \theta_{nt}$. Finally, the third component describes the situation where the item location is forgotten and the subject guesses according to a uniform distribution. The model has 3 parameters p_t , p_{nt} and σ, which can be estimated to fit the experimental data.

Attraction model. In this model the subjects' reports are described by a unimodal von Mises distribution centered on a location intermediate between the target and non-target items. This displacement would occur as a result of the attraction of coding bumps in our more detailed model of Figure 1. This model drops one of the components, the possibility of having swap errors, and introduces a bias b in the mean, representing the attraction effect:

$$f_{EXP}(\Delta\theta) = p_t\, \phi_\sigma(\Delta\theta + b) + p_u \frac{1}{2\pi}$$

Since nearby items were separated by different distances δ_i , the bias b_i in individual trials was constrained to be a fraction of δ_i: $b_i = b'\,\delta_i$, and we estimated the constant factor b'. In total, the model has 3 parameters p_t, σ and b', which can be estimated to fit the measured data.

Attraction+swap model. Finally, both errors might co-exist: in some trials the two features of the stimulus are misbound, but in any case reports (to target or to non-target items) are biased towards the nearby stimulus. This model is the same as the swap model but with one more parameter for the bias:

$$f_{EXP}(\Delta\theta) = p_t\, \phi_\sigma(\Delta\theta + b) + p_{nt} \frac{1}{n}\sum_i^n \phi_\sigma(\Delta\theta_i^* - b) + p_u \frac{1}{2\pi}$$

Note that the bias b (as above, $b_i = b'\,\delta_i$) affects equally both the responses to target and non-target items. This model has 4 parameters p_t , p_{nt} , σ ,and b' which can be estimated to fit the experimental data.

Results

Predictions from the computational model

We used an existing computational model (Compte et al. 2000, Edin et al. 2009) to study vsWM of several simultaneously presented items. For simplicity, we considered only the memory storage of locations at equal eccentricity, so that the item locations could be labeled by an angle θ. The model consists of a one-dimensional network of neurons connected in a topographic manner (Figure 1A), so that neurons encoding nearby locations have stronger connections than neurons encoding far apart locations. This structure enables the network to sustain stimulus selective activity during a delay period (Compte et al. 2000). When plotting the activity of excitatory neurons organized according to their selectivity (Figure 1B, C), the sustained spiking corresponding to a memory trace is visualized as a spatially localized bump of activity in the network (y-axis) that is persistent over time (x-axis). The continuous topographical structure of the network connectivity implies that memory traces maintained simultaneously are not independent and interfere with each other. It further implies that the interference is dependent on the relative locations of the angles memorized, more interference being expected for nearby items than for far-apart items. Possible types of interference of memory traces are attraction (Figure 1B), repulsion and extinction (Figure 1C). To study the effects of interference on vsWM for several items we started by considering two items and we systematically changed the angle $\Delta\theta$ separating them. We measured memory bias as the angular distance between cued locations and memory locations encoded in network activity 0.5 s after stimulus extinction (*Materials and Methods*). Further we defined memory bias as being positive when it reflected attraction between memory traces and negative when it reflected repulsion between memory traces. Figure 2A shows that there is a large attraction effect for angles smaller than 60, and an intermediate repulsion effect for intermediate angles, which disappears as $\Delta\theta$ increases. Our simple model cannot match quantitatively the conditions of a real cortical circuit and hence we do not know in what range of $\Delta\theta$ we should expect the different behaviors, attraction and repulsion. However, we do know that for small angles between items we should have an attraction effect while for very large angles we should have no effect. Based on this we sought to mainly test our model using items very close by or in relative isolation, where we would not need to search for subject-dependent angles leading to repulsion. Hence, the first prediction we aimed at testing in behavioral experiments was that vsWM for adjacent locations should show biases consistent with a perceived attraction between the two items. We refer to this prediction as the *Prediction of attraction biases*. We have however also checked *a posteriori* our experimental data for evidence of the predicted repulsive effects at intermediate inter-item distances (see "Testing repulsive biases").

We then studied how interference affected precision in our network model when the number of items to be memorized (the load) increased. We measured the standard deviation over trials of report errors σ , in simulation series where different number of items (from 1 to 4) where presented to the network for memorization. We considered two cases. In the first case, we minimized interference by keeping distances between items large (*far case*). In the second case, the items were located at random (*random case*). We found that σ depended markedly on load in the random case, while it remained relatively constant as load changed in the far case (Figure 2B). This was because when items were randomly placed, the probability of having items separated in the range of interference (Figure 2A) increased with load. When this probability was only allowed to change minimally with load, as in the far case, σ remained practically constant.

This effect can be demonstrated in the shape of psychometric curves. We used the same simulations as above to derive psychometric curves showing the proportion of items that are judged counter-clockwise to a probed location (Materials and Methods), as a function of angular distance between probed location and item location (Figure 2C, D). For the simulations where only far items were probed, the psychometric curves changed minimally with load (Figure 2C). For the simulations where items were randomly placed, the psychometric curves for loads 3 and 4 showed greater difference (Figure 2D). The different slopes of the psychometric curves reflect different memory precisions for loads 3 and 4, consistent with greater interference of neighboring bumps in load 4 trials. So, our second prediction was that the previously reported loss of precision with load (Bays and Husain 2008) would largely depend on the relative positioning of the items to be memorized, being minimized when the minimal distances between the items in the visual stimuli are large. This prediction will be referred to as the *Prediction of conditional dependence of precision on load*.

Testing the prediction of attraction biases

To test the predictions from the model we used the behavioral experiment illustrated in Figure 3A. The experimental paradigm was adapted from a previously reported paradigm (Bays and Husain 2008) used to investigate the loss of precision with load in a vsWM task in humans. For each trial the subjects were required to keep in mind the locations of 3 or 4 colored dots positioned on an invisible circle (stimulus). After presentation of a visual mask, and in some trials after an additional short delay period (1-3 s), one colored dot reappeared on the invisible circle (probe) and the task was to judge whether it had been displaced clockwise or counter-clockwise. The average accuracy on this task was of 70% correct. All subjects performed significantly above chance level, with accuracies ranging from 59% to 79%.

We conducted two tests of the *Prediction of attraction biases*. For the first test we used the trial types depicted in Figure 3B and labeled them as far (encircled in black) and outwards trials (encircled in green). In the far trials all items were located far apart from each other. In the outwards trials the probed item was presented within a visual angle of 4.2 deg from another item, and it was displaced outwards (or away) from the nearby item (see *Materials and Methods*). In such trials, if the predicted attraction between bumps of activity corresponding to neighboring items occurred (Figures 1B, 2A), we expected the memory of any one of these two adjacent items to be biased towards the middle point between them. As a result, a probe displaced outwards from the corresponding target, whose memorized location has been attracted to the neighboring item, would appear to have been subject to a larger displacement than the actual one. This would help the subject to judge correctly the displacement as *outwards* as opposed to *inwards*. This is schematically depicted in Figure 3D. The bell-shaped curves in Figure 3D represent the probability distributions of the locations stored in memory over multiple trials of two fixed cue stimulus configurations, corresponding to far and outwards trial categories, respectively. One can see that the distance between the mean location of the remembered item and the location of the probe is smaller for far trials (distance 1) than for outwards trials (distance 2). The location of the probed item defines an area under the tail of the probability function which is larger for the far trials (area 1) than for the outwards trials (area 2), and this determines the probability of incorrectly judging the direction of displacement of the probe. This should result in better performance for outwards trials than in a control condition without interference, like in far trials. This is indeed what we observed in our behavioral data set: the fraction of behavioral errors for far trials was significantly larger than that for outwards trials ($p = 0.01$) (Figure 3C). However, the effect observed could have occurred before the delay period, during encoding of the visual stimulus. We rejected this explanation by testing for a difference between trials with and without intervening delay between visual stimulation and response. We found a significant interaction between trial type (far or outwards) and delay ($p = 0.03$) and no significant difference between trial types for no-delay trials (Figure 3C).

For the second test of the Prediction of attraction biases we used the trial types depicted in Figure 3E, and labeled them as counter-clockwise (encircled in red) and clockwise (encircled in blue) trials. In both trial types the probed item was located adjacent to another item. For counter-clockwise item trials the probed item was located counter-clockwise to the neighboring item, and for clockwise item trials the opposite was verified. If attraction occurred, we expected the memory to be biased and the psychometric curves of the two trial types should be horizontally displaced instead of centered at zero probe displacement. The predicted displacement would be clockwise (counter-clockwise) for counter-clockwise (clockwise) item trials, indicating that nearby items were perceived attracted to each other. The data confirmed this prediction (Figure 3F). The two psychometric curves were significantly different from each other ($p < 0.0001$) and the effect appeared with delay, as verified by a significant interaction ($p < 0.0001$) between trial type and delay. Note that the magnitude of the attractive bias was indicative of a partial attraction, not a complete merge of the memories (mean distance between close by items was 3.2 ± 0.14 deg of visual angle, so a complete merge would correspond to a horizontal displacement by 1.6 ± 0.14 deg of visual angle in Fig. 2E).

Testing the prediction of conditional dependence of precision on load

To test this prediction we used two different trial types having in common that the probed item was not in close vicinity to any other item (more than 50 deg along the circle). These different trial types result from the following considerations on the experimental design (for details see *Materials and Methods*). We designed the experiment such that each load condition included a balanced number of trials with probed item far from or close to neighboring items. The former trials (probed item far) contained a balanced number of trials with non-probed items in a far or close configuration, giving rise to the two trial types used in this section. Further, a relatively large part of the circle was covered by the items in each trial by experimental design, in order to minimize possible effects of focusing the attention on a restricted arc. Given these constraints, the two trial types had different inter-item distance properties in relation to load, which we took advantage of to test our second model prediction. In one trial type (far non-probed items) the minimal distance from the probed item to other simultaneously presented items was relatively invariant with load (Figure 4A) and therefore these trials are referred to as balanced trials. In the other trial type (close non-probed items) the minimal distance between the probed item and other items varied markedly between loads (Figure 4B) and therefore they are referred to as unbalanced trials. Note that the labels balanced and unbalanced refer to the distance between probed item and the nearest item being practically invariant (balanced) or varying significantly (unbalanced) across loads. This difference is summarized in Figure 4C showing the mean of the minimal distances for the two loads, which is the same for balanced trials but differs for unbalanced trials. With this set of trials that dissociate load changes from changes in inter-item distances, we went on to test behavioral performance in the task to validate the model's prediction. We found that there was a significant interaction of trial type (balanced/unbalanced) and probe displacement on the fraction of correct responses ($p = 0.05$). Further, we found no difference between the psychometric curves for load 3 and 4 for balanced trials (Figure 4D) but a difference emerged ($p = 0.03$) for unbalanced trials (Figure 4E). The difference between the psychometric curves for loads 3 and 4 in unbalanced trials corresponded to a loss of precision with load (Figure 4F). Precision is here defined as the inverse of the standard deviation of the cumulative normal curves fitted to the data (Bays and Husain 2008), and it quantifies the slope of the psychometric curve at zero probe displacement. This loss of precision was not observed when the distances were balanced across loads (Figure 4F), thus confirming our second Prediction. The observed differential loss of precision with load for unbalanced trial types appeared with delay: We verified that there was a significant interaction between delay, displacement and trial type ($p = 0.05$) and that for the cases with no delay there was no interaction between trial type and displacement or load. That is, the differences in psychometric curves observed in trials with delay were not present with no delay.

Testing a swap-error model

An alternative explanation for the results in Figure 3C and F is that, in some error trials, the subjects swapped the colors and locations of the two memorized nearby items (Bays et al. 2009, Pertzov et al. 2012, Ma et al. 2014). Misremembering the binding between color and location would also result in a reduced fraction of errors for outwards trials. Intuitively, in trials where the color and locations memories are swapped, the perceived displacement of the probe would be large (the distance between items plus the actual displacement) and therefore the response would be correct with higher probability. Thus, we carried out another experiment to contrast this misbinding hypothesis with the memory attraction hypothesis supported by our computational model.

To check for evidence of swap errors in our experimental context, we collected behavioral data in a variant of the original paradigm (Figure 5A and *Materials and Methods*). In this task, nine participants had to report the remembered locations by controlling a cursor. We quantified behavioral performance with the standard deviation of the error-to-target distribution, which was 3.6 ± 0.6 degrees of visual angle across subjects (range: 2 to 7.5 deg). If we excluded trials for which the error to target exceeded 45 degrees along the circle, the error-to-target standard deviation was 2.8 ± 0.4 degrees of visual angle (range: 1.5 to 5.8 deg).

First, we checked that the results shown in Figure 3 were also verified in the modified experimental paradigm. Indeed, we found that there was a significant difference between the reported errors for the counter-clockwise and clockwise trial types (Figure 4B, $p< 0.0001$). Similar as in Figure 3, this data was consistent with attraction of the two memories. We were able to measure the specific fraction of a perfect merge verified in the data. We did this by normalizing the mean error in each trials to the distance between close stimuli. The subjects that showed a significant effect (5 out of 9) presented 26% ± 8% (39% ± 6%) of the attraction expected for a total merge of the memories in clockwise (counter-clockwise) trials.

We then fitted behavioral reports with statistical models that included Gaussian-like distributions around the target memory items (*Materials and Methods*) using a custom expectation maximization algorithm based on (Bays et al. 2009). For all tested models, the dispersion parameter σ estimated from trials with close probed items ($\sigma = 7.63 \pm 0.88$ deg along the circle, n=9) did not differ significantly from that estimated from trials with far probed items (paired t-test, $p> 0.05$, n=9), suggesting that differences in precision between isolated and clustered memory items (Figure 3C) were not due to different memory resolutions in these two situations. Instead, we tested the hypothesis that these differences occurred as a result of memory biases caused by neighboring memories, and we contrasted 3 different models (*Materials and Methods*): an *attraction model* where responses to the target stimulus experienced a mean bias towards the neighboring memory; a *swap model*, in which responses to target stimuli were unbiased, but in some trials responses clustered around the neighboring non-target item; and an *attraction+swap model*, which combined the two situations: a fraction of swap responses and a mean bias toward neighboring memories

(Figure 5C). Note that for the swap model we only considered swaps between close-by items. We compared the estimated maximum likelihoods of each model using differences in the Akaike information criterion (AIC, *Materials and Methods*). We calculated this difference between all the models and the best model. The best model (the one with the lowest AIC) was the *attraction model* for all but one participant, for which the *attraction+swap model* had the lowest AIC (ΔAIC for the *swap model* was 11.7, i.e. a relative likelihood < 0.0001). We excluded this subject to calculate the average information loss of the *swap* and *merge+swap* models relative to the *attraction model* for the other participants. The swap model was the worst of the three statistical models tested (Figure 5D). Adding up AICs for these 8 participants, the relative likelihood of the *swap model* compared to the *attraction model* was below 10^{-4}. These results lead us to discard an explanation based on swap errors alone for the memory attraction that we demonstrated in Figure 3.

Testing repulsion biases

Our model also predicts repulsion for intermediate distances between close-by items (Figure 1B). This is a result of the limited divergence of inhibitory connections in the network (medium-range inhibitory connectivity, see Materials and Methods). We could test this prediction in our second experiment. As shown in Figure 6, the interaction between two nearby memories transitioned from attraction to repulsion as the inter-item distance grew, matching qualitatively our network simulations (Figure 1B). We computed the memory bias from the psychometric curve fit for each subject (Materials and Methods) and plotted it against distance between items (Figure 6A). Across subjects, the attractive memory bias of the psychometric curve decreased significantly (one-tailed paired t-test, $p = 0.02$, $n = 9$) from very close memories (3.-3.5 deg of visual angles, memory bias 95% confidence interval [0 0.7] deg, permutation test $p = 0.05$) to slightly more distant ones (4.2 deg of visual angles), at which point the memory bias became marginally negative (memory bias 95% confidence interval [−1.2 0.1] deg, permutation test $p = 0.07$). In addition, we tested significant memory biases within subjects (Materials and Methods), and we found that the number of subjects with a significant repulsive (attractive) memory bias increased (decreased) with distance between items (Figure 6B, multinomial regression model $p = 0.035$, *Materials and Methods*), indicating a consistent but individually-specific dominance of repulsion for intermediate distances.

Discussion

In the current study we investigated the neural circuit mechanisms of vsWM limitations by formulating predictions from a specific neural circuit hypothesis and by testing them in new behavioral experiments. Specifically, we confirmed model-predicted attractive and repulsive biases in the recollection of items located nearby in space, and we found that the model-predicted reduction in vsWM precision caused by the presence of nearby memorized items could explain the previously reported decrease of vsWM precision with load (Bays and Husain 2008). Taken together, our results support the encoding of vsWM in sustained activity of topographically organized neural circuits.

Item similarity, interference and WM

With this work we contribute to two partially overlapping debates on the behavioral aspects of visual WM. One of these debates revolves around the impact of similarity and interference between items, between items and distractors, and items and landmarks on WM performance. Several studies have demonstrated such effects in vsWM in the presence of landmarks (Werner and Diedrichsen 2002), WM with distractors (Kerzel 2002, Macoveanu et al. 2007, Van der Stigchel et al. 2007, Herwig et al. 2010), memory of sequential items (Papadimitriou et al. 2015), vsWM with memory manipulation (Oberauer and Kliegl 2006), WM of colors (Johnson et al. 2009, Lin and Luck 2009, Elmore et al. 2011, Brady and Alvarez 2011), WM of spatial frequency (Viswanathan et al. 2010, Huang and Sekuler 2010, Mazyar et al. 2012, van den Berg et al. 2012,), WM of sizes (Brady and Alvarez 2011) and WM of orientation (Johnson et al. 2009, van den Berg et al. 2012). However, these studies found discrepant results concerning the impact of item similarity and interference. To our knowledge we are the first to demonstrate a similarity effect for WM of simultaneously memorized spatial locations: the attraction effect of neighboring items. We have provided evidence of a detrimental effect of similarity interference on performance, but we identified one specific condition under which the similarity effect results in vsWM performance enhancement: when the test is presented away from the nearby memorized item (Figure 3C). This is consistent with an attraction of the representations of memorized nearby locations. The analogy between the attraction of memories and the previously reported attraction between a memory and a distractor (Herwig et al. 2010, Macoveanu et al. 2007) and between a memory and an irrelevant previous memory (Papadimitriou et al. 2015) suggests that distractors compete for a representation in the same memory circuits as actual memories, similar to the hypothesis of current neural models of vsWM (Compte et al. 2000, Macoveanu et al. 2007, Cano-Colino et al. 2013).

Conceptually, the very existence of similarity effects has led some authors (Elmore et al. 2011, van den Berg et al. 2012) to interpret them as support for a resources model of WM (Wilken and Ma 2004, Ma et al. 2014), which in its most basic formulation states that WM can be seen as a resource shared between the memory representations of the different items. Indeed, similarity effects are not accommodated naturally in the alternative model, the slots model of WM, which states that one memorizes each item independently until a maximal number of items is reached (Luck and Vogel, 1997, 2013). As some authors have noted, however, similarity or interference effects would not pose any problem to the slots model if they primarily occurred in the encoding phase, not the mnemonic phase of the task (Johnson et al. 2009, Lin and Luck 2009). In our experiments, similarity effects are not present when there is no delay period and the task is otherwise identical. This suggests that spatial interference of memorized locations occurs during the maintenance of information in WM and not during the encoding of information. An alternative explanation for the results in Figure 3C,F is that the participants remembered in some trials the colors of two nearby items swapped (Bays et al. 2009, Pertzov et al. 2012, Ma et al. 2014). To have an idea about how prevalent this type of errors was in our experimental setup, we ran an additional experiment. We found clear evidence that swap errors alone cannot explain the prediction of attraction biases and so we conclude that attraction of memory traces is a more plausible explanation for our results. Note however, that the amount of swap errors is probably closely related to the specifics of the task and previous studies that found substantial evidence for swap errors did not use vsWM but tasks based on WM of color (Bays et al. 2009) or orientation (Pertzov et al. 2012).

WM precision with load

A second debate concerns the relation between precision of vsWM and number of items to memorize (WM load), and its implications for the nature of WM. Some authors found a decrease of precision with load (Bays and Husain 2008, Bays et al. 2009) supporting the resources model (Wilken and Ma, 2004) of WM, while others found a saturation of precision with load (Zhang and Luck 2008) supporting models of the family of the slots model (Luck and Vogel 1997, Zhang and Luck 2008). Crucially, in these slots models information about further items cannot enter WM after reaching a maximum number of memorized items. Much ongoing research on WM limitations has focused on resolving the dichotomy between these two alternatives providing new experimental evidence and leading to further development of algorithmic models, including hybrid models with characteristics from the slots and resources models (Alvarez and Cavanagh 2004, Xu and Chun 2006, Bays and Husain 2008, Zhang and Luck 2008, Bays et al. 2009, Anderson et al. 2011, Buschman et al. 2011, Elmore et al. 2011, van den Berg et al. 2012, Luck and Vogel 2013, Ma et al. 2014). A parallel line of research is focusing on the circuit mechanisms of vsWM in biologically detailed network models (Compte et al. 2000, Macoveanu et al. 2007, Edin et al. 2009, Wei et al. 2012, Bays 2014) that are typically hard to classify into any of these abstract model categories. We took one such biologically detailed model and we found that the interference between items causes, on average, loss of memory precision (see also Wei et al. 2012). As the number of items in a constrained area increases, the probability of having interference between memories increases and hence a loss of precision with load is observed. The model thus predicts that the decrease of vsWM precision with load depends largely on the relative location of the items. Our experimental results were consistent with a distance-dependent relation between precision and load, showing both a reduction of precision with load (Figure 4E) and a lack thereof (Figure 4D) on the same behavioral data, depending on a selection of trials based on inter-item distance. This suggests that inter-item distance could be a factor explaining the conflicting results in the literature (Bays and Husain 2008, Zhang and Luck 2008). Furthermore, our experiments showed that the relationship between spatial memory precision and load emerged through the delay. This suggests that explanations based on the processes of memory encoding and decoding (Bays 2014) need to incorporate also the role of memory maintenance mechanisms.

WM model

The network model was used with the same parameters as in (Edin et al. 2009), without further tuning. We did not seek a quantitative match between the angles or times used for the behavioral experiments and model simulations. Such a match can be sought by changing parameters of the model, for example increasing the size of the network would make the values of angular distances and times in the model approach those of the experiments, at the cost of slower simulations. Such procedure would make model testing unpractical, without providing any significant conceptual advantage. Hence, we searched for qualitative robust predictions to test experimentally. Consistent with this, Wei and coauthors (Wei et al. 2012) working in parallel in a similar model derived predictions qualitatively in agreement with ours but based on different activity patterns. Indeed, their model differs from ours fundamentally in that it features a normalization regime where the same number of active neurons is split among the number of items encoded, with the overall population activity invariant with load (see also Bays 2014). This is not the regime of operation of our network, which shows graded rate responses and mean firing rates increasing with load (Edin et al. 2009). Another difference between the models is that our model, but not the model of Wei et al. (2012), predicts repulsion between memory traces. Our experimental results (Figure 6) show evidence for repulsion, hence supporting our model. Further exploration of the regimes where the two models operate should provide new discriminating predictions to test against experimental data in the future. Johnson and coauthors (Johnson et al. 2009) also proposed a firing rate model explaining color similarity effects based on a specific decoder mechanism, in contrast with our model which allocates the mechanism in the dynamics of the circuit during the maintenance phase. Our experimental results for vsWM show that the similarity effects appear with delay and therefore are not originated during the encoding or decoding phases of the task. This is consistent with interference during the active maintenance of memory. We note however that different mechanisms might be behind the effects described for color (Johnson et al. 2009) or orientation (Bays 2014) WM tasks. Finally, our model did not simulate all components of the tasks: Our tasks demanded the binding of two different features (color and position), while the model was only simulating the storage of position. This is partly because of the lack of a consensual model for feature binding in working memory, but also because the behavioral effects that we are reporting proved not to depend crucially on such binding. Indeed, we demonstrate in our last experiment that the attraction effect is independent of swap errors. This result justifies interpreting our data with a simplified model representing only location information. However, a complete understanding of this task will require explicitly simulating the binding component.

Our results advance our understanding of vsWM in terms of its neuronal circuit underpinnings by providing evidence for a critical assumption of an explicit computational model of vsWM. Namely, that vsWM is supported by a network of neurons organized according to a continuous topography in terms of internal connectivity and external inputs received. This topographical connectivity enables the model to sustain a continuous attractor mechanism, on which memories of neighboring items interfere (Amari 1977). Recently, direct experimental evidence from neural activity in the prefrontal cortex of monkeys performing a single-item spatial working memory task has been obtained in favor of this continuous attractor mechanism (Wimmer et al. 2014). Here, the consistency of our experimental results with the model predictions in the case of multi-item working memory lends further support to the continuous attractor as the basis of vsWM. Further, the model explains parsimoniously behavioral effects that cannot be consistently integrated within the prevalent algorithmic models for vsWM. This underscores the potential of using a circuit-based framework to interpret experimental results on the mechanisms of vsWM.

Funding

This work was supported by the Ministry of Economy and Competitiveness of Spain, the European Regional Development Fund (Ref: BFU2009-09537, BFU2012-34838) and the Karolinska Institutet Strategic Neuroscience Program. R.A. was supported by the Generalitat de Catalunya (Beatriu de Pinós program, 2007BP-B100135). The work was carried out at the Esther Koplowitz Centre, Barcelona.

Acknowledgments

We thank Fredrik Edin, Torkel Klingberg, Fiona McNab and Chantal Roggeman for code, data used in exploratory pilot analyses and valuable discussions of this work.

References

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19:716–723, 1974

Alvarez G, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychol Sci. 15:106-111, 2004

Amari SI. Dynamics of pattern formation in lateral-inhibition type neural fields. Biol Cybernetics. 27:77–87, 1977

Anderson D, Vogel E, Awh E. Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. J Neurosci. 31:1128–1138, 2011

Baddeley A. Working memory. New York: Oxford University Press, 1986

Bays P. Noise in neural populations accounts for errors in working memory. J Neurosci. 34: 3632-3645, 2014

Bays P, Catalao R, Husain M. The precision of visual working memory is set by allocation of a shared resource. J Vis. 9:1–11, 2009

Bays P, Husain M. Dynamic shifts of limited working memory resources in human vision. Science. 321:851–854, 2008

Brady T, Alvarez G. Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. Psychol Sci. 22:384-92, 2011

Burnham K, Anderson D. Multimodel inference understanding AIC and BIC in model selection. Sociological methods & research. 33:261–304, 2004

Buschman T, Siegel M, Roy J, Miller E. Neural substrates of cognitive capacity limitations. Proc Natl Acad Sci USA. 108:11252–11255, 2011

Cano-Colino M, Almeida R, Compte A. Serotonergic modulation of spatial working memory: predictions from a computational network model. Front Integr Neurosci. 7:71, 2013

Compte A, Brunel N, Goldman-Rakic P, Wang X-J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb Cortex. 10:910-923, 2000

Constantinidis C, Franowicz M, Goldman-Rakic PS. Coding specificity in cortical microcircuits: A multiple-electrode analysis of primate prefrontal cortex. J Neurosci. 21:3646–3655, 2001

Conway A, Kane M, Engle R. Working memory capacity and its relation to general intelligence. Trends Cogn Sci. 7:547–552, 2003

Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39:1-38, 1977

Eaton J, Bateman D, Hauberg S. GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform, 2009

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A. Mechanism for top-down control of working memory capacity. Proc Natl Acad Sci USA. 106:6802–6807, 2009

Elmore L, Ma W, Magnotti J, Leising K, Passaro A, Katz J, Wright A. Visual short-term memory compared in rhesus monkeys and humans. Curr Biol. 21:975–979, 2011

Funahashi S, Bruce C, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J Neurophysiol. 61:331–349, 1989

Georgopoulos A, Schwartz A, Kettner R. Neuronal population coding of movement direction. Science. 233:14161419, 1986

Herwig A, Beisert M, Schneider W. On the spatial interaction of visual working memory and attention: Evidence for a global effect from memory-guided saccades. J Vis. 10:1–10, 2010

Huang J, Sekuler R. Distortions in recall from visual memory: two classes of attractors at work. J Vis. 10:1–27, 2010

Inoue M, Funahashi S. Prefrontal delay-period activity is affected by visual cues presented outside the memory field. Neuroreport. 13:2097–2101, 2002

Johnson J, Spencer J, Luck S, Schöner G. A dynamic neural field model of visual working memory and change detection. Psychol Sci. 20:568–577, 2009

Kastner S, DeSimone K, Konen C, Szczepanski SW K, Schneider K. Topographic maps in human frontal cortex revealed in memory-guided saccade and spatial working-memory tasks. J

Neurophysiol. 97:3494–3507, 2007

Kerzel D. Memory for the position of stationary objects: disentangling foveal bias and memory averaging. Vis Res. 159-167, 2002

Lara AH, Wallis JD. Executive control processes underlying multi-item working memory. Nat Neurosci. 17:876-883, 2014

Lee D, Reis B, Seung H, Tank D. Nonlinear network models of the oculomotor integrator. In Bower J, editor. Computational neuroscience, Trends in Research. Plenum Press, 1997

Lin P, Luck S. The influence of similarity on visual working memory representations. Vis Cogn. 17:356–372, 2009

Luck S, Vogel E. The capacity of visual working memory for features and conjunctions. Nature. 390:279–281, 1997

Luck S, Vogel E. Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn. Sci. 17:391–400, 2013

Ma W, Husain M, Bays P. Changing concepts of working memory. Nat. Neurosci. 17:347-356, 2014

Macoveanu J, Klingberg T, Tegnér J. A biophysical model of multiple item working memory: a computational and neuroimaging study. Neuroscience. 141:1611–1618, 2006

Macoveanu J, Klingberg T, Tegnér J. Neuronal firing rates account for distractor effects on mnemonic accuracy in a visuo-spatial working memory task. Biol Cybern. 96:07–419, 2007

Mazyar H, van den Berg R, Ma W. Does precision decrease with set size? J Vis. 12:1–16, 2012

Oberauer K, Kliegl R. A formal model of capacity limits in working memory. J Mem Lang. 55:602–626, 2006

Papadimitriou C, Ferdoash A, Snyder LH. Ghosts in the machine: memory interference from the previous trial. J Neurophysiol 113:567-577, 2015

Pertzov Y, Dong MY, Peich MC, Husain M. Forgetting what was where: the fragility of object-location binding. PLoS ONE. 7:e48214, 2012

R Development Core Team. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, 2013

Schluppeck D, Curtin C, Glimcher P, Heeger D. Sustained activity in topographic areas of human posterior parietal cortex during memory-guide saccades. J Neurosci. 26:5098-5108, 2006

Tuckwell H. Introduction to theoretical neurobiology. Cambridge University Press, 1988

Van den Berg R, Shin H, Chou W, George R, Ma W. Variability in encoding precision accounts for visual short-term memory limitations. Proc Natl Acad Sci USA. 109:8780–8785, 2012

Van der Stigchel S, Merten H, Meeter M, Theeuwes J. The effects of a task-irrelevant visual event on spatial working memory. Psychonomic Bullet Rev. 14:1066–1071, 2007

Venables, W, Ripley B. Modern Applied Statistics with S. 4th ed. New York: Springer, 2002

Viswanathan S, Perl D, Visscher K, Kahana M, Sekuler R. Homogeneity computation: how interitem similarity in visual short-term memory alters recognition. Psychon Bull Rev. 17:59–65, 2010

Warden M, Miller E. The representation of multiple objects in prefrontal neuronal delay activity. Cereb Cortex. 17:41–50, 2007

Wei Z, Wang X-J, Wang D. From distributed resources to limited slots.in multiple-item working memory: a spiking network model with normalization. J Neurosci. 32:11228–11240, 2012

Werner S, Diedrichsen J. The time course of spatial memory distortions. Memory & cognition. 30:718-730, 2002

Wilken P, Ma W. A detection theory account of change detection. J Vis. 4:1120–1135, 2004

Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci. 173:431-439, 2014

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

Zemel R, Dayan P, Pouget A. Probabilistic interpretation of population codes. Neural Computation. 10:403–430,1998

Zhang W, Luck S. Discrete fixed-resolution representations in visual working memory. Nature. 453:233–235, 2008

Figure legends

Figure 1: The biophysical network model. (A) Schematic representation of the ring structure of the network model (left) and of the connectivity structure (right) between excitatory (triangles) and inhibitory neurons (grey circles). Neurons encoding similar angles were strongly connected as illustrated by the width of the lines connecting cells. Connections onto excitatory neurons are indicated with a solid line and onto interneurons with a dashed line, excitatory connections are indicated in black and inhibitory in grey. (B) Example activity of excitatory neurons in the network, when items were located in the vicinity of each other leading to attraction of the memory traces. (C) Example activity of excitatory neurons in the network in a trial with 3 presented items, illustrating the loss of a memory trace during the delay period.

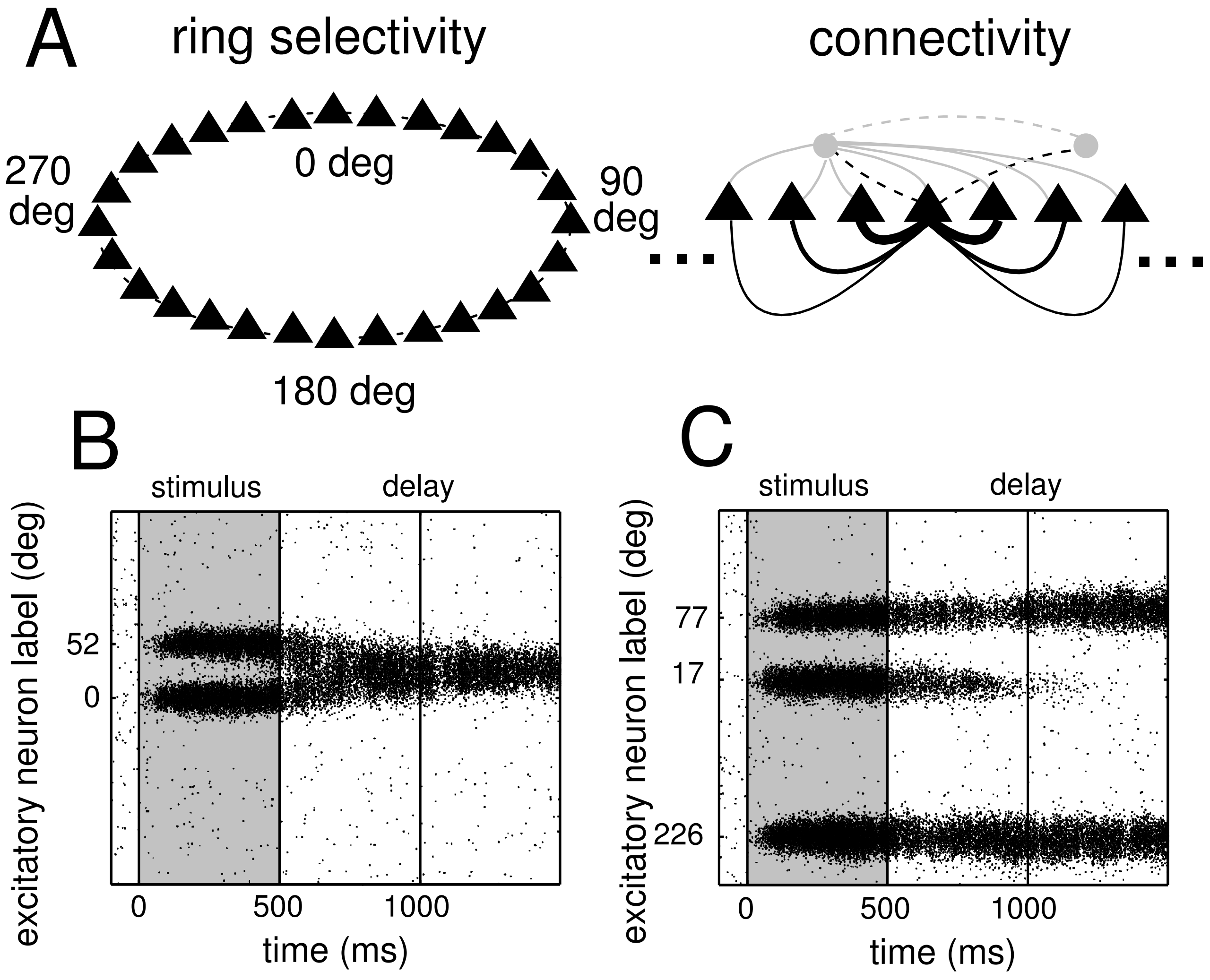
Figure 2: The biophysical network model predicts behavioral effects in multi-item vsWM tasks. (A) Memory bias as a function of angle between two items simultaneously presented. The results are averages over 100 simulations and are based on memory traces after 500 ms from stimulus offset. Memory biases towards the other presented item (attraction) were defined as positive, while biases away from the other presented item (repulsion) were defined as negative. The bias for small angles is easier to explore experimentally and leads to the formulation of the Prediction of attraction biases. (B) Standard deviation error of the memory trace after 500 ms as a function of load. The standard deviation error was relatively constant for far items (circles, dashed line) and increased with load for randomly located items (triangles, solid line), leading to the Prediction of conditional dependence of precision on load. (C) Proportion of probes judged to be displaced counter-clockwise from the memorized item. The results are for far items and loads 3 (black) and 4 (gray) and were fitted with a probit model with displacement of the probe as independent variable. (D) Same as (C) but for randomly located items. Panels (C) and (D) use the same simulations as in (B) and show that for far items there is no decrease in precision with load, which is observed for randomly located items. This observation also leads to the Prediction of conditional dependence of precision on load.

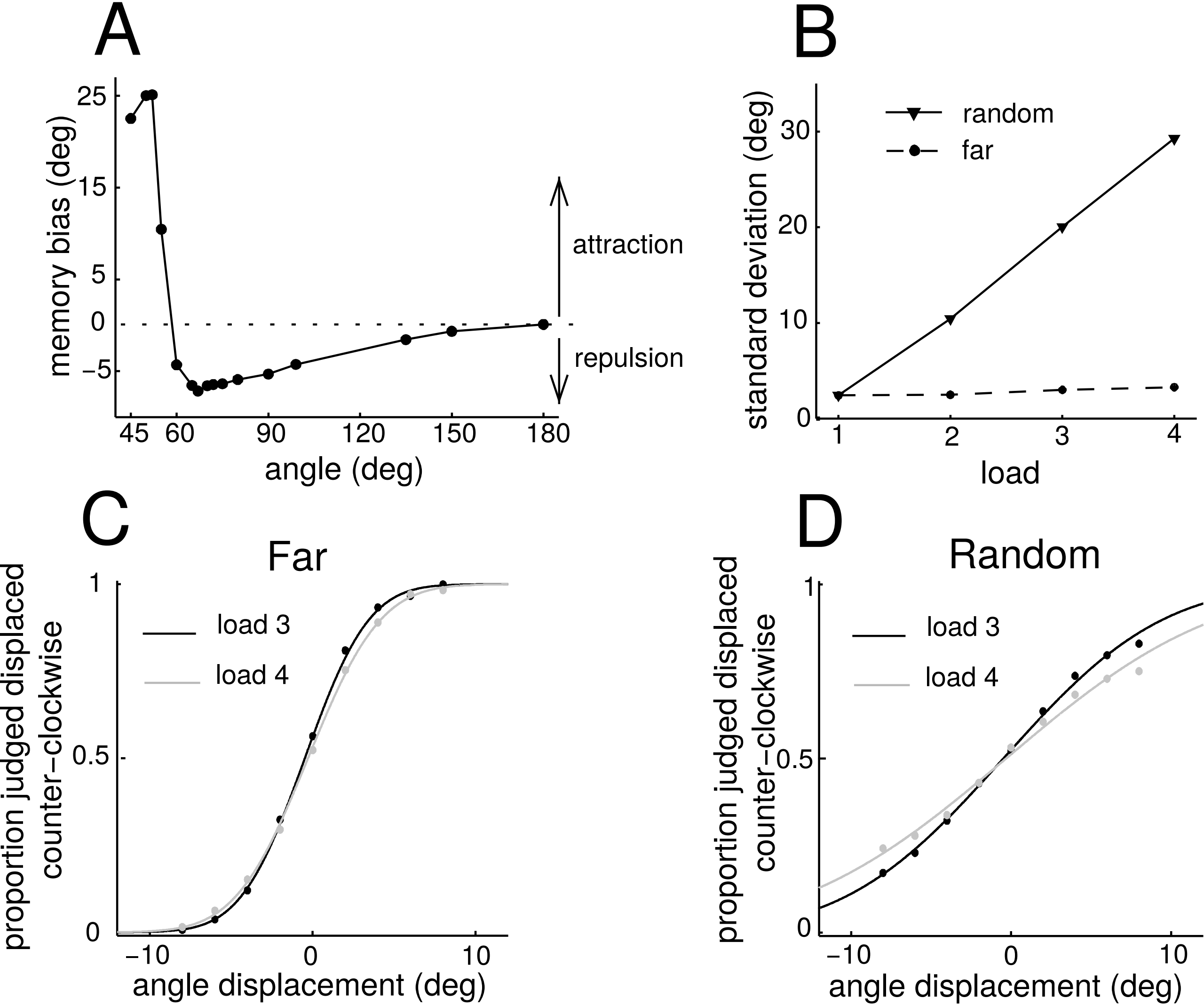
Figure 3: Behavioral data supports the model-derived Prediction of attraction biases. (A) Schematic illustration of the paradigm used in the behavioral experiment. (B) Illustration of the sorting of trials according to relative positions of the items. In one case, items were far from each other (far trials, framed in black). In the other case, the target item was presented close to another item and was displaced away from its neighbor during probing (outwards trials, framed in green). (C) Fraction of errors averaged over participants (n=8) in 48 trials of each trial type (delay/no delay and far/outwards). Data was analyzed using a probit model. Significant differences are indicated with a *. There was a significant interaction between delay and trial type. For no delay trials there was no difference between the fraction of errors for far and outward trials, while there was a significant difference for delay trials. Error bars indicate standard errors of the mean. (D) Schematic illustration of the mechanism thought to underlie the decrease in errors for outward trials compared with far trials. The bell-shaped curves represent the probability distribution of the remembered locations. The probed item defines an area under the probability function. This area is the probability of incorrectly judging the direction of displacement of the probe and is larger for far than outward trials ($a2<a1$). The distance between location of the item and location of the probe is larger for outward trials ($d1<d2$). Hence, the probability of a correct response in outwards trials is larger than in far trials, as observed experimentally. (E) Illustration of another sorting of trials, all containing the probed item in the vicinity of another item. Trials were sorted according to the clockwise (blue) or counter-clockwise (red) location of the probed item relative to the neighboring item. (F) Psychometric curves for clockwise (blue) and counter-clockwise (red) trials were horizontally displaced in relation to each other. Curves resulted from a probit model fit to data from all participants (n=8). The results of C and F are consistent with the Prediction of attraction biases.

Figure 4: Behavioral data supports the model-derived Prediction of conditional dependence of precision on load. (A and B) Histograms of the distances between the target or probed item to the nearest non-probed item for loads 3 (contour only bars) and 4 (filled bars) for the case of balanced or invariant distances across load, A, or for the case of unbalanced or varying distances across load trials, B (see Results). Each combination of load and trial type (balanced/unbalanced) included 384 trials. (C) Mean distances from the target to the nearest neighbor for loads 3 and 4 and for balanced (*t*riangles) and unbalanced (circles) distances. Error bars indicate standard deviations. (D) Psychometric curves for load 3 (black) and 4 (gray) for the case of balanced distances. Curves resulted from a probit model fit to data from all participants (n=8). (E) The same as in D for unbalanced distances. (F) Precision derived from panels D, E decreased with load for unbalanced distances (circles) while it remained unchanged for balanced distances (triangles). Error bars indicate standard errors of the mean.

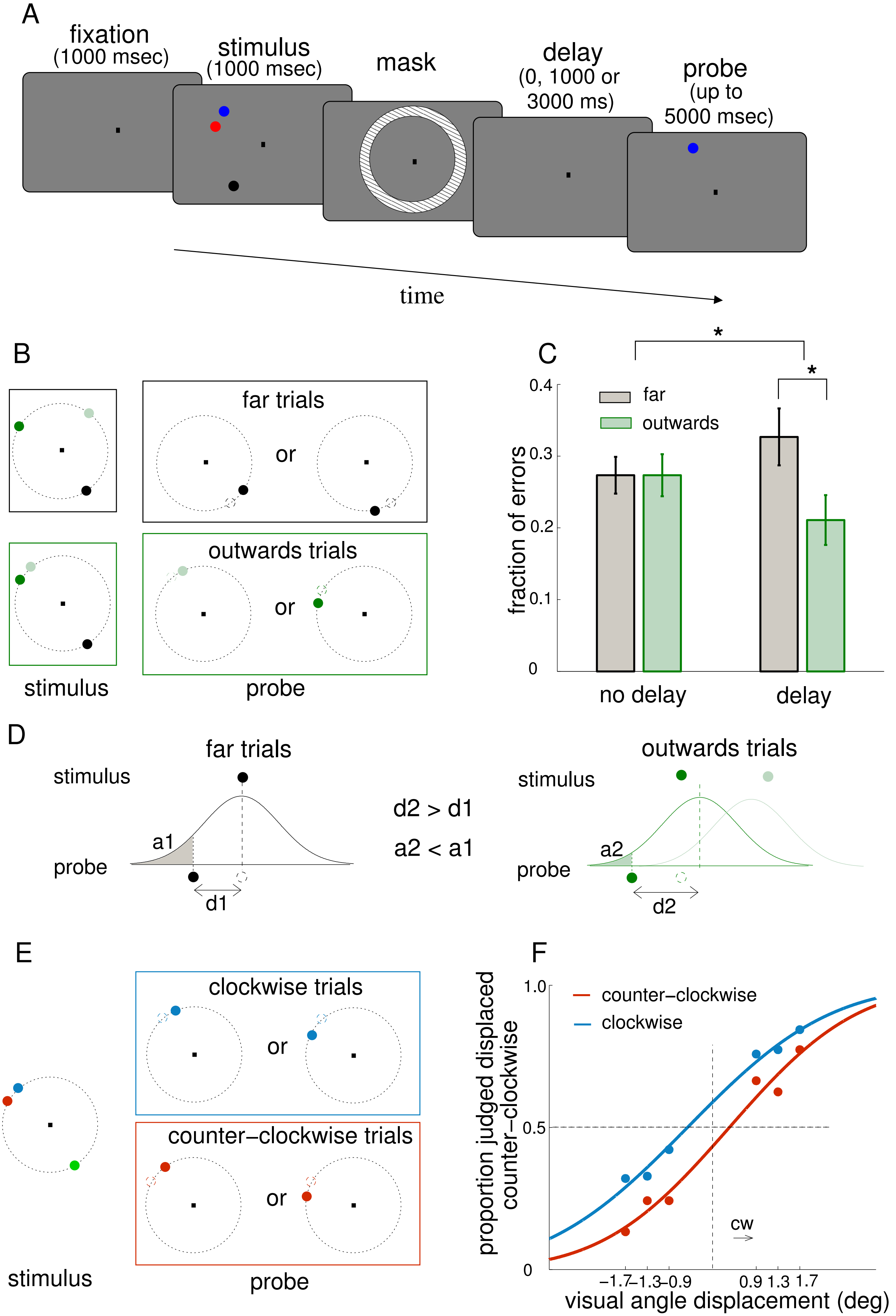
Figure 5: Behavioral data suggests that attraction of memory representations and not swap errors is responsible for memory biases observed in close trials. (A) Schematic illustration of the modified experimental paradigm, where participants indicated the remembered target location upon appearance of a colored cue in the center of the screen. (B) Top: distributions of error to target for clockwise (gray) and counter-clockwise (black) trials differed significantly ($p < 0.00005$, data from all participants n=9), revealing an attractive bias. Bottom: Cumulative proportion of errors to target from the distributions in top panel, to compare with psychometric curves in Figure 2E. Data was fitted with a cumulative normal function. (C) Schematic illustration of the probability density function for each of the 3 models tested. *Swap* (black), attraction (dark gray) and attraction+*swap* model (light gray). (D) Average information loss ΔAIC across subjects (n=8) for *swap* and attraction+*swap* models compared to the attraction model, the best model for data from these participants.

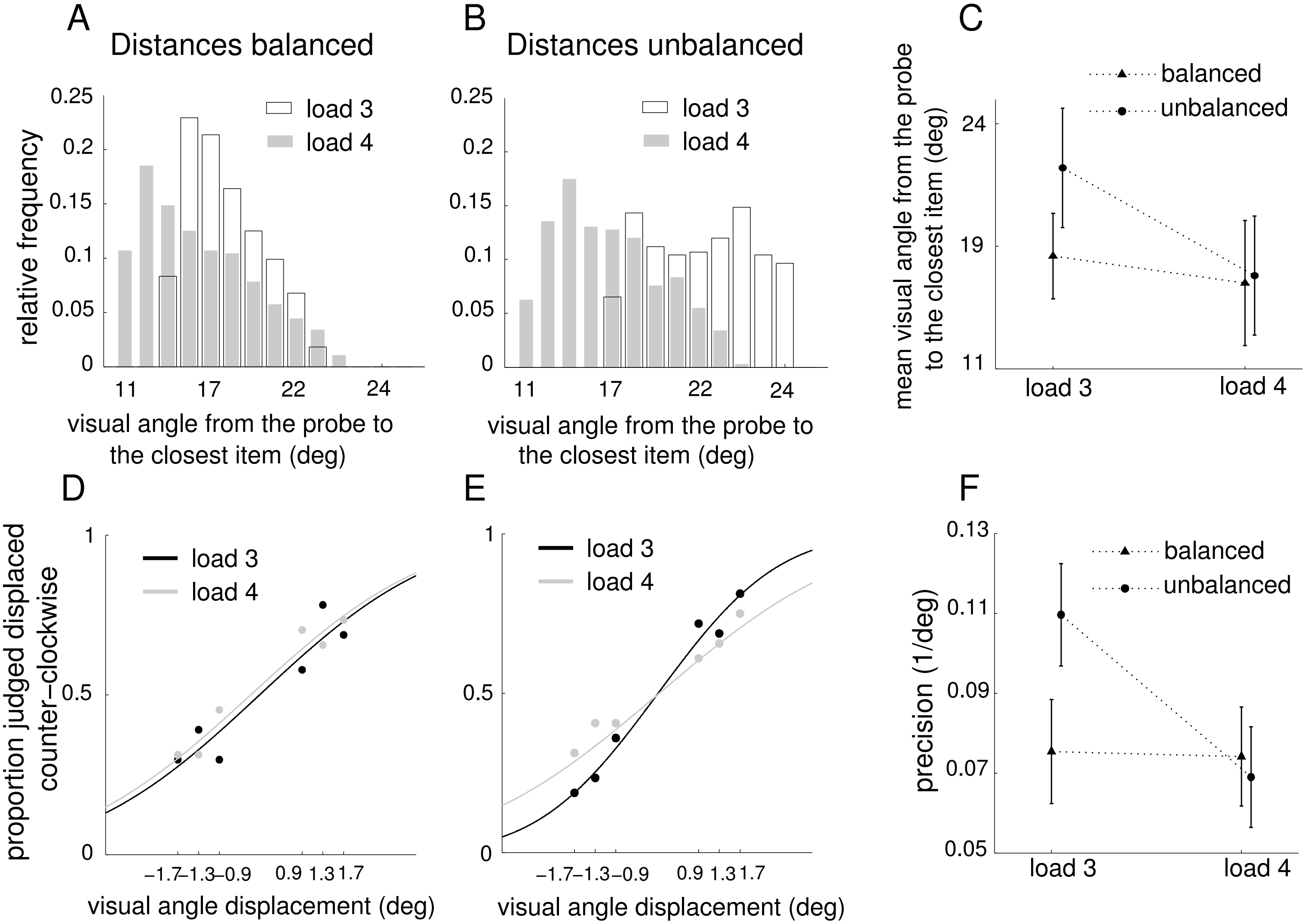
Figure 6: Memory repulsion emerges for intermediate distances between close-by items. (A) Subject-averaged memory bias (*Materials and Methods*) for trials with different distances between memorized close-by items (x-axis). Shadows indicate bootstrap-derived 95% confidence intervals. Stars denote significant difference as evaluated with one-tailed paired t-test at $p < 0.05$. (B) Number of subjects with significant (t-test $p < 0.05$) attractive and repulsive memory bias in trials with different inter-item distance.


A
ring selectivity
0 deg
90 deg
180 deg
270 deg
connectivity
B
stimulus
delay
excitatory neuron label (deg)
52
0
0
500
1000
time (ms)
C
stimulus
delay
excitatory neuron label (deg)
77
17
226
0
500
1000
time (ms)

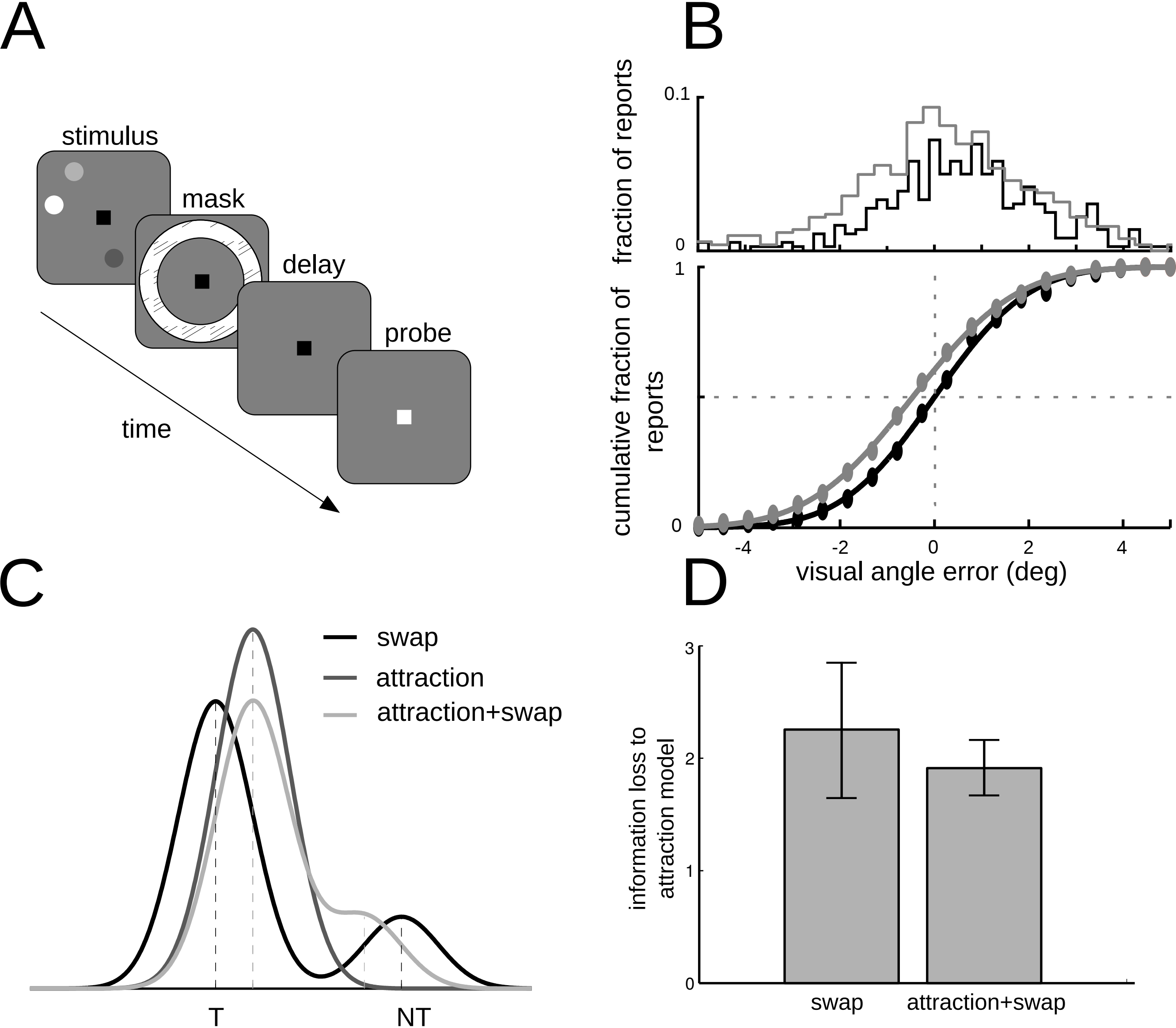


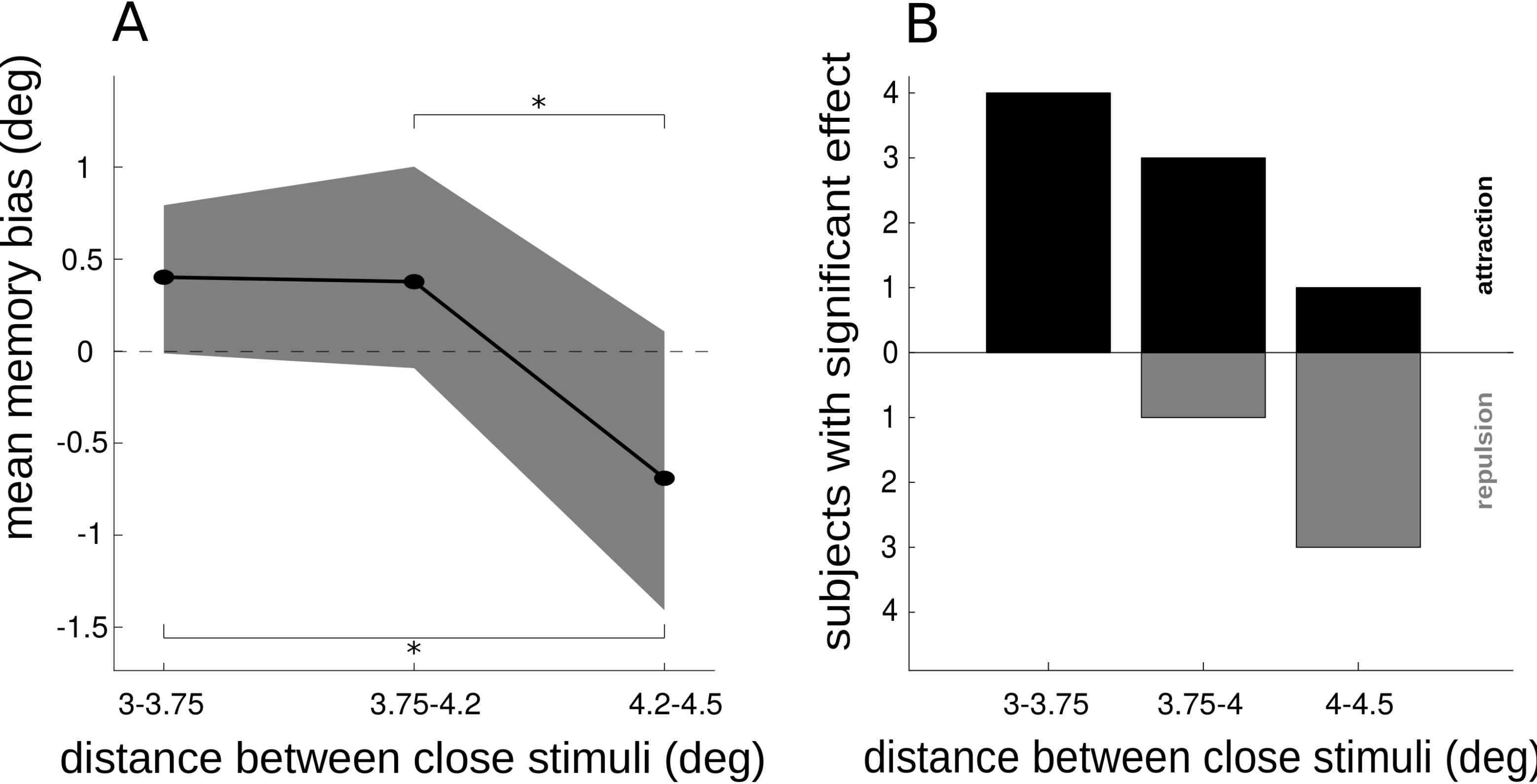
A
memory bias (deg)
25
15
5
0
−5
attraction
repulsion
45
60
90
120
150
180
angle (deg)
B
standard deviation (deg)
random
far
30
20
10
0
1
2
3
4
load
C
Far
proportion judged displaced counter−clockwise
load 3
load 4
1
0.5
0
−10
0
10
angle displacement (deg)
D
Random
proportion judged displaced counter−clockwise
load 3
load 4
1
0.5
0
−10
0
10
angle displacement (deg)


A
fixation
(1000 msec)
stimulus
(1000 msec)
mask
delay
(0, 1000 or 3000 ms)
probe
(up to 5000 msec)
time
B
far trials
or
outwards trials
or
stimulus
probe
C
fraction of errors
0.4
0.3
0.2
0.1
0
far
outwards
*
*
no delay
delay
D
far trials
stimulus
probe
a1
d1
d2 > d1
a2 < a1
outwards trials
stimulus
probe
a2
d2
E
clockwise trials
or
counter–clockwise trials
or
stimulus
probe
F
proportion judged displaced counter–clockwise
1.0
0.5
0
counter–clockwise
clockwise
cw
−1.7 −1.3 −0.9
0.9 1.3 1.7
visual angle displacement (deg)



A
Distances balanced
relative frequency
load 3
load 4
0.25
0.2
0.15
0.1
0.05
0
11
17
22
24
visual angle from the probe to the closest item (deg)
B
Distances unbalanced
load 3
load 4
0.25
0.2
0.15
0.1
0.05
0
11
17
22
24
visual angle from the probe to the closest item (deg)
C
mean visual angle from the probe to the closest item (deg)
balanced
unbalanced
24
19
11
load 3
load 4
D
proportion judged displaced counter-clockwise
load 3
load 4
1
0.5
0
−1.7 −1.3 −0.9
0.9 1.3 1.7
visual angle displacement (deg)
E
load 3
load 4
1
0.5
0
−1.7 −1.3 −0.9
0.9 1.3 1.7
visual angle displacement (deg)
F
precision (1/deg)
balanced
unbalanced
0.13
0.11
0.09
0.07
0.05
load 3
load 4


A
stimulus
mask
delay
probe
time
B
fraction of reports
0.1
0
cumulative fraction of reports
1
0
-4
-2
0
2
4
visual angle error (deg)
C
swap
attraction
attraction+swap
T
NT
D
information loss to attraction model
3
2
1
0
swap
attraction+swap


A
mean memory bias (deg)
1
0.5
0
-0.5
-1
-1.5
*
*
3-3.75
3.75-4.2
4.2-4.5
distance between close stimuli (deg)
B
subjects with significant effect
4
3
2
1
0
1
2
3
4
attraction
repulsion
3-3.75
3.75-4
4-4.5
distance between close stimuli (deg)