Concurrent maintenance of both veridical and transformed working memory representations

Güven Kandemir[1,2]*, Michael J. Wolff[3,4]*, Aytaç Karabay[1,5], Mark G. Stokes[3], Nikolai Axmacher[6], & Elkan G. Akyürek[1]

[1] Department of Experimental Psychology, University of Groningen, The Netherlands

[2] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, The Netherlands

[3] Department of Experimental Psychology, University of Oxford, United Kingdom

[4] Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany

[5] Department of Psychology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[6] Department of Neuropsychology, Faculty of Psychology, Ruhr University Bochum, Germany

*shared first authorship

Running head: Transformation and representation in working memory

28/10/2022

Word count: 11245

Address correspondence to:

Elkan Akyürek

Department of Experimental Psychology, University of Groningen

Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Email: e.g.akyurek@rug.nl

Telephone: +31 (0)50 3636406

Abstract

In a dynamic environment, the already limited information that human working memory can maintain needs to be constantly updated to optimally guide behaviour. Indeed, previous studies showed that working memory representations are continuously being transformed during delay periods leading up to a response. This goes hand-in-hand with the removal of task-irrelevant items. However, does such removal also include veridical, original stimuli, as they were prior to transformation? Here we aimed to assess the neural representation of task-relevant transformed representations, compared to the no-longer-relevant veridical representations they originated from. We applied multivariate pattern analysis to electroencephalographic data during maintenance of orientation gratings with and without mental rotation. During maintenance, we perturbed the representational network by means of a visual impulse stimulus, and were thus able to successfully decode veridical as well as imaginary, transformed orientation gratings from impulse-driven activity. On the one hand, the impulse response reflected only task-relevant (cued), but not task-irrelevant (uncued) items, suggesting that the latter were quickly discarded from working memory. By contrast, even though the original cued orientation gratings were also no longer task-relevant after mental rotation, these items continued to be represented next to the rotated ones, in different representational formats. This seemingly inefficient use of scarce working memory capacity was associated with reduced probe response times and may thus serve to increase precision and flexibility in guiding behaviour in dynamic environments.

Keywords: working memory; mental rotation; EEG; decoding; impulse perturbation

Introduction

Working memory (WM) is more than a short-term storage for sensory input; it comprises the ability to actively manipulate and modify information as well (Baddeley, 2003). This is essential, since our environment is constantly in flux, and we need to organize perceptual information in that dynamic context. Indeed, one of the key features of WM is that it can adapt and alter its contents to anticipate change. For example, the egocentric location of an object will change constantly as we move around, yet we can easily point at its current spatial location even when it is temporarily out of sight (Rieser et al., 1986), showing that previous sensory input can be recalled and transformed to predict its changed state. Thus, the ability to not only maintain but also add information through imagery, and to update existing information in WM, is essential to guide both current and future behaviour (Rainer et al., 1999).

The brain seems to use the same neural substrate to maintain both current and previous sensory inputs, as well as mentally imagined items. Imagery and sensory-driven perception activate spatially overlapping regions in visual cortex, and share a common coding scheme (Stokes et al., 2009). Furthermore, items maintained in WM as well as items that are only imagined are both represented in early visual brain regions, with common activity patterns resembling those generated by direct visual input (Albers et al., 2013). These neural commonalities make a functional interpretation seem appealing (but see also Linke & Cusack, 2015; Iamshchinina et al., 2021), and several authors have highlighted the similarity of WM, imagery, and perception (Dijkstra et al., 2019; Pearson, 2019), all of which may use primarily sensory brain regions, such as visual cortex, as a "blackboard" (Roelfsema & De Lange, 2016).

Loaded with such wide-ranging demands, an important question is how the brain keeps its blackboard clean and fit for re-use. Not only that, but it also needs to be able to distinguish between current sensory input and maintained or imagined information. This is further compounded by the fact that it is not sufficient to retain only stimulus-related information, since WM ultimately serves to guide behaviour (Stokes et al, 2013; van Ede et al., 2019). In many cases, information that was once encoded to reflect certain properties of the environment may thus have to be updated or even superseded to relate to a possible course of action. On the one hand, it could be an economical strategy to free up WM capacity by removing or overwriting information that is no longer behaviourally relevant. On the other hand, preserving such information might carry the adaptive advantage that transformations could be re-applied on their originals, or even adjusted if the need arose. To arbitrate between these alternatives was the purpose of the present study.

There is at least some prior evidence from fMRI to suggest that once an item is transformed (e.g., mentally rotated) it replaces its original altogether (Christophel et al., 2015). Similarly, MEG recordings taken during mental rotation suggest that a gradual change occurs from the original representation into a rotated one (Trübutschek et al., 2019). By contrast, a recent study by Iamshchinina and colleagues (2021) showed that perceptual representations in primary visual cortex lasted throughout the mental rotation interval. It was nevertheless not the primary aim of these studies to track the possible simultaneous maintenance of perceived and transformed items in WM, so that the evidence is not clear-cut: In the studies by Christophel and colleagues (2015) and Trübutschek and colleagues (2019), the relationship between original and rotated items was constant across trials, which complicates any attempt to assess the presence of either item independently. For instance, between two representations rotated either 0 or 120 degrees, neural decoding of intermediate values cannot be definitely attributed to either representation. Conversely, in the study by Iamshchinina and colleagues (2021) there was only a single item being manipulated, so the original item may only have been found to persist as a consequence of presentation history, rather than it being part of WM proper.

A compounding, more general challenge to studying transformations in WM is the recent discovery that neural activity alone may not reflect the full breadth of WM operations. WM maintenance may not rely on unbroken chains of ongoing neural activity (LaRocque et al., 2013), as was previously thought (e.g., Curtis & D'Esposito, 2003; Kamiński et al., 2017). Rather, it may utilize activity-silent or quiescent brain states that are mediated by short-term changes in functional connectivity (Mongillo et al., 2008). Furthermore, it has been suggested that there may be different functional states in WM relating to passive maintenance and active, attentional updating (Olivers et al., 2011; Trübutschek et al., 2019), and both these states are presumably traversed when stored information is transformed. Importantly, although active attentional updating of WM is associated with easily measurable neural activity, this is not necessarily the case for passive maintenance (Kamiński & Rutishauser, 2020; but see also Muhle-Karbe, Myers, & Stokes, 2021; Stokes, Muhle-Karbe, & Myers, 2020). It is thus conceivable that traditional, activity-based measurement approaches miss part of the picture.

In our study, we sought to overcome these issues and decisively assess and compare WM states before and after mental transformation. First, we implemented a design in which original and transformed items were sufficiently independent from each other, so that they could be examined individually. Second, although activity-quiescent states remain intrinsically difficult to measure non-invasively, it has recently been confirmed

that functional connectivity can be illuminated by driving a standardized impulse signal through the network, as the response to that stimulation will partially reflect the momentary state of the network, independent from the focus of attention (Buonomano & Maass, 2009; Stokes, 2015; Wolff et al., 2015; Wolff, et al., 2017; Rose et al., 2016). Therefore, we implemented a perturb-and-measure approach (Wolff et al., 2015; 2017) by presenting task-irrelevant impulse stimuli during WM maintenance before and after mental rotation of randomly oriented gratings. Based on the research to date, two contrasting hypotheses were formulated: It may be that both the original and the transformed WM items are maintained and can be similarly decoded through impulse perturbation. Alternatively, one may hypothesize that WM only stores task-relevant information, so that once an item is transformed, only the resultant representation is kept, and the original item that is no longer relevant is rapidly discarded.

To preview the principal findings, before rotation, only the cued, relevant WM item could be successfully predicted from impulse-evoked EEG activity, unlike the uncued item, which seemed to have been rapidly purged from the WM system, replicating earlier findings (Wolff et al., 2017; 2020a; 2020b). Prior to the response probe, from the second impulse stimulus that followed the rotation instruction, the imagined rotation product could be decoded. Intriguingly, the original orientations could also still be decoded. The continued presence of these obsolete originals in the WM network suggests that transformations in WM rely on relatively elaborate 'double' encoding, which was associated with faster probe response times. Thus, the brain may prioritize representational precision and behavioural flexibility over storage capacity in spite of its scarcity.

Method

Participants

Thirty students of the University of Groningen (17 female, M_{Age} = 21.3; $Range_{Age}$ = 18 – 31; all right-handed) volunteered to participate in the study in exchange for course credits or monetary reward (€8 per hour). The participants were selected from a larger group by means of a pre-screening procedure. Pre-screening consisted of a behavioural task, lasting approximately one hour, which was otherwise identical to the EEG session. The selection criteria were pre-determined and communicated to the participants; at least 70% task accuracy and a response time of less than 700 ms on average. The sample size was based on earlier studies with a similar design (e.g., Wolff et al., 2017). The study was conducted in accordance with the Declaration of Helsinki (2008), and was approved by the Ethical Committee of the Behavioural and Social Sciences Faculty

of the University of Groningen (Study ID = 18029-SP). All participants provided written informed consent before taking part.

Apparatus and Stimuli

Participants were seated in a fully lit testing chamber at a viewing distance of approximately 60 cm from the screen, a 17" Samsung 797DF CRT monitor, set at a refresh rate of 100 Hz and a resolution of 1024 by 798 pixels. The stimuli were created and presented with the freely available Psychtoolbox 3 extension for Matlab (Brainard, 1997; Kleiner et al., 2007). A custom two-button response box connected via a USB interface was used to collect behavioural responses.

As shown in Figure 1, the background remained grey throughout the experiment (RGB = 128, 128, 128), and a black fixation dot with a white outline (0.25° of visual angle) was present in the centre of the screen throughout the trials. The memory items were circles containing sine-wave gratings with 6 different orientations ranging from 15° to 165° with an interval of 30°. At the beginning of each trial, two orientation gratings were presented, each centred at 6.5° of visual angle (at 60 cm viewing distance) from the centre of the screen on the horizontal axis. All gratings were presented at 20% contrast, each again with a diameter of 6.5° of visual angle. The spatial frequency of the gratings was set to 0.65 cycles per degree, while their phase was randomized within and across trials. The cue stimuli were arrows of 0.5° of visual angle ">>", pointing in the direction of the stimulus that was to be cued, presented in black Arial font 0.5° of visual angle above the centre of the screen. The direction of rotation was cued by an arrow of 1° of visual angle pointing clockwise or counter-clockwise (which was absent if no rotation was required), fitted to the middle of a 90° wide radian cut out from a circle with a diameter of 8° of visual angle. The angle of rotation was presented numerically (0, 30 or 60°) in bold black Arial font at a visual angle of 0.5° above the centre of the screen. The complete rotation instruction including the direction and the angle covered 2.75° of visual angle on the screen. The impulse stimulus consisted of three partially overlapping circles each with a diameter of 9.75°. The centre-to-centre distance of each circle was 6.5°. The probe stimulus was presented at the centre of the screen and was always identical to one of the 6 sine-wave gratings used as memory items.

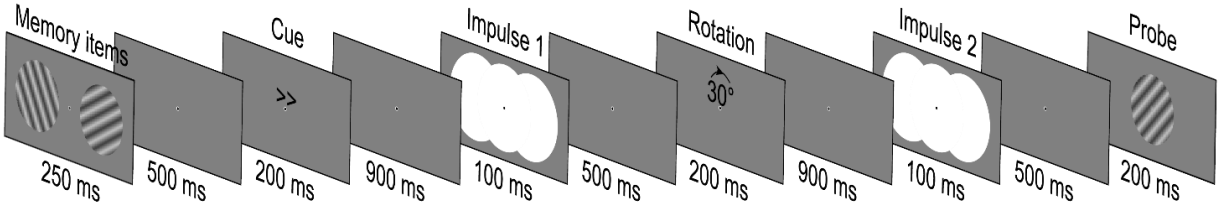

Figure 1. The design of a single experimental trial. Participants maintained one of two memory items in memory, indicated by a retro-cue. Subsequently, this item was rotated (0, 30, or 60°) and eventually compared to the probe stimulus. Impulse stimuli were task-irrelevant and served to elicit an EEG response by perturbing the representational network.

Procedure

Each participant first completed 48 practice trials that were identical to the experimental trials prior to EEG recording. During practice, all participants were instructed and trained to keep their gaze on the fixation dot at all times, and to blink only during or after response. Fast and accurate responses were encouraged. The whole experiment consisted of 1440 trials, spread across 4 consecutive sessions separated by breaks with a duration that was determined by the participants themselves. In each session, participants completed 24 blocks containing 15 trials each. Within these blocks, trials continued without interruption. On average, each participant took 4 hours to complete the task, including breaks.

Each trial began with the presentation of a fixation dot for 700 ms, followed by the presentation of the memory array, which consisted of two orientation gratings that appeared on both sides of the visual field for 250 ms. The orientations of the gratings were randomly selected without replacement from a uniform distribution, such that each of the six orientations was presented the same number of times throughout the experiment. Next, a blank screen with only the fixation dot was presented following stimulus offset for 500 ms. A retro-cue was then presented for 200 ms, indicating which of the two orientation gratings had to be retained in memory. The direction of the cue (left or right) was randomized, but evenly distributed across all conditions. After a delay of 900 ms the impulse signal was on display for 100 ms, and was followed by a blank delay interval of 500 ms. Then the rotation instruction was displayed for 200 ms, followed by another 900 ms delay. Like the cues, the rotation angles (0, 30 or 60°, clockwise and counter-clockwise) were randomized but evenly distributed across conditions. The second impulse was then presented for 100 ms, followed by a delay of 500 ms. Lastly, the probe was on display for 200 ms. Participants were asked to judge, as quickly as possible, whether the probe was the same as the relevant WM item, which was either the rotation product in rotation

trials, or the original cued item on trials that did not require rotation (i.e., 0°). The probe was randomized but matched the relevant WM item in 50% of cases, while in the other 50% of cases, the probe was sampled randomly from the other possible orientations. The participants could report their answer by pressing one of the two buttons on the response box. After each response, feedback was given by a smiley that was presented for 200 ms, where a happy face indicated that the response was correct. The keys on the response box were counterbalanced across participants.

EEG Acquisition and Pre-processing

The EEG signal was recorded at a sample rate of 1000 Hz from 62 Ag/AgCl sintered electrodes deployed with a 10-20 international layout. The average of all electrodes was used as the reference during recording, with a ground electrode placed on the sternum. The data were recorded with BrainVision Recorder software, and a TMSI Refa 8-64/72 amplifier. Eye movements were tracked via bipolar electrooculography with vertical electrodes above and below the left eye and two horizontal electrodes on ipsilateral sides of both eyes. The impedance at all electrodes was kept below 10 kΩ.

Offline, the data were re-referenced to the mastoids, downsampled to 500 Hz and bandpass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB (Delorme & Makeig, 2004). The data were epoched relative to Impulse 1 and Impulse 2 (-150 ms to 600 ms). Epochs with excessive variance, voltage drifts, or muscle or eye movement artefacts were identified visually and removed from all subsequent analyses. In total, 11.3% of epochs were rejected.

Finally, the data were reformatted to best capture the time-locked neural dynamics evoked neural responses to the impulse stimuli. The approach was the same as used previously (Wolff et al., 2020a, 2020b). For the full *spatiotemporal* analyses, the voltage traces from the 17 posterior channels of interest (P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2) were extracted relative to the impulse onsets from 100 to 400 ms. Then, the mean voltage within each time window was removed for each trial and channel separately, which removes drift (as in conventional baselining) and isolates the dynamic, stimulus-evoked neural response. The voltage traces were then down-sampled to 100 Hz, and combined with the channel dimension, resulting in 510 data features (30 time-points x 17 channels) per trial for each impulse.

For the *time-course* analyses of neural dynamics, the same posterior channels as above were included. Here we used a 100 ms sliding window from -50 ms to 550 ms relative to impulse onsets. Similar as above, the data were down sampled to 100 Hz within each time-window, and the mean voltage was removed. The resulting

10 time-points were then combined with the channels, resulting in 170 data features. The analyses (described below) were done on each time-point that the time-window was centred on, separately. The time-courses were smoothed with a gaussian smoothing kernel (*SD* = 16 ms).

For the *searchlight* analyses, the whole spatiotemporal window was used, but the analyses were repeated iteratively for each electrode and its closest two neighbours (thus 30 time-points x 3 channels), across all 62 electrodes.

EEG Analyses

Linear discriminant contrast

We used cross-validated representational similarity analysis (RSA) with Mahalanobis distance (Nili et al., 2014), also termed linear discriminant contrast (LDC; Walther et al., 2016), to investigate the contributions of different task-related models to the neural codes evoked by Impulse 1 (before rotation instructions) and Impulse 2 (after rotation instructions). We used an 8-fold cross-validation approach to compute the squared Mahalanobis distance (MD^2) between condition pairs of all condition combinations of interest at Impulse 1 and Impulse 2 (see further below), using the following formula:

$$MD^2 = (P_A - P_B)_{train} \times \sum_{train}^{+1} \times (P_A - P_B)_{test}^{T}$$

The data was randomly split into 8 folds using stratified sampling that ensures a roughly equal number of trials in each fold. The *test* data consisted of a single left out fold and the *train* data consisted of the remaining 7 folds, which was also used to estimate the noise covariance matrix ($\sum$). The number of trials of each condition was equalised through random subsampling within the test fold and train folds. P_A and P_B were the trial-averaged patterns of conditions A and B of the subsampled trials of the *test* and *train* data. The noise covariance matrix was estimated from the subsampled *train* data by subtracting the trial-averaged activity patterns of each condition from all trials of the corresponding condition. The covariance matrix was calculated using a shrinkage estimator (Ledoit & Wolf, 2004) on the resulting condition-mean centred trial by feature matrix. The pseudoinverse of the covariance matrix ($\sum^{+1}$) was then used to compute the MD^2 between all possible condition differences resulting in an n-conditions by n-conditions representational dissimilarity matrix (RDM). This procedure was repeated for all test and train fold combinations. The 8-fold partitioning and the subsampling within test and train data was randomised. To ensure stable results, the above procedure was repeated a total of 100 times for each subject, with random folds and random subsampling each time. The resulting data RDMs

were averaged across repetitions and folds, resulting in a single RDM per subject and impulse. These RDMs were z-scored and then regressed against specific task models of interest.

The conditions that were considered and the task models were different for each impulse. At Impulse 1 we considered the cued item (6 orientations), the uncued item (6 orientations), and the cued location (left or right). The pairwise MD^2s between all 72 unique condition combinations resulted in a 72 × 72 data RDM for each subject. The task models we tested at Impulse 1 comprised the cued location model (left or right), parametric coding models for the cued and uncued memory item (absolute circular distance between cued items, and between uncued items), as well as a parametric coding model of the generalization between cued and uncued items (absolute circular distance between cued and uncued). All models (except the cued location model) were subdivided and tested separately within the same and across different cueing conditions to test for (cued) location specificity. All task model RDMs were z-scored and regressed simultaneously against the z-scored data RMD of each subject using multiple regression.

At Impulse 2 we considered the cued item (6 orientations), the rotation instructions (5 conditions), and the cued location (left or right), which resulted in a 60 × 60 RDM for each subject. We first tested the cued location and the rotation instructions task models simultaneously using multiple regression. The rotation instructions models consisted of 5 separate models (one for each condition), each testing same vs. different instructions. The resulting model fits/beta values of the rotation models were averaged for plotting and statistical analyses. We furthermore tested parametric coding models of the cued item, the rotated item, and the generalization between them. Since the cued and the rotated item models are statistically related, the models were fit on the residual data RDMs of the other. This meant that the cued item model was fit to the residual variance of the data RDM that was not accounted for by the rotated item model data, and vice versa. The cued-rotated generalization model was fit to the residual variance of the data RDM that was not accounted for by either the cued item model or the rotated item model. As in the case of Impulse 1, all models (except for the cued location model) were subdivided and tested separately for within and across cued location conditions. All model and data RDMs, as well as the residual variance RDMs, were z-scored before regression.

Correlation of trial-wise decoding strength

We tested whether the trial-wise strength in neural activation patterns related to the cued item correlated with the corresponding neural activation patterns of the rotated item. We used the same decoding approach as in Wolff et al. (2020b) to obtain trial-wise decoding strengths for the cued and the rotated item at Impulse 2.

This entailed using an 8-fold cross-validation decoding approach using Mahalanobis distance. The data was randomly split into 8 folds using stratified sampling (ensuring a roughly equal number of trials for each orientation in each fold). The covariance matrix (with shrinkage estimator; (Ledoit & Wolf, 2004)) was computed using the 7 folds of the train data. The number of trials of each orientation within train data was then equalised via random subsampling. The averaged patterns of each orientation of the train data were convolved with a half cosine basis set raised to the 5^{th} power to pool information across similar orientations (Myers et al., 2015). The Mahalanobis distance between each test trial and averaged orientation patterns of the train data were then computed. This procedure was repeated for all test and train fold combinations. The resulting 6 distances for each trial were summarized into a single "decoding strength" value by computing the cosine vector mean of the absolute circular distance between test trial's orientation and averaged orientation patterns of the train trials. To get reliable estimates, the above procedure was repeated 100 times (with random folds and subsamples), resulting in 100 decoding strength values for each trial and subject, which were subsequently averaged. The cued and the rotated items were decoded separately at Impulse 2. To ensure that the decoding results were not statistically related to one another, we excluded trials in which participants were instructed not to rotate the cued item. The Pearson correlation between the trial-wise decoding strengths of the cued and the rotated item was computed for each subject separately, Fisher's z-transformed, and statistically tested against 0 (see statistical testing section).

Decoding generalization before and after mental transformation

We were interested if the neural pattern related to the cued item before mental rotation (at Impulse 1) generalized to the neural pattern after mental rotation (at Impulse 2). To test this, we trained a classifier using the same approach as described in the previous section on the neural pattern in all the trials of the cued item at Impulse 1 and tested it only on the rotation trials at Impulse 2 (i.e., excluding the trials in which subjects were instructed not to rotate the item), using 8-fold cross-validation, random subsampling within test and train folds, and 100 repetitions. Given the cue-specific coding scheme observed at Impulse 1, the classifier was trained and tested within each cued location condition separately and averaged afterwards. The test trials at Impulse 2 were either labelled with the cued item, to test temporal generalization of the cued item over time (Wolff et al., 2020b), and after mental rotation, or labelled with the rotated item, to test whether the same neural pattern that codes for the cued item was re-used to code for the rotated item.

Relationship between the neural and behavioural data

We were interested if the quality of WM content predicted behavioural performance (probe response times and accuracy) on a trial-by-trial basis within subjects. We used the trial-wise decoding strengths of the cued/original and the rotated item at Impulse 2 (excluding the "0 rotation" trials) and tested if they predicted trial-wise fluctuations in accuracy and response times. First, we regressed the decoding strengths against each other to obtain the residuals of each, to ensure that the decoding strengths of the cued/original and the rotated item were uncorrelated and explained unique aspects of the behavioural measure in question. We then used the residuals of the cued/original and the rotated item decoding strengths as regressors to predict behavioural accuracy (logistic regression) and log-transformed response times (linear regression) within each subject. The resulting regression weights were then tested for significance in the expected direction of facilitation.

Simulations of neural patterns

We simulated several plausible effects to compare the pattern of results of simulated data with the results of the actual data. Activity patterns were simulated by randomly drawing 20 values from the standard normal distribution two times. To simulate a parametric pattern for the circular memory items, one of the two patterns was convolved with the sine of the memory item in question, and the other with the cosine of the same memory item of that trial, before adding both signals together. Trial-wise noise was added to the signal by randomly drawing 20 values from the standard normal distribution and then multiplying it with a random value drawn from a normal distribution with mu = 10 and SD = 3. This was done separately for each trial to simulate trial-wise fluctuations in neural noise levels. The trial-wise noise patterns were added to the signal patterns comprising the overall simulated neural signal.

We simulated and analysed the following scenarios that we thought could be expected in this task:

A. Rotated only: A new activity pattern for the rotated item is present in the signal, while the code of the original item completely disappears.

B. Partial rotation: An activity pattern is present in every trial that represents an item that is only partially rotated (half-way). Neither the original, nor the fully rotated item are represented in the signal. This simulates the possibility that subjects failed to fully transform the item, while dropping the original item from memory.

C. Same coding schemes: The original and the rotated item use the same signal pattern and both are present in the data.

    i. The patterns of both items are simultaneously present in every trial.

ii. The pattern of only one of the items is present in a given trial, simulating the possibility that subjects may have only sometimes followed instructions and mentally rotated the item.

D. Unique coding schemes: The original and the rotated items use unique and independent coding schemes and are both present in the data.

i. Both patterns present in every trial.

ii. Only the pattern of one of the items is present in a given trial.

For simplicity we did not consider cue-specific effects in the simulations (i.e., whether or not the coding schemes generalize across cued-locations or not). We ran each scenario 100 times, with randomised signal and noise patterns each time (simulating 100 subjects). The simulated data was subsequently analysed in the same way as the real data of Impulse 2 (LDC and correlation of trial-wise decoding strengths).

Statistical significance testing

We used non-parametric tests to assess statistical significance in all cases. We tested for statistical significance of the neural analyses results by randomly shuffling the conditions in question 1,000 times and using the resulting null-distribution to conduct a *t*-test. In the case of the LDC and trial-wise decoding analyses this meant that the analyses were re-run with randomised condition labels resulting in 1,000 "null" model fits/decoding values per subject. These were transformed into null distributions of *t* values by computing the t-value across subjects for each of the 1,000 values, which was then used for a *t* test against 0 of the actual model fit/decoding value. For the time-course analyses, a cluster-based permutation test (1,000 permutations) was used to correct for multiple comparisons over time with a cluster-forming threshold of $p < 0.05$. All tests involving the statistical significance of the neural analyses (decoding strengths or model fits) were two-sided.

To test for statistical significance of the correlation between trial-wise decoding values of the cued and the rotated item, the trial-wise decoding values were randomly shuffled 10,000 times, each time obtaining the Fisher's z-transformed correlation value. These were transformed into a t-value distribution which was then used to perform a t-test (two-sided).

The relationship between neural and behavioural data was tested for statistical significance by shuffling the trials 10.000 times and repeating the regression analyses each time. The resulting null-distribution of the beta values was used to perform a t-test. Since we expected higher trial-wise decoding-strengths to result in

better performance (higher accuracy, faster response time), the tests involving the behavioural relationship of the neural data were one-sided.

We also computed Bayes Factors (BF) to complement p-values. We used the Bayesian implementation of the non-parametric Wilcoxon signed-rank test with 10.000 samples and the Cauchy prior with the default scale of 0.707, as implemented in JASP (JASP Team, 2018).

Data and code availability

All data and MATLAB code used to generate the results and figures of this manuscript will be publicly available from the Open Science Framework at https://osf.io/3hdpc upon peer-reviewed publication. Updated scripts and functions in both MATLAB and Python related to the manuscript will also be available at https://github.com/mijowolff/veridical-and-transformed-representations-in-wm.

Results

Behavioural Results

Behavioural measures consisted of response accuracy (% correct), reflecting the comparison of the probe and the memory item, and reaction time (RT), reflecting the time from probe onset until the button press. Mean accuracies and median RTs were analysed as a function of rotation condition (-60, -30, 0, 30, 60). Behavioural analyses were conducted in the freely available JASP program (JASP Team, 2018).

The results were typical for rotation tasks (e.g., Wexler et al., 1998; Searle & Hamm, 2017), and are shown in Figure 2. The accuracy in the task was highest when no rotation took place and declined as the rotation magnitude increased, regardless of direction. Repeated measures ANOVA confirmed that accuracy in at least one condition differed from the others, $F(4, 116) = 61.454$, $p < 0.001$, $\eta^2 = 0.679$, $BF_{10} > 1000$. Post hoc paired *t*-tests revealed that performance differed between all rotation conditions, unless rotation magnitude was identical ($t_{-60, -30}(29) = -5.127$, $p < 0.001$, $BF_{10} = 588.67$; $t_{-60, 0}(29) = -12.613$, $p < 0.001$, $BF_{10} > 1000$; $t_{-60, 30}(29) = -5.589$, $p < 0.001$, $BF_{10} > 1000$; $t_{-60, 60}(29) = 1{,}444$, $p = 1$, $BF_{10} = 0.476$; $t_{-30,0}(29) = -7.487$, $p < 0.001$, $BF_{10} > 1000$; $t_{-30, 30}(29) = -0.463$, $p = 1$, $BF_{10} = 0.217$; $t_{-30, 60}(29) = 6.570$, $p < 0.001$, $BF_{10} > 1000$; $t_{0, 30}(29) = 7.024$, $p < 0.001$, $BF_{10} > 1000$; $t_{0, 60}(29) = 14.057$, $p < 0.001$, $BF_{10} > 1000$; $t_{30, 60}(29) = 7.033$, $p < 0.001$, $BF_{10} > 1000$; two-sided, Bonferroni-corrected; Fig. 2A). The distribution of accuracies across rotation conditions clearly showed that greater rotation magnitudes negatively influenced recall accuracy, though accuracy remained well above chance in all cases.

Rotation magnitude similarly influenced RT, $F(2.601, 75.429) = 16.989$, $p < 0.001$, $\eta^2 = 0.369$, $BF_{10} > 1000$ (with Greenhouse-Geisser correction), with an increase in RT as rotation magnitude increased ($t_{-60, -30}(29) = -2.376$, $p = 0.191$, $BF_{10} = 10.993$; $t_{-60, 0}(29) = 7.282$, $p < 0.001$, $BF_{10} > 1000$; $t_{-60, 30}(29) = 3.119$, $p = 0.023$, $BF_{10} = 26.003$; $t_{-60, 60}(29) = 0.367$, $p = 1$, $BF_{10} = 0.212$; $t_{-30, 0}(29) = 4.905$, $p < 0.001$, $BF_{10} = 438.915$; $t_{-30, 30}(29) = 0.743$, $p = 1$, $BF_{10} = 0.353$; $t_{-30, 60}(29) = -2.009$, $p = 0.468$, $BF_{10} = 1.013$; $t_{0, 30}(29) = -4.163$, $p < 0.001$, $BF_{10} = 80.012$; $t_{0, 60}(29) = -6.915$, $p < 0.001$, $BF_{10} > 1000$; $t_{30, 60}(29) = -2.752$, $p = 0.069$, $BF_{10} = 14.801$; two-sided, Bonferroni-corrected; Fig. 2B).

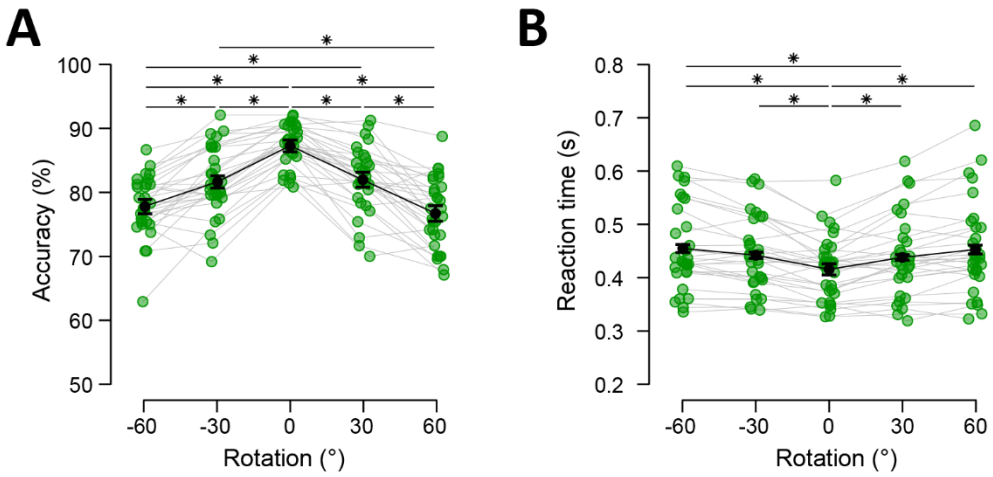

Figure 2. Behavioural performance as a function of rotation. **(A)** Mean accuracy in percent correct. **(B)** Reaction times (means of medians) in seconds. Green dots represent individual data points, black dots reflect averages, and error bars represent within-subject 95% confidence intervals (Morey, 2008). Significant pairwise differences are indicated with asterisks (* $p < 0.05$, two-sided, Bonferroni-corrected).

EEG results

LDC at Impulse 1; before transformation/after cue

The RDM of Impulse 1, depicting the MD^2 between all cued, uncued, and cued location combinations, is shown in Figure 3A. The cued location model showed a significant effect (spatiotemporal: $p < 0.01$, $BF_{10} > 1000$, two-sided; time-course: $p < 0.01$, -55 ms to 550 ms, cluster-corrected, two-sided, Fig. 3B), likely driven by a shift in spatial attention toward the cued item in WM (e.g., Wolff et al., 2017). The cued item coding model was statistically significant within cued location (spatiotemporal: $p = 0.02$, $BF_{10} = 11.117$, two sided; time-

course: $p < 0.01$, 92 ms to 300 ms, cluster-corrected, two-sided), cluster-corrected, two-sided, but not across (spatiotemporal: $p = 0.712$, $BF_{10} = 0.219$, two-sided), and the difference between within and across cued location conditions of the cued item coding model was statistically significant (spatiotemporal: $p = 0.042$, two-sided; time-course: $p < 0.01$, 140 ms to 292 ms , cluster-corrected, two-sided; Fig. 3C, left), though the Bayesian evidence for or against an effect was ambiguous ($BF_{10} = 1.671$). The uncued item coding models showed no significant effects and Bayesian evidence against effects (spatiotemporal, within cued location: $p = 0.354$, $BF_{10} = 0.276$; across cued location: $p = 0.426$, $BF_{10} = 0.242$; difference: $p = 0.242$, $BF_{10} = 0.394$, two-sided; Fig. 3C, middle). These results replicate previous findings in EEG of cue-specific neural coding of the cued WM item when it was laterally presented, and no detectable trace of the uncued item (Wolff et al., 2017, 2020a). None of the generalization models between the cued and uncued item reached the statistical significance threshold (spatiotemporal, within cued location: $p = 0.536$, $BF_{10} = 0.238$; across cued location: $p = 0.058$, $BF_{10} = 1.349$; difference: $p = 0.368$, $BF_{10} = 0.257$, two-sided; Fig. 3C, right). However, the trend of negative generalization between cued and uncued items across cued locations (meaning that the cued and the uncued item were presented at the same location in different trials) could suggest a possible suppression of the uncued item (van Loon et al., 2018; Wan et al., 2020).

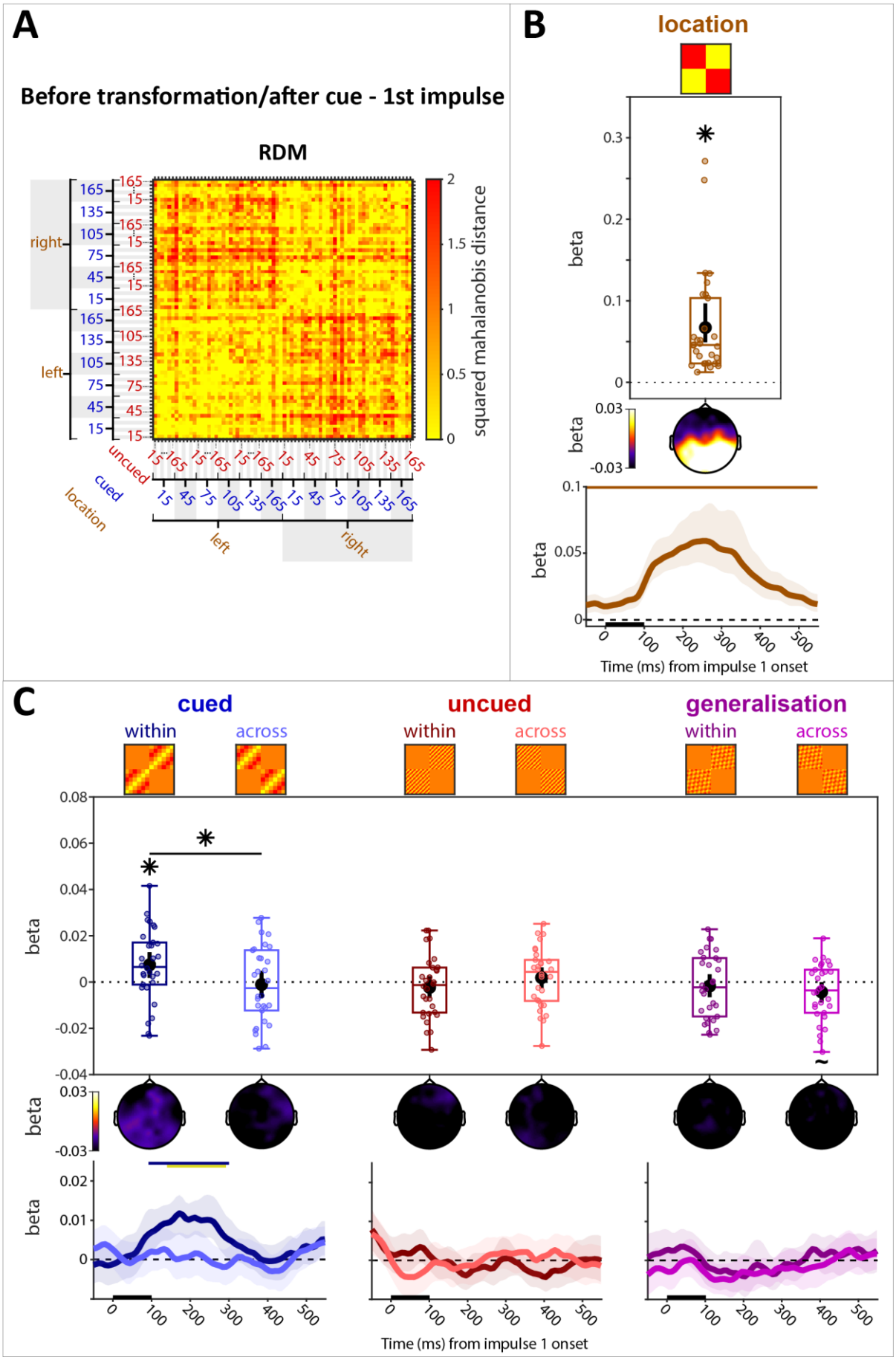

Figure 3. LDC before transformation & after the retro-cue at Impulse 1. **(A)** Average RDM. **(B)** Model fit (beta) of cued location condition. Top: Spatiotemporal; Middle: Searchlight; Bottom: Time-resolved. **(C)** Model fits of WM items (cued, uncued) and their generalization. Models are separated into within-cued location condition (within) and across-cued location condition (across). Same convention as (B). Error bars (spatiotemporal) and error shadings (time-resolved) are 95% C.I. Statistically significant ($p < 0.05$, two-sided) and marginally significant ($p < 0.1$, two-sided) model fits and differences of model fits between within and across locations are marked with (*) and (~), respectively. Statistically significant time-course clusters ($p < 0.05$, two-sided, cluster-corrected) are marked with coloured bars at the top (corresponding colour for significant model fits, yellow for differences between within and across locations).

Simulations of predicted neural effects after mental transformation in WM

We simulated plausible changes in the neural coding schemes of WM content after mental transformation (see methods). The spatiotemporal analyses for the simulated data and the data at Impulse 2 (see below) were the same. The pattern of results for each scenario are shown in Fig. 4. Note that almost every considered scenario resulted in a qualitatively unique pattern of results. The exception is the partial rotation of the original item (Fig. 4B) and the simultaneous representation of both the original and the rotated item in every trial, both using the same coding schemes (Fig. 4Ci), which cannot be distinguished with the analyses we employ. As seen in the next section below, the pattern of results of the actual data at Impulse 2 after transformation best resembled the scenario depicted in Fig. 4Di: the simultaneous maintenance of both items in every trial (original and rotated), with unique coding schemes.

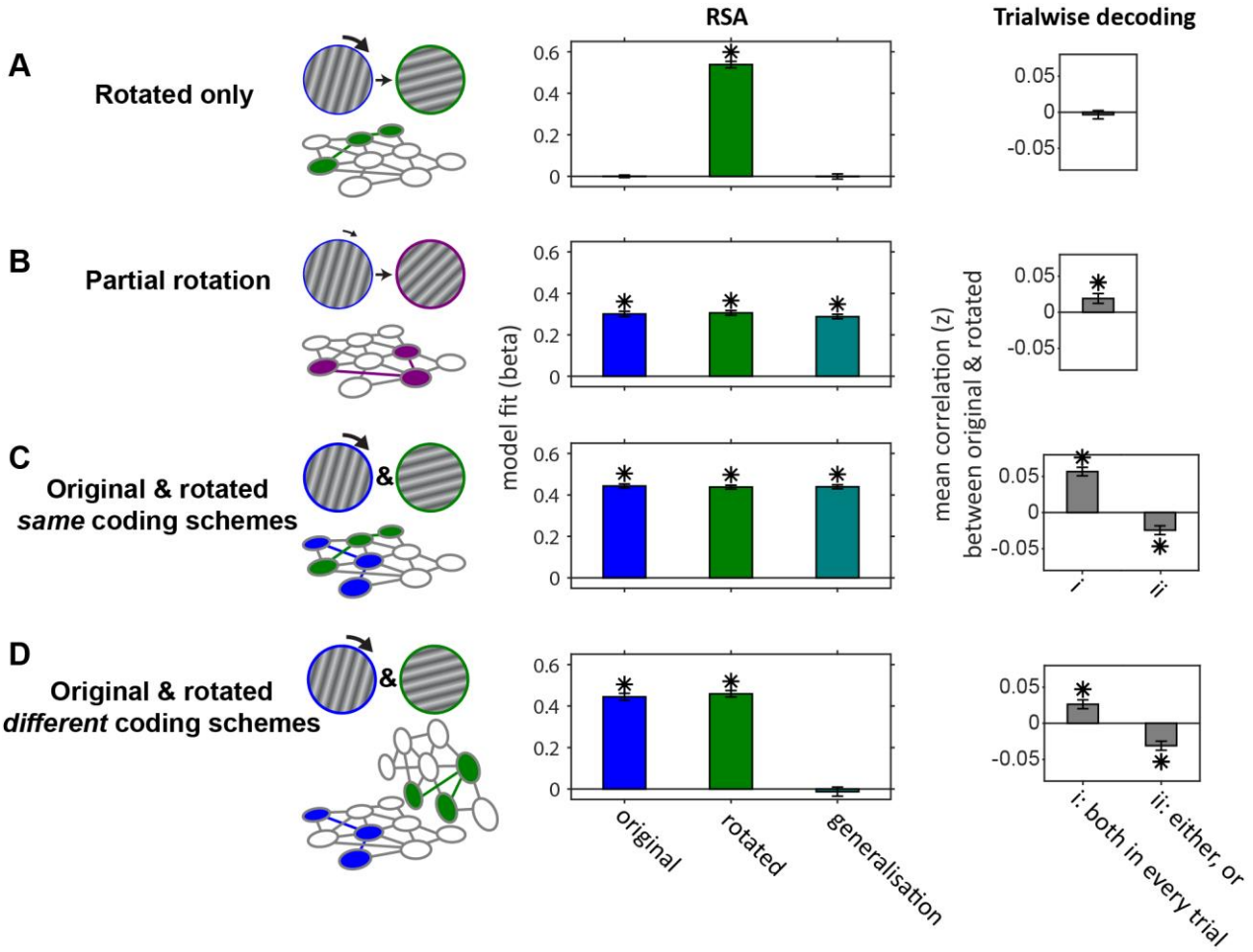

Figure 4. Simulation results (N = 100) of plausible maintenance scenarios of the original cued item and the rotated item after mental transformation. The left panel shows the schematics of each scenario, the middle panel the model fits of the original item, the rotated item, and the generalisation between them, and the right panel the mean correlation (Fisher's z) between the trial-wise decoding strengths of the original and the rotated item. **(A)** Rotated only: Only the fully rotated item is present in the WM network, and the original item is removed completely. **(B)** Partial rotation: The original item is only halfway rotated in each trial and the original item is no longer present. Note that the analyses would incorrectly imply that both the original and the fully rotated items are simultaneously present in the network using the same coding schemes (i.e., generalisation between them) **(C)** Original & rotated, *same* coding schemes: The original and the (fully) rotated items are both coded using the same coding schemes. i) both are present in every trial (note the same pattern of results as B). ii) either one, or the other is represented in a given trial. **(D)** Original & rotated, *different* coding schemes: The original and the rotated item are represented in unique coding schemes, i) simultaneously on every trial, or ii) only one of them is randomly represented on every trial). Error bars are 95% C.I. Asterisks denote statistically significant ($p < 0.05$, two-sided) results of the simulated data.

LDC at Impulse 2; after transformation

The RDM of Impulse 2, depicting the MD^2 between all cued item, rotation instructions, and cued location condition combinations, is shown in Figure 5A. The cued location model was significant (spatiotemporal: $p < 0.01$, $BF_{10} > 1000$, two-sided; time-course: $p < 0.01$, 68 ms to 452 ms, cluster-corrected, two-sided), as were both rotation instruction coding models (spatiotemporal, within cued location: $p < 0.01$, $BF_{10} > 1000$; time-course: -55 ms to 20 ms, 52 ms to 550 ms, $p < 0.01$, cluster-corrected, two-sided; spatiotemporal, across cued location: $p < 0.01$, $BF_{10} > 1000$, two-sided; time-course: 36 ms to 550 ms, $p < 0.01$, cluster-corrected, two-sided), with no difference between them (spatiotemporal: $p = 0.478$, $BF_{10} = 0.326$, two-sided; Fig. 5B). Even though the original orientation of the cued item was not behaviourally relevant at Impulse 2 anymore, both cued item coding models were statistically significant for both within cued location (spatiotemporal: $p < 0.01$, $BF_{10} = 11.126$, two-sided; time-course: 236 ms to 428 ms, $p < 0.01$, cluster-corrected, two-sided) and across cued location (spatiotemporal: $p = 0.044$, $BF_{10} = 1.454$, two-sided; time-course: 108 ms to 148 ms, $p = 0.012$, 268 ms to 308 ms, $p = 0.044$, cluster-corrected, two-sided Fig. 5C, left), though Bayesian evidence was ambiguous for the latter. This suggests that the neural code of the cued item was less spatially specific at Impulse 2 after transformation, in contrast to the code at Impulse 1 before transformation. However, while there was no difference between them in the spatiotemporal analysis ($p = 0.448$, $BF_{10} = 0.248$), the time-course analysis showed a significant difference between within- and across-models of the cued item (332 ms to 380 ms, $p < 0.01$, cluster-corrected, two-sided), suggesting some spatial specificity nonetheless.

The rotated item coding models were both statistically significant (within cued location, spatiotemporal: $p < 0.01$, $BF_{10} = 53.742$; time-course: 116 ms to 428 ms, $p < 0.01$, cluster-corrected, two-sided; across cued location, spatiotemporal: $p < 0.01$, $BF_{10} = 251.767$, two-sided; time-course: 100 ms to 420 ms, cluster-corrected, two-sided), and were not different from each other (spatiotemporal: $p = 0.68$, $BF_{10} = 0.196$, two-sided; Fig 5C, middle). The generalization coding models between the cued and the rotated item were not significant with Bayesian evidence for no effect (spatiotemporal: within cued location: $p = 0.892$, $BF_{10} = 0.198$; across cued location: $p = 0.704$, $BF_{10} = 0.210$; difference: $p = 0.912$, $BF_{10} = 0.213$, two-sided; Fig. 5C, right), though the time-course analysis revealed a significant cluster in the difference (time-course: 68 ms 124 ms, $p < 0.01$, cluster-corrected, two-sided).

Overall, these results provide evidence for the presence of both the cued, and the mentally rotated items in the WM network, which are both coded using distinct coding schemes that do not cross-generalise, in line with the simulation results of scenario 4D above.

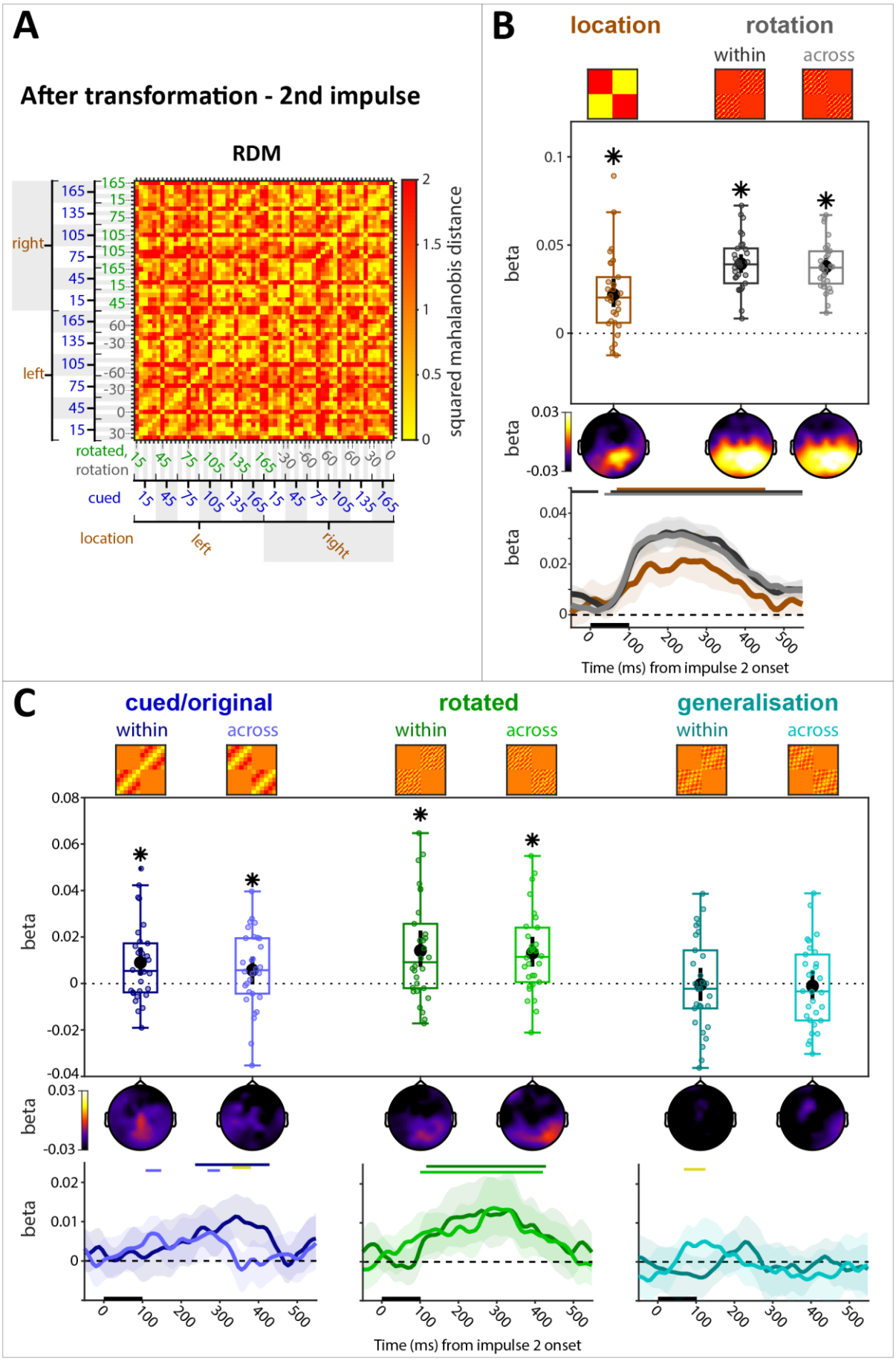

Figure 5. LDC after transformation at Impulse 2. **(A)** Average RDM. **(B)** Model fit (beta) of cued location condition and rotation condition (separately for within cued location and across cued location). Top: Spatiotemporal; Middle: Searchlight; Bottom: Time-course. **(C)** Model fits of the original cued item, the rotated item, and their generalization. Models are separated into within cued location condition (within) and across cued location condition (across). Same convention as (B). Error bars (spatiotemporal) and error shadings (time-course) are 95% C.I. Statistically significant ($p < 0.05$, two-sided) model fits and differences of model fits between within and across locations are marked with asterisks (*) for spatiotemporal. Statistically significant clusters ($p < 0.05$, two-sided, cluster-corrected) are marked with coloured bars at the top (corresponding colour for significant model fits, yellow for differences between within and across locations) for the time-course.

Correlation between trial-wise decoding strengths of the original and the rotated item after transformation

The results presented above are based on trial-averaged data and do not rule out the possibility that although evidence for both the cued/original and the rotated item was found, subjects only maintained one of the two in individual trials. We tested this by correlating the trial-wise decoding strengths of the cued and the rotated item (excluding no-rotation trials). A negative correlation would be evidence that only one of them was coded in any one trial, while the simultaneous maintenance of both items in each trial would result in a positive correlation due to variable noise levels of each trial. While it did not reach statistical significance, there was a trend of a positive correlation between the decoding strengths of the cued and the rotated item ($p = 0.084$, $BF_{10} = 1.087$, two-sided; Fig. 6). Explicitly testing whether the correlation was negative provided strong evidence against it ($BF_{10} = 0.074$, one-sided). This provides evidence that there was no trade-off in the neural strengths of the items across trials and suggests that both items may have been present simultaneously in the neural data in at least some trials, which fits with scenario 4Di above.

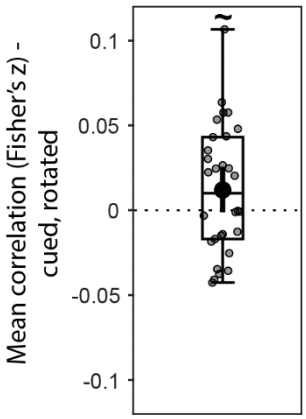

Figure 6. Mean correlation (Fisher's z transformed) between the trial-wise decoding strengths of the cued/original item and the rotated item after transformation at Impulse 2. Error bars are the 95% C.I. of the mean. The marginally significant effect ($p < 0.1$, two sided) is marked with (~).

Generalization of coding schemes before and after transformation

We tested if the coding scheme used for the cued item before rotation (Impulse 1) generalized to the coding scheme after rotation (Impulse 2). Training the classifier on the cued item at Impulse 1 resulted in significant cross-generalization when tested on the cued item at Impulse 2 of rotation trials ($p = 0.028$, $BF_{10} = 2.648$ two-sided; Fig. 7, left). This is evidence that even in the face of mental manipulation, the code of the original, non-manipulated item persists, and its coding scheme remains relatively stable over time. We also tested if the classifier trained on the cued item at Impulse 1 generalised to the mentally rotated item at Impulse 2. We found no evidence for this ($p = 0.726$, $BF_{10} = 0.209$, two-sided; Fig. 7, right). This again suggests that the mental rotation of the cued item resulted in a new coding scheme for the rotated item, without removing the original item from the WM network.

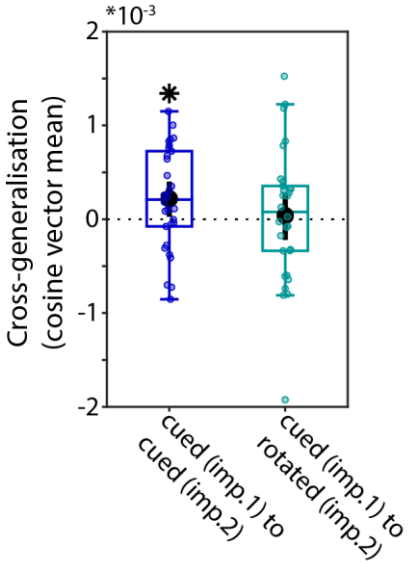

Figure 7. Cross-generalization of coding schemes between the original item before transformation at Impulse 1 and the original item and the rotated item after transformation at Impulse 2. The "0 rotation" condition is excluded. Error bars are 95% C.I. Statistically significant ($p < 0.05$, two-sided) cross-generalization is marked with an asterisk (*).

Relationship between trial-wise decoding strengths and WM performance

We tested if the trial-wise decoding strengths of the cued/original and the rotated item after transformation at Impulse 2 predicted trial-wise fluctuations in performance. Using logistic regression, we found

no evidence that accuracy was predicted by either the cued/original item or the rotated item on a trial-by-trial basis (cued: $p = 0.452$, $BF_{10} = 0.210$; rotated: $p = 0.098$, $BF_{10} = 1.027$, one-sided, Fig. 8A), though the latter showed a non-significant trend. Trial-wise variance in log-RT was predicted by both, however (cued: $p = 0.022$, $BF_{10} = 3.072$; rotated: $p = 0.010$, $BF_{10} = 9.555$, one-sided, Fig. 8B) as revealed by linear regression.

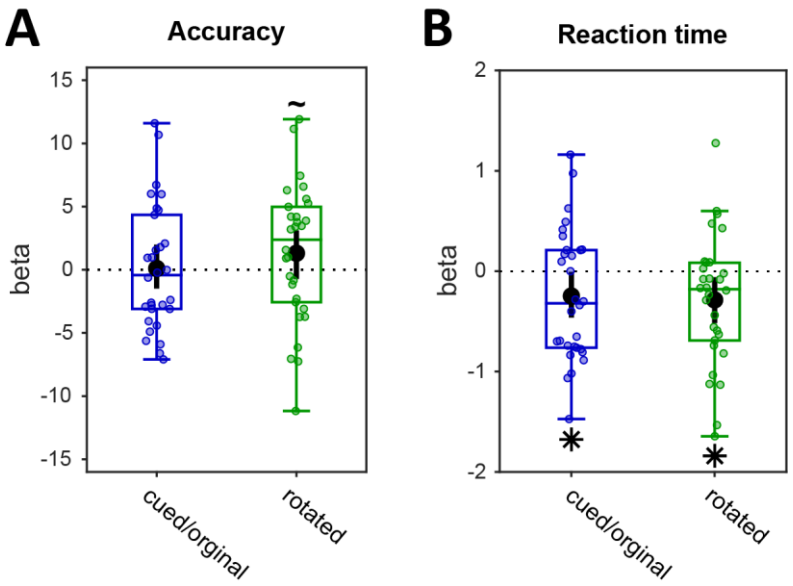

Figure 8. Relationship between trial-wise decoding at impulse 2 and performance. **(A)** Regression weights (beta) of logistic regression between item decoding and accuracy. **(B)** Regression weights (beta) of linear regression between item decoding and log-transformed reaction times. The "0 rotation" condition is excluded. Error bars are 95% C.I.. ***** $p < 0.05$, ~ $p < 0.1$, one-sided.

Discussion

For working memory to guide behaviour in the flexible manner required by the dynamic environment we live in, it cannot rely on the storage and static maintenance of sensory input alone, but also needs to be able to manipulate and update information when necessary. Here, we assessed how working memory represents items that were not directly formed by sensory input, but were imagined by variable degrees of mental rotation. From the EEG response to an impulse stimulus presented during WM maintenance we were able to decode not only visually presented items, but also imaginary ones. We discovered that the originally presented memory items, although rendered task-irrelevant after mental rotation, continued to be represented next to, and independent from, the imagined rotation products. By contrast, uncued items seemed to be purged rapidly from WM, and were not decodable, as was also previously observed (Wolff et al., 2017).

Representational coding of veridical and imagined items

As predicted by synaptic theories of WM (Mongillo et al., 2008; Zucker & Regehr, 2002), the visual impulse allowed an external readout of WM contents in the present study. Previous studies that have employed the impulse perturbation technique have shown its efficacy for stimuli that are presented in both the visual and auditory modality, and which are encoded in WM (Wolff et al., 2017; 2020a; 2020b). Here we found that the impulse is also effective for imagined items, supporting the idea that these items are similarly maintained in WM, and that they may equally utilise activity-quiescent brain states. The results furthermore suggest that the recovery of information in WM by means of a visual impulse is not dependent on the encoding of previous, direct sensory input, extending the scope of this approach. At the same time, the EEG patterns associated with the veridical (cued) item and the rotated item showed that their representations might differ in certain ways.

Although the cued item was clearly retained throughout the trial, change was observed in its neural representations across different time points. At the second impulse, the representation of the cued orientation was no longer spatially specific. Earlier studies using multiple impulse signals reported that the memory content remained spatially specific in time (Wolff et al., 2020b), which suggests that the loss of spatial specificity could be related to the present transformations. In a recent animal study, Panichello and Buschman (2021) observed that spatially specific memory content was transformed to another state with the involvement of the prefrontal cortex, once this memory representation became relevant for behaviour, and that this representation no longer carried spatial information (cf. Ester et al., 2009; Fukuda et al., 2016; Stokes et al., 2013). Likewise, in the current study, the cued orientation might have been transformed to a new state once it was relevant for mental rotation, and in which spatial information would not have a functional role. At the same time, we did observe cross-generalization of the cued item at Impulse 2 with that of Impulse 1, before rotation, suggesting that the coding scheme is otherwise still similar. Future research might clarify the role of transformations in the observed loss of spatial specificity by adding a systematic manipulation of the task-relevance of spatial information.

The results furthermore provided evidence that veridical and imagined items share the same representational substrate in visual processing areas of the brain, as both were sensitive to the visual impulse signal, and decodable from posterior electrodes alone. However, we also observed that there was no representational similarity between the original, cued orientation and the rotation product at the second impulse. That is, a given orientation angle (e.g., 30°) was represented differently, depending on whether it was previously

presented on the screen, or the end product of mental rotation (e.g., 60° rotated 30° counter-clockwise). This finding seems to contradict those of the fMRI study by Stokes and colleagues (2009) on mental imagery, in which mental imagery and visual perception activated representations in the same areas of the visual cortex. Similar correspondence was later reported for items held in WM (Albers et al., 2013), and Christophel and colleagues (2015) also observed that original and rotated colour patches seemed to share a similar coding scheme in visual and parietal cortices. It is possible that the discrepancy between the currently observed lack of representational similarity between rotated and original items, and the similarity observed by the aforementioned authors is due to the inherent temporal blurring of fMRI as well as their analysis approach, in which multiple time points were combined. It is also possible that subtle differences in the level of abstraction may have led to differences in coding scheme, as has been reported for auditory stimuli before (Linke & Cusack, 2015). Nevertheless, our results also show important commonalities between rotated and original items in that both were revealed in the impulse response, which is in line with these studies.

The representational difference between the original and rotated item that we observed contradicts a strict interpretation of memory models that define working memory as an activated and attended portion of long-term memory (e.g., Cowan, 1999; 2005; Oberauer, 2002; 2009). In such models, a single memory system houses all representations. The representations only differ with regard to their activity level; high levels correspond to items held in WM and in the focus of attention, while low levels correspond to items held in long-term memory. Accordingly, since in our design the visually presented items as well as the rotation products consisted of the same 6 orientation angles, these representations should have cross-generalized: Mental rotation should have reactivated the same, shared orientation representations from the low activity state as visually presented items would have. As indicated, our results showed instead that the veridical and imagined representations did not cross-generalize. Since veridical items did generalize over time, from the first to the second impulse, despite a loss of spatial specificity, the lack of cross-generalization between veridical and transformed representations cannot be ascribed to a lack of power. More generally, this lack of cross-generalization also supports the idea that our participants performed mental rotation by actually generating new images, rather than recalling discrete items from long-term memory.

Capacity limits

In view of the scarcity of WM space, possibly the most striking outcome of the present study was the observation that obsolete original items were retained in WM next to the task-relevant rotation products that

were derived from them. It is conceivable that this liberal use of WM space is also one of the factors that make mental rotation a relatively difficult task. The rotation process itself is certainly not trivial, as was evident from the progressive reduction in performance for increasing angles of rotation that we presently observed, and which is commonly found in mental rotation paradigms (e.g., Wexler et al., 1998; Searle & Hamm, 2017). After the rotation itself, the way in which both the now-irrelevant starting point and the outcome of this transformation process are maintained in WM may further compound that difficulty.

By contrast, previous work has shown that items that are no longer task-relevant, such as those that are uncued, are quickly purged, and can no longer be decoded even after impulse perturbation (Wolff et al., 2017). This has supported the view that WM may employ an active purging mechanism to get rid of information that is no longer needed. Alternatively, though, such information may also simply fade from WM, due to its inherently volatile nature, in the absence of active reinforcement (e.g., periodic refreshing). The latter account is in line with predictions from a recent computational model of WM, based on calcium-mediated short-term synaptic plasticity (Pals et al., 2020). The current results nevertheless provide further evidence that a purging mechanism may indeed exist, since the fate of irrelevant items, that is, uncued items and original items after rotation, was not the same. This may indicate that the brain treated them differently, such that the former, but not the latter item was actively removed.

There may be a good reason why the brain seems to maintain the obsolete original items. In the current task, holding onto the original item as well as the rotation instructions allows the participants to recreate or double-check the rotation product. There was some evidence that this helped to improve task performance; lower probe response times were associated with better decoding of both original and rotated items. In everyday scenarios it may also often make sense to remember more than the end-product of a mental transformation. For instance, if we predict the future location of a temporarily occluded vehicle in the environment, it would be useful to do so flexibly, to be able to make use of different estimates of its speed. Such flexibility requires the retention of the original input (the location of the vehicle) and the transformation (estimated distance covered based on speed), so that they can be used again and adjusted as needed. Thus, we propose that the maintenance behaviour observed in our experiment might reflect the prioritization of adaptive flexibility over WM storage capacity, despite the scarcity of the latter.

In this context it may be worth noting that in the study of WM, a lot of research has been devoted to charting WM capacity limits: The number of items (e.g., Miller, 1956; Cowan, 2001), the organization of information in individual properties and compound objects (e.g., Luck & Vogel, 1997; Xu, 2002), and the nature of WM capacity itself in terms of continuous resources or discrete slots (e.g., Zhang & Luck, 2008; Bays et al., 2009). The insights gained from this long-standing and important line of research remain highly relevant to this date. However, the present work suggests that a full understanding of WM cannot reflect on storage capacity alone. It also needs to develop a perspective on how the available storage capacity in WM may be utilized to support adaptive behaviour. The current results suggest that at times, the already strongly limited capacity of WM is filled very rapidly. From a strict capacity perspective, there would be little reason to assume that WM load at any one time after the initial retro-cue in our experiment would be more than a single item, yet the data showed differently. It is crucial to further our understanding of WM by examining the conditions that may foster such capacity-costly behaviour.

Conclusion

In line with synaptic theories of WM, we found that the representations of mentally rotated items and of visually presented items rely on the same neural substrate, as both could be decoded during maintenance from the EEG impulse response, even though their coding schemes appeared to be different. Importantly, in doing so the brain seems willing to sacrifice already-scarce WM capacity to support flexible behaviour in dynamic environments, as we observed that the original, no-longer task-relevant items continued to be maintained concurrently with transformed ones. This finding prompts the question of how much WM capacity is commonly ‘lost’ by this striking tendency to hold on to obsoleted information.

Acknowledgments

This research was in part funded by an Open Research Area grant to EGA (NWO 464.18.114), NA (DFG project number 396894956), and MGS (ESRC ES/S015477/1).

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology, 23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829–839. https://doi.org/10.1038/nrn1201

Bays, P. M., Catalao, R. F., & 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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience, 10*(2), 113–125. https://doi.org/10.1038/nrn2558

Christophel, T. B., Cichy, R. M., Hebart M. N., & Haynes J. D. (2015). Parietal and early visual cortices encode working memory content across mental transformations. *Neuroimage, 106*, 198-206. https://doi.org/10.1016/j.neuroimage.2014.11.018.

Cowan, N. (1999). An embedded-processes model of working memory. In: A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (p. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*(1), 87–185. https://doi.org/10.1017/s0140525x01003922

Cowan, N. (2005). *Working memory capacity*. Hove, United Kingdom: Psychology Press. https://doi.org/10.4324/9780203342398

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences, 7*(9), 415-423. https://doi.org/10.1016/s1364-6613(03)00197-9

Dijkstra, N., Bosch, S. E., & van Gerven, M. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences, 23*(5), 423–434. https://doi.org/10.1016/j.tics.2019.02.004

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience, 29(48),* 15258-15265. https://doi.org/10.1523/JNEUROSCI.4388-09.2009

Fukuda, K., Kang, M. S., & Woodman, G. F. (2016). Distinct neural mechanisms for spatially lateralized and spatially global visual working memory representations. *Journal of Neurophysiology, 116(4),* 1715-1727. https://doi.org/10.1152/jn.00991.2015

Iamshchinina, P., Kaiser, D., Yakupov, R., Haenelt, D., Sciarra, A., Mattern, H,m Luesebrink, F., Duezel, E., Speck, O., Weiskopf, N., & Cichy, R. M. (2021). Perceived and mentally rotated contents are differentially represented in cortical depth of V1. *Communications Biology, 4,* Article 1069. https://doi.org/10.1038/s42003-021-02582-4

JASP Team. (2018). JASP (version 0.9) [computer software]. Amsterdam, The Netherlands: University of Amsterdam.

Kamiński, J., Sullivan, S., Chung, J. M., Ross, I. B., Mamelak, A. N., & Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature Neuroscience, 20*(4), 590-601. https://doi.org/10.1038/nn.4509

Kamiński, J. and Rutishauser, U. (2020). Between persistently active and activity-silent frameworks: Novel vistas on the cellular basis of working memory. *Annals of the New York Academy of Sciences, 1464,* 64-75. https://doi.org/10.1111/nyas.14213

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1-16.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *Journal of Cognitive Neuroscience, 25*(1), 127-142. https://doi.org/10.1162/jocn_a_00305

Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management, 30*(4), 110–119. https://doi.org/10.3905/jpm.2004.110

Linke, A. C., & Cusack, R. (2015). Flexible information coding in human auditory cortex during perception, imagery, and STM of complex sounds. *Journal of Cognitive Neuroscience, 27*(7), 1322–1333. https://doi.org/10.1162/jocn_a_00780

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science, 319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61-64. https://doi.org/10.20982/tqmp.04.2.p061

Muhle-Karbe, P. S., Myers, N. E., & Stokes, M.G, (2021). A Hierarchy of Functional States in Working Memory. *Journal of Neuroscience, 41*(20), 4461-4475. https://doi.org/10.1523/JNEUROSCI.3104-20.2021

Myers N. E., Rohenkohl, G., Wyart, V., Woolrich, M. W., Nobre, A. C., & Stokes, M. G. (2015). Testing sensory evidence against mnemonic templates. *eLife, 4,* Article e09000. https://doi.org/10.7554/eLife.09000

Nili, H., Wingfield, C., Walther, H., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4), Article e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation: Advances in Research and Theory, 51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences, 15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004

Pals, M., Stewart, T. C., Akyürek, E. G., & Borst, J. P. (2020). A functional spiking-neuron model of activity-silent working memory in humans based on calcium-mediated short-term synaptic plasticity. *PLoS Computational Biology, 16*(6), Article e1007936. https://doi.org/10.1371/journal.pcbi.1007936

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature, 592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience, 20*(10), 624–634. https://doi.org/10.1038/s41583-019-0202-9

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective coding for objects in primate prefrontal cortex. *Journal of Neuroscience, 19*(13), 5493–5505. https://doi.org/10.1523/JNEUROSCI.19-13-05493.1999

Rieser, J. J., Guth, D. A., & Hill, E. W. (1986). Sensitivity to perspective structure while walking without vision. *Perception, 15*(2), 173–188. https://doi.org/10.1068/p150173

Roelfsema, P. R., & de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual Review of Vision Science, 2*, 131–151. https://doi.org/10.1146/annurev-vision-111815-114443

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), 1136–1139. https://doi.org/10.1126/science.aah7011

Searle, J. A., & Hamm, J. P. (2017). Mental rotation: An examination of assumptions. *WIREs Cognitive Science, 8*(6), 1–14. https://doi.org/10.1002/wcs.1443

Stokes, M. G., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience, 29*(5), 1565-1572. https://doi.org/10.1523/JNEUROSCI.4657-08.2009

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron, 78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding rich spatial information with high temporal resolution. *Trends in Cognitive Sciences, 19*(11), 636–638. https://doi.org/10.1016/j.tics.2015.08.016

Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (2020). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition, 28*(5-8), 420–432. https://doi.org/10.1080/13506285.2020.1825141

Trübutschek, D., Marti, S., Ueberschär, H., & Dehaene, S. (2019). Probing the limits of activity-silent non-conscious working memory. *Proceedings of the National Academy of Sciences U. S. A., 116*(28), 14358-14367. https://doi.org/10.1073/pnas.1820730116

van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience, 22*(3), 477–483. https://doi.org/10.1038/s41593-018-0335-6

van Loon, A.M., Olmos-Solis, K., Fahrenfort, J.J., & Olivers, C. N. L. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *Elife, 7*, Article e38677. https://doi.org/10.7554/elife.38677.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage, 137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Wan, Q., Cai Y., Samaha, J., & Postle, B. R. (2020). Tracking stimulus representation across a 2-back visual working memory task. *Royal Society Open Science, 7*, Article 190228. https://doi.org/10.1098/rsos.190228

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition, 68*(1), 77–94. https://doi.org/10.1016/s0010-0277(98)00032-8

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience, 9*, Article 123. https://doi.org/10.3389/fnsys.2015.00123

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience, 20*(6), 864–871. https://doi.org/10.1038/nn.4546

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyürek, E. G. (2020a). Unimodal and bimodal access to sensory working memories by auditory and visual impulses. *Journal of Neuroscience, 40*(3), 671–681. https://doi.org/10.1523/JNEUROSCI.1194-19.2019

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020b). Drifting codes within a stable coding scheme for working memory, *PLoS Biology, 18*(3), Article e3000625. https://doi.org/10.1371/journal.pbio.3000625

Xu, Y. (2002). Encoding color and shape from different parts of an object in visual short-term memory. *Perception & Psychophysics, 64*(8), 1260–1280. https://doi.org/10.3758/bf03194770

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zucker, R. S., & Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual Review of Physiology, 64*, 355–405. https://doi.org/10.1146/annurev.physiol.64.092501.114547