Cognition 245 (2024) 105739

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Internal attention modulates the functional state of novel stimulus-response associations in working memory

Silvia Formica [a,*], Ana F. Palenciano [b], Luc Vermeylen [c], Nicholas E. Myers [d], Marcel Brass [a], Carlos González-García [b]

[a] Berlin School of Mind and Brain, Department of Psychology, Humboldt Universität zu Berlin, Berlin 10117, Germany
[b] Mind, Brain, and Behavior Research Center, University of Granada, Granada 18071, Spain
[c] Department of Experimental Psychology, Ghent University, Ghent 9000, Belgium
[d] School of Psychology, University of Nottingham, Nottingham NG7 2RD, UK

ARTICLE INFO

Keywords:
Working memory
Attention
Instructions
Drift-diffusion modeling
Retro-cues

ABSTRACT

Information in working memory (WM) is crucial for guiding behavior. However, not all WM representations are equally relevant simultaneously. Current theoretical frameworks propose a functional dissociation between 'latent' and 'active' states, in which relevant representations are prioritized into an optimal (active) state to face current demands, while relevant information that is not immediately needed is maintained in a dormant (latent) state. In this context, task demands can induce rapid and flexible prioritization of information from latent to active state. Critically, these functional states have been primarily studied using simple visual memories, with attention selecting and prioritizing relevant representations to serve as templates to guide subsequent behavior. It remains unclear whether more complex WM representations, such as novel stimulus-response associations, can also be prioritized into different functional states depending on their task relevance, and if so how these different formats relate to each other. In the present study, we investigated whether novel WM-guided actions can be brought into different functional states depending on current task demands. Our results reveal that planned actions can be flexibly prioritized when needed and show how their functional state modulates their influence on ongoing behavior. Moreover, they suggest the representations of novel actions of different functional states are maintained in WM via a non-orthogonal coding scheme, thus are prone to interference.

1. Introduction

To achieve our goals, we constantly need to internally maintain and manipulate information that is no longer available in the environment, independently of sensory stimulation. The cognitive system assumed to support these computations is commonly referred to as working memory (WM) (Baddeley, 2012; Baddeley & Hitch, 1974). This ability is essential to guide adaptive behavior and as such it plays a crucial role in a vast range of higher cognitive functioning, such as decision making, planning, and reasoning (Oberauer, 2009).

Given its central role, extensive research has focused on understanding the underlying cognitive and neural mechanisms of WM, and to characterize some relevant features of its architecture, such as its capacity limitations (Cowan, 2010; Fukuda, Awh, & Vogel, 2010; Luck & Vogel, 1997). One crucial and, until recently, relatively overlooked aspect of WM is its flexibility. Namely, when holding multiple items in WM, the level of priority of each of these can be dynamically adjusted as a function of their behavioral relevance (de Vries, Slagter, & Olivers, 2019; Olivers, Peters, Houtkamp, & Roelfsema, 2011; Souza & Oberauer, 2016). Recent theoretical proposals highlight the role of endogenous attention not only in selecting currently relevant WM representations, but also in the goal-oriented modulation of their functional state (i.e., their functional attributes dependent on momentary task relevance). In other words, directing attention to a specific WM item also triggers the reconfiguration of its format, from a purely mnemonic trace to an action-guiding representational state that is optimized to efficiently perform the task at hand (Myers, Stokes, & Nobre, 2017; Nobre & Stokes, 2019; Nobre & Van Ede, 2018).

Empirical evidence supporting the existence of different functional states of concurrently held memoranda is rapidly growing. WM items that are currently prioritized to exert their influence on the ongoing behavior are referred to as 'active', in contrast to 'latent' items that are

* Corresponding author.
E-mail address: silvia.formica@hu-berlin.de (S. Formica).

https://doi.org/10.1016/j.cognition.2024.105739
Received 2 June 2023; Received in revised form 22 January 2024; Accepted 4 February 2024
Available online 9 February 2024

merely maintained but should not drive behavior (Stokes, Muhle-Karbe, & Myers, 2020). Crucially, task demands might require latent items to become active and drive behavior at a later time, raising the need for a rapid and flexible prioritization mechanism (de Vries et al., 2019). Research in the field of visual WM has consistently shown that relevant representations of visual memories can be dynamically modulated by means of attentional prioritization (Olivers et al., 2011; Sprague, Ester, & Serences, 2016). For example, one out of several encoded visual items can be retrospectively brought in the focus of attention, and serve as active target template during a visual search task (de Vries, van Driel, Karacaoglu, & Olivers, 2018; de Vries, van Driel, & Olivers, 2017; van Driel, Ort, Fahrenfort, & Olivers, 2019; van Loon, Olmos-Solis, Fahrenfort, & Olivers, 2018). The distinction between active and latent representations can be potentially supported by qualitatively distinct neural states, although the exact mechanisms allowing for such differential maintenance remain elusive (Stokes et al., 2020). In a recent study, it has been demonstrated that both active and latent items can be decoded from patterns of EEG activity, but only the quality of the representation of the currently active item has a direct influence on behavior (Muhle-Karbe, Myers, & Stokes, 2021). Analogously, Henderson, Rademaker, and Serences (2022) showed that both prioritized and unprioritized items could be decoded from the neural activity, but differently across brain areas, suggesting that the active items gets recoded into a more action-oriented format including the specification of a motor plan and recruiting motor rather than visual areas (Henderson et al., 2022).

Despite the theoretical prominence of this proposal in visual WM, the role of endogenous attention in prioritizing WM representations has been seldom investigated when such representations go beyond visual memories, as in the case of planned actions and stimulus-response (SR) associations. Specifically, given the role of WM representations to inform ongoing behavior, the functional relevance of action plans in WM needs to be flexibly modulated in order to achieve the desired goal. Research in the field of novel SR implementation has already suggested the existence of different representational formats in which novel task sets can be held in WM. This evidence comes from studies in which different readouts from memory encourage different maintenance strategies. As such, representations of planned actions can range from a more declarative format in which the memoranda are maintained without the intention to implement the action, to a more procedural, action-oriented representation, capable of driving the implementation of the action (Brass, Liefooghe, Braem, & De Houwer, 2017; Formica, González-García, & Brass, 2020; González-García, Formica, Liefooghe, & Brass, 2020; González-García, Formica, Wisniewski, & Brass, 2021; Liefooghe, Wenke, & De Houwer, 2012; Meiran, Pereg, Kessler, Cole, & Braver, 2015b, 2015a; Oberauer, 2009). According to this literature, SRs in WM are also sensitive to top-down biases, resulting in the same memoranda being encoded and held in different representational formats depending on the expected task demands (i.e., execution or mere maintenance) (Whitehead & Egner, 2018b, 2018a). Moreover, the interplay between the representational state of novel action plans and attentional prioritization has recently been investigated, suggesting that attention plays a role in shifting between representational formats (González-García et al., 2020). Notably, research in this field mostly treated these representational states as dichotomous alternatives and investigated how different expected readouts favored the use of one or the other state. Novel SRs held in declarative format are expected to exert limited to no interference on the ongoing behavior (Liefooghe et al., 2012). However, this perspective is hard to reconcile with studies using visual material, reporting mixed evidence as to whether prospectively relevant (but currently latent) items interfere with ongoing behavior, showing either an effective shielding of the latent items (Mallett & Lewis-Peacock, 2018; van Loon, Olmos-Solis, & Olivers, 2017), or their interfering effect (Carlisle & Woodman, 2019). These findings suggest that, rather than a sharp dichotomy, representational states of WM contents might range from active to latent on a continuous spectrum, resulting in more or less interference of the latent items on ongoing behavior depending on task demands.

Therefore, in the current study we first aimed at investigating whether internal attention to encoded novel SRs could effectively modulate their prospective relevance, and thus their influence on ongoing behavior. To this goal, we developed a task wherein, after the encoding of four novel SR mappings, a retro-cue would tag one pair of action plans as relevant for the first of two subsequent probes (i.e. active SRs). Crucially, the retro-cue also provided information with respect to the functional relevance of the second pair of encoded SRs, which could be either tagged as irrelevant and thus dropped from WM (Drop condition), certainly relevant (100% probability) for a second probe (Prepare condition), or potentially relevant (50% probability) for the second probe (Maintain condition). Importantly, the Prepare and Maintain conditions did not differ with respect to the amount of information being held in WM (4 SRs), but only in the functional state of latent items. We hypothesized that the functional state of the latent set of SRs (from 0%, to 50%, to 100% probability of becoming active) would affect the efficiency of the implementation of the currently relevant SRs: specifically, we predicted that the higher the relevance of latent items, the more detriment they would cause to the implementation of the active SRs.

If this hypothesis were to be confirmed, we aimed at further investigating such impact of latent items on active ones. To do so, we devised the first probe after the retro-cue to be bivalent (i.e., including two stimulus features). Each of the two presented features was associated with a specific response according to the initially encoded SRs. Therefore, the probe could be congruent, if both stimulus features elicited the same response, or incongruent, if the response associated with the latent SR prompted an opposite response with respect to the active item. In line with our hypotheses, we observed an interaction of congruency with the functional state of latent items: the more likely a latent SR is to become relevant in the future, the more it will interfere with ongoing behavior.

2. Methods

The experimental procedure was preregistered before data collection https://aspredicted.org/qe5mt.pdf. Task files, raw data and analysis scripts are available at https://osf.io/phxq4/.

2.1. Participants

Participants were recruited through the online platform Prolific Academic Website (https://www.prolific.co/) and received a compensation of £5.50 per hour. To incentivize a good performance, we offered an additional reward of £1 to participants with accuracies within the first quartile of the sample. For ethical reasons, all participants eventually received the additional reward. Eligible participants needed to be of age between 18 and 35 years, and fluent in English to ensure a full understanding of the task instructions. Sample size was estimated a-priori to detect a small effect size (Cohen's d = 0.25) with 80% power in a one-sample *t*-test (see Data Analysis section), resulting in a sample of 101. An initial group of 109 participants performed the task, 4 of which were excluded based on our population-level rejection criterion, namely error rate (see below for details). Therefore, our final sample consisted of a total of 105 participants (52 women, 51 men, 2 non-binary; 13 left-handed, 92 right-handed) with a mean age of 24.89 years (SD = 4.89).

2.2. Material

The stimuli set consisted of 360 shapes evenly distributed along a continuous shape wheel (Li, Liang, Lee, & Barense, 2020) and 360 colors sampled from an analogous color wheel (built using the hue values from the HSV model after fixing the saturation and value parameters at 1). Critically, these stimuli were employed so that a large number of novel SRs (i.e. combination of a specific color/shape with a response, see below) could be created for each trial, while controlling the similarity

between stimulus features. The experiment was programmed in JsPsych v.6.1.0 (de Leeuw, Gilbert, & Luchterhandt, 2023).

2.2.1. Procedure

Each trial started with an encoding screen (3000 ms) consisting of two shapes with black contour and white filling, and two colored dots (Fig. 1). The pairs of stimuli were arranged so that each occupied the upper or lower half of the screen; the locations occupied by shape and colors were counterbalanced across trials. The two shapes and the two colors presented in the encoding screens always lay 180° away from each other on the shape and color wheels, respectively, ensuring maximal discriminability and constant distance across trials.

Participants were instructed to memorize the shapes and colors, and to associate their location on the screen with a specific response effector. Namely, the shape and color on the left side of the screen were to be associated with the left index finger (positioned on the keyboard key 'd'), whereas right side stimuli corresponded to right index finger response (keyboard key 'k'). This setup resulted in 4 novel SRs for each trial: one shape associated with a left response, one shape associated with a right response, one color associated with a left response, and one color associated with a right response. Each shape and color pair appeared twice throughout the whole experiment. However, the same set of four stimuli never appeared in the same encoding screen more than once, but rather each shape pair at its second presentation would be associated with a color pair orthogonal with respect to the color pair it was associated with on the first iteration. Again, the scope of this manipulation was to ensure maximal discriminability between stimuli, while maximizing also the novelty of the encoding screens across trials. Immediately after the encoding screen, four gray dots were displayed for ~16 ms in the locations previously occupied by the stimuli, to act as post-stimuli masks and reduce the role of iconic memory traces (Myers, Chekroud, Stokes, & Nobre, 2018).

After an 800 ms interval with a centrally displayed fixation cross, a feature retro-cue appeared on screen for 500 ms. This consisted of two strings of characters arranged on two rows and indicating which stimulus feature (i.e., shape, color) was going to be relevant in a first bivalent probe (indicated by the word on the upper row) and a second univalent probe (indicated by the word on the lower row). The upper row of the retro-cue was 100% valid, whereas the second row differed across three conditions and was intended to determine the functional state of the encoded SRs. Specifically, in the *Drop* condition, the first and second probes required responding to the same stimulus feature (e.g., *shape-shape* retro-cue), rendering the unmentioned stimulus feature (in this example, *color*) irrelevant for the remainder of the trial so that, in principle, participants could *drop* the corresponding mappings. In other words, the uncued stimulus feature had 0% probability of becoming active for the second probe. In a second case, the two probes required attending to two different stimulus features (e.g., *shape-color* retro-cue, prompting a response based on *shape* for the first probe and on *color* for the second probe). In this condition, which we refer to as *Prepare*, both features were equally task-relevant insofar as they were needed to correctly perform the task, although at different moments in the trial. Therefore, participants could prepare upfront to attend to each of the two stimulus features in the two probes, and both features were actively tagged as relevant throughout the trial. In this condition, the stimulus features indicated in the second row of the retro-cue had 100% probability of becoming active for the second probe. Finally, in a last condition, the retro-cue informed participants only regarding the feature relevant to respond to the first probe, remaining agnostic with respect to the demands of the second probe (e.g., *shape-????* retro-cue). In this case, participants still had to maintain all the initially encoded information, as the second probe was equally likely to require attending to the already probed stimulus feature (in our example, *shape*) or the opposite one (*color*). We refer to this as the *Maintain* condition, as the unmentioned stimulus feature needed to be maintained at least until the second retro-cue, in which it would be relevant in 50% of trials, but participants needed not to actively prepare for it, in contrast with the *Prepare* condition. Altogether, we expected this retro-cue manipulation to modulate the functional state of the irrelevant feature during the first probe.

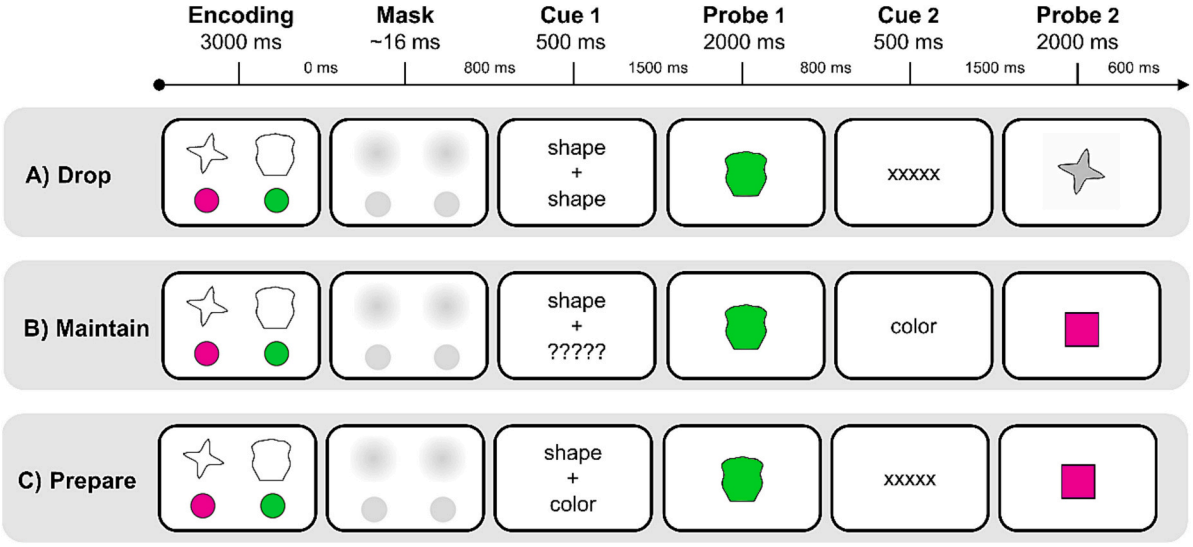

Fig. 1. Experimental paradigm.
Participants encoded four novel SR mappings at the beginning of each trial. In the encoding screen, two shapes and two colors were presented, each associated with a left or right response depending on their location on the screen. Later, a first retro-cue (100% valid) provided information regarding the relevant feature in Probe 1 (word in the upper row of the retro-cue) and Probe 2 (word in lower row). Specifically, the retro-cue could A) display the same word (e.g. shape-shape) and thus render only one feature relevant for both probes, while the other could be dropped from WM (Drop condition), B) provide information about the first probe only (e.g. shape-?????), which rendered one feature relevant but prevented participants from dropping the uncued one, since it could become relevant later on the trial (Maintain condition), or C) indicate with 100% validity one feature to be relevant for the first probe and the other for the second probe (e.g., shape-color, Prepare condition). For Probe 1, participants had to provide the response associated to the cued feature (e.g. shape), while ignoring the uncued one (e.g. color, which could entail a compatible or incompatible response, or no response in the case of neutral trials). Finally, a second probe was presented after a second retro-cue, that could indicate the currently relevant stimulus feature (Maintain condition), or show an uninformative string of characters ('xxxxx', Drop and Prepare conditions).

After the retro-cue, a 1500 ms fixation cross preceded the appearance of the first probe. The probe was displayed for a maximum of 2000 ms or until a response was provided via keyboard press. Probe 1 consisted of a bivalent stimulus (i.e., a colored shape) presented in the middle of the screen. The specific characteristics of the bivalent stimulus in Probe 1 determined the second crucial manipulation of our design: *Congruency*. The levels of this manipulation reflected the response plan associated with the currently latent stimulus dimension. In the *Neutral* condition, the irrelevant stimulus dimension was not associated with any response (i.e., if participants were asked to respond to the shape of the probe, its color would be a neutral gray; if the target feature was color, this was presented in a neutrally-shaped square). The *Congruent* condition consisted of a bivalent stimulus which contained two features associated to the same response, as opposed to *Incongruent* probes in which the two features prompted incompatible responses.

After 800 ms of fixation, a second retro-cue appeared, for a duration of 500 ms. This consisted of one single string of characters presented in the middle of the screen. In the *Maintain* condition, this retro-cue indicated the stimulus feature to attend to for the second probe (i.e., *shape* or *color*), whereas in the *Drop* and *Prepare* conditions it consisted of a meaningless string of Xs (*XXXXX*). This was done to achieve overall perceptual and timing similarity across the different WM conditions, while ensuring participants were fully informed on the task demands raised by the second probe before its appearance. After a 1500 ms fixation interval, Probe 2 appeared for 2000 ms or until response. Probe 2 consisted of either a shape with a gray filling, or a colored square, thus highlighting only one stimulus feature (i.e., same as in the Neutral condition in Probe 1). Finally, a 600 ms intertrial interval was presented after Probe 2.

The main task consisted of a total of 228 trials, equally distributed across conditions. Among these, in 12 trials ($\sim$ 5%) the first probe displayed a catch stimulus. With respect to the cued feature, catch probes featured an intermediate value to the ones encoded at the beginning of the trial (e.g., if the encoded colors were blue and yellow, the catch probe would be green), while the irrelevant feature had neutral characteristics (i.e., square shape or gray color), similar to neutral trials. In these occasions, participants had to respond by pressing both keys. This was done to ensure participants encoded 4 SR mappings and did not use any strategy to limit load and encode only a subset of them. In fact, if participants attempted to reduce the number of items to encode by only focusing on the shape and color on one side of the screen (e.g., encoding only the two items on the left) they would have failed to identify the novel color/shape and instead interpret it as the color/shape assigned to the right side of the screen but not encoded. This would result in chance-level accuracy scores for these trials.

Before the main task, participants completed a practice phase to familiarize with the task. First, participants performed 10 trials of a simplified version of the task, where they had to provide associated responses to either the shape or the color of univalent stimuli. Of these 10 trials, 2 of them were catch trials, and participants were instructed on how to respond to these. Participants repeated this phase until they achieved at least 80% of accuracy. In a second practice phase, participants performed 18 trials (3 catch) of the actual task. Again, they repeated this phase until at least 80% of accuracy was achieved. In all cases, feedback on the responses was provided after each trial.

2.3. Data analysis

All analyses and results reported pertain to responses to Probe 1. Our hypotheses concerned the performance in response to the currently relevant feature depending on the priority status of the currently irrelevant feature, and therefore were addressable only during Probe 1. Responses to Probe 2 are considered as part of the data cleaning procedure, but performance is not evaluated across experimental conditions. This is due to the complexity of the task design during Probe 2, which made it difficult to reliably assess differences.

2.4. Preregistered analyses

2.4.1. Data cleaning procedure

We were only interested in participants' performance in response to regular trials. Therefore, catch trials (mean accuracy: 69% $\pm$ 19%, indicating that on average participants were aware that catch trials contained features not matching any of the encoding items) were discarded from further analyses.

Participants were excluded if their error rate in response to regular trials of either Probe 1 or Probe 2 exceeded 40%. This resulted in 4 participants being discarded, leaving a sample of 105 participants for the remaining analyses.

All participants completed a total of 216 regular trials, 24 for each experimental condition. In line with our preregistered exclusion criteria, we discarded trials with an incorrect response to Probe 2 (on average, 19.92 $\pm$ 14.94 trials per participant were discarded, range = [2–72]). This was done to only include trials in which SRs were successfully maintained throughout the trial. Additionally, omissions errors in Probe 1 (i.e., no response provided within the response deadline of 2000 ms) were also discarded (on average, 2.23 $\pm$ 2.45 trials per participant were discarded, range = [0−13]).

Data resulting from the above cleaning procedures entered the analyses for Error rates, distributional analyses, and drift diffusion modeling. On average, 21.54 $\pm$ 2.56 trials per condition and participant were retained for these analyses.

With respect to analyses on RTs, as preregistered, we only wanted to consider trials with correct responses. Therefore, from the dataset obtained with the aforementioned cleaning procedures, we additionally excluded trials with errors in Probe 1 (on average, 8.14 $\pm$ 7.01 trials per participant were discarded, range = [0–35]). Furthermore, we only retained Probe 1 responses with a RT within 2 standard deviations from the mean, separately computed for each participant and experimental condition (on average, 1.09 $\pm$ 0.60 trials per participant were discarded, range = [0–3]). This cleaning procedure resulted in an average of 19.55 $\pm$ 3.34 trials per condition and participant entering the RTs analyses.

All statistical analyses were performed using the Pingouin package, version 0.5.3 (Vallat, 2018) running in Python 3. Bayes Factors (BF) for *t*-tests were obtained from the output of the t-test function of the Pingouin package (using the default prior parameters), whereas for ANOVAs we computed Bayes Factors with the JASP software, again using the default prior parameters (Jasp Team, 2019).

2.4.2. Error rates analysis

To assess the quality of WM representations, we focused our analyses on error rates in response to Probe 1, as we were primarily interested in the main effect of the functional state of the irrelevant SRs (i.e., Functional State) and its interaction with the congruency of the probe.

Error rates were analyzed in a 3 $\times$ 3 repeated measures ANOVA, using Functional State (Drop, Maintain, Prepare) and Congruency (Neutral, Congruent, Incongruent) as within-subject factors. Our main preregistered prediction consisted of a significant interaction between the two factors Functional State and Congruency on error rates. To further investigate this interaction in our two conditions of interest, namely, Prepare and Maintain, we computed three additional indices. The *congruency effect* was defined as the difference in errors between congruent and incongruent trials; the *congruency benefit* as the difference between neutral and congruent trials; and the *congruency cost* as the difference between incongruent and neutral trials. We then compared these indices between the two Functional State conditions by means of one-tailed paired-samples t-tests, assuming larger values in the Prepare condition compared to Maintain condition. Sample size calculations in the preregistration were based on these contrasts. Greenhouse-Geisser correction was adopted whenever the sphericity assumption resulted to be violated in the Mauchly's test.

2.4.3. Drift diffusion modeling

The predicted changes in error rate across Functional States could arise from a number of non-mutually exclusive mechanisms, such as needing more time to adequately configure an appropriate task set in the Prepare condition, or a greater ability to accumulate relevant information when less SRs have to be maintained in an active Functional State. To better understand how the implementation process is carried out, and to adjudicate between potential effects of the functional state of the irrelevant SRs, we modeled our data using the hierarchical drift diffusion model (HDDM) toolbox (Wiecki, Sofer, & Frank, 2013). HDDM treats two-alternative decision-making as a process of evidence accumulation towards one of two decisional boundaries over time (Ratcliff & Rouder, 1998). Although originally developed to model decision processes in perceptual tasks, it has been fruitfully applied also in the WM domain, where the evidence being accumulated in favor of one of two alternatives rely on some internally held representation (Ratcliff, Smith, Brown, & Mckoon, 2016). It has the advantage of taking into account the available data in its entirety, thus retrieving parameters to fit both correct trials and errors RTs distributions. Crucially, such parameters can be mapped onto psychologically meaningful processes, making it possible to infer which cognitive operations are affected by the experimental manipulations. In its simplest formulation, the decision process can be described by 4 parameters: the *drift rate (v)*, reflecting the pace and efficiency of evidence accumulation; the *non-decision time* (t_0), encompassing all cognitive processes not directly associated to the decision itself, such as perceptual and motor operations; the *decision threshold (a)*, referring to the amount of evidence required to reach a specific decision; and the *starting point (z)*, indicating whether there is a systematic bias towards one of the two options. The HDDM is considered hierarchical insofar as it first uses data from all the participants to estimate group-level mean parameters, and then it uses these group-level priors to constrain the estimation of the subject-specific parameters. Consequently, it allows for more stable results, even with fewer data per participant with respect to the traditionally used algorithms (Lerche, Voss, & Nagler, 2017). The estimation of model parameter distributions within the HDDM toolbox relies on a Markov-chain Monte Carlo sampling procedure (Gamerman & Lopes, 2006). We used a chain with 10,000 samples; the first 1000 samples were discarded as burn-in, to allow for the sampling procedure to settle around a value after an initial more exploratory sampling. To reduce autocorrelation in the retained samples, we additionally discarded every second sample.

In a deviation from our preregistered plan, for the modeling analyses we focused on the main contrast of interest and reduced the design size. The rationale of this choice was to simplify the modeling procedure by reducing the number of conditions and therefore ensure better model fit to our empirical data. Moreover, the Prepare and Maintain conditions are matched in terms of working memory load (i.e., in both, four SRs needed to be maintained, with different priority levels), making them interesting to compare with our modeling approach without the confound of a different number of encoded items. Therefore, we investigated the effects of Functional State (Prepare vs Maintain) and Congruency (Congruent vs Incongruent) on performance in Probe 1, by fitting a total of 7 alternative models. Besides the 4 models that were included in the preregistration, we fitted 3 additional models. The rationale for exploring the fitting performance of these new models was based on the observation of the empirical RTs distributions and on recent literature that relies on drift rate (rather than non-decision time) to model congruency effects (Ulrich, Schröter, Leuthold, & Birngruber, 2015; White, Servant, & Logan, 2017). Moreover, previous modeling studies investigating the beneficial effect of retro-cues on memory performance showed an increase in drift rate and a reduction in non-decision time (Shepherdson, Oberauer, & Souza, 2018). For the current study, we hypothesized our two independent variables to impact either the rate of evidence accumulation (v) or the non-decision time (t_0). The former would imply that the manipulation of the independent variable results in a decrease in the quality of the evidence entering the decision process, whereas a modulation of t_0 would reflect a slowing in the non-decision phase, that could encompass perceptual, retrieval, and motor processes associated with the decision. Therefore, the model fitting was carried out by keeping fixed for all alternative models the parameters *a* and *z*, as we had no specific hypotheses on how our experimental manipulations would impact them. On the contrary, for each of the different models we implemented (Table 1) we allowed different combinations of the parameters *v* and t_0 to vary depending on the two experimental manipulations.

Model fitting was initially assessed based on the Deviance Information Criterion (DIC), where lower values indicate better fit of the model to the empirical data. Next, we explored model fit in more detail. First, we ran a posterior predictive check: 500 datasets were generated from random parameter values drawn from the posterior distributions. From these 500 datasets we calculated the mean accuracy, mean RT and mean of the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles for correct and error distribution. We then inspected the mean squared error (MSE) which quantifies the misfit between the mean of the predicted values compared to the observed values. Second, to get a more condition-specific and visual assessment of the model fit, we simulated data for all conditions (5000 trials per condition, based on the mean of the posterior parameter estimates) and qualitatively compared the predicted and observed RT distributions. Model 4 and Model 7 resulted to have the lowest DIC values. However, differences in DIC scores lower than 5 are traditionally considered inconclusive (Cain & Zhang, 2019) thus we turned to the MSE to pick one among the two. To further corroborate our decision, we took into account also the considerations by Ulrich et al. (2015), highlighting the inconsistency of mapping the effects of Congruency onto the non-decision time parameter. As a conclusion, we decided to consider Model 7 as the best fitting model, and we report results accordingly. Nevertheless, drift rate results are consistent across the two models.

2.5. Exploratory analyses

2.5.1. Distributional analyses

Given the HDDM results (see Results section), we further conducted distributional analyses to better describe the nature of interference in the current task. This approach captures RTs and accuracy differences between conditions across response speed, highlighting whether the effect of interest is larger for faster or slower responses (Dittrich, Kellen, & Stahl, 2014). To this end, we calculated the delta functions for RT and accuracy of each Functional State. First, we separated the data for each Functional State in quintiles (based on RT), and then, for each quintile, we computed the RT or accuracy difference (i.e., the delta) between congruent and incongruent trials. Next, we used Bayesian Linear Mixed Effects modeling to investigate whether there were differences in the slopes of the delta functions between Functional States. We predicted the RT or accuracy delta values to be modulated by the fixed effects of quintile and Functional State. We used a full random effects structure (i.e., quintile, Functional State, and their interaction were specified as

Table 1
Fitted HDDM models.

Model #	Functional State	Congruency	Preregistered	DIC value	MSE
Model 1	t_0	t_0	yes	1940.64	1.20
Model 2	v	t_0	yes	1864.28	1.20
Model 3	v, t_0	t_0	yes	1865.32	1.20
Model 4	v	v, t_0	yes	1782.24	1.21
Model 5	t_0	v	no	1870.65	1.18
Model 6	v, t_0	v	no	1791.03	1.20
Model 7	v	v	no	1785.76	1.20

The table lists the seven hDDM models that were fitted. For each model, a different combination of the parameters v and t_0 were allowed to vary depending on the experimental condition. Parameters a and z were always fixed. The table additionally indicates for each model whether it was included in the preregistration, and the two considered measures of model fit (i.e. DIC and MSE).

random slopes). The default (i.e. flat) priors of the brms package were used (Burkner, 2015). All Rhats were below 1.1 suggesting that the chains converged. We report the 95% confidence interval (highest density interval) and probability of direction (*pd*, the effect is considered to exist if $pd > 97.5$, Makowski, Ben-Shachar, Chen, & Lüdecke, 2019).

2.5.2. Reaction times analysis

While our main dependent variable of interest was error rates, we performed the same 3 × 3 rmANOVA with reaction times (RTs) for completeness.

3. Results

3.1. Preregistered analyses

3.1.1. Error rates analysis

The 3 × 3 repeated measure ANOVA on error rates revealed the predicted main effect of Functional State ($F_{1.34,139.21} = 39.39$, $p < .001$, $\eta p2 = 0.27$, $BF_{10} > 150$), and a main effect of Congruency ($F_{1.26,\ 131.58} = 54.07$, $p < .001$, $\eta p2 = 0.34$, $BF_{10} > 150$) as well. Participants were overall less accurate when responding to Incongruent compared to Congruent and Neutral probes, and error rate increased across Functional States. Crucially, as predicted, we observed a significant Functional State x Congruency interaction ($F_{2.13,222.06} = 17.71$, $p < .001$, $\eta p2 = 0.14$, $BF_{10} > 150$). The effect of Congruency was significant across all three levels of the variable Functional State (Drop: $F_{1.67,\ 173.29} = 7.06$, $p = .001$, $\eta p2 = 0.064$; Maintain: $F_{1.53,\ 159.45} = 39.9$, $p < .001$, $\eta p2 = 0.277$; Prepare: $F_{1.25,\ 129.81} = 38.1$, $p < .001$, $\eta p2 = 0.268$; see Fig. 2, Table 2).

To further characterize the observed interaction and investigate the effect of latent SRs depending on their function state, we compared the congruency effect, benefit, and cost in error rates between Maintain and Prepare conditions, since in both conditions all four SRs needed to be held in WM up to Probe 2. We found a significantly larger congruency effect ($t_{104} = 3.53$, $p < .001$, Cohen's $d = 0.40$, $BF_{10} = 67.90$) and a larger cost ($t_{104} = 3.08$, $p = .001$, Cohen's $d = 0.37$, $BF_{10} = 18.07$) while preparing compared to merely maintaining the irrelevant SRs. We also observed a larger benefit in the Prepare condition, although the Bayes Factor revealed inconclusive evidence for this contrast ($t_{104} = 1.85$, $p = .033$, Cohen's $d = 0.24$, $BF_{10} = 1.13$).

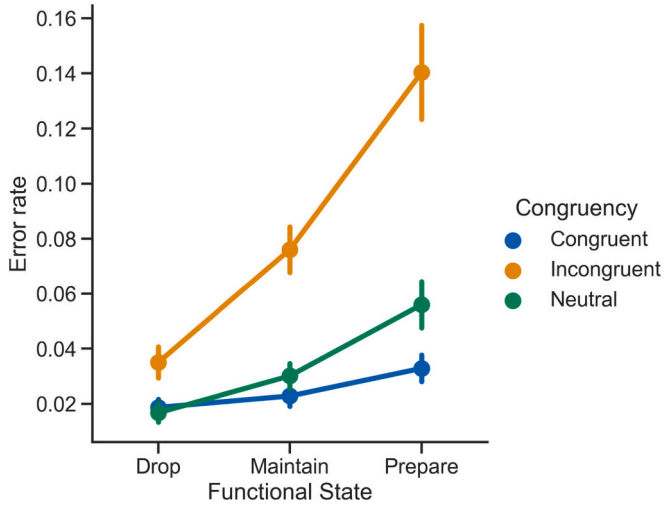

Fig. 2. Error rates results.
Error rates in Probe 1 as a function of the Functional State and the Congruency of the irrelevant feature. Error bars indicate s.e.m.

Table 2
Descriptive statistics on error rates.

Functional State	Congruency	Mean	SD
Drop	Congruent	0.019	0.031
	Neutral	0.017	0.035
	Incongruent	0.035	0.058
Maintain	Congruent	0.023	0.039
	Neutral	0.030	0.046
	Incongruent	0.076	0.085
Prepare	Congruent	0.033	0.049
	Neutral	0.056	0.086
	Incongruent	0.140	0.175

The table contains mean error rates and corresponding standard deviations for each experimental condition.

3.1.2. Drift diffusion modeling

What cognitive processes underlie these effects? To better characterize the observed findings in the error rates analyses, we performed a drift diffusion modeling analysis. After evaluating model fit with DIC and MSE metrics (see Methods), the best model was the one where both Functional State and Congruency were mapped onto drift rate. We found a main effect of Functional State on drift rate, $b = -0.33$, 95% HDI = [−0.42, −0.23], $pd = 100\%$, showing that the rate of evidence accumulation was lower in the Prepare condition relative to the Maintain condition. We also found a main effect of congruency, $b = -0.35$, 95% HDI = [−0.45, −0.24], $pd = 100\%$, revealing slower evidence accumulation in incongruent trials relative to congruent trials. In contrast to the error rates ANOVA, we did not find evidence for an interaction of Functional State and Congruency, $b = 0.05$, 95% HDI = [−0.08, 0.18], $pd = 76.04\%$, suggesting that the difference in evidence accumulation for congruent and incongruent trials was similar across functional states (Fig. 4).

Further assessment of model fit by posterior predictive checks revealed that while the model adequately captured the correct RT distribution, there was a misfit for the error RT distribution, especially for later quantiles of the error RT distribution, as evidenced also by the large mean square error (MSE) between observed and predicted values of these quantiles (Fig. 5). Importantly, this observation was constant across different models. We further simulated data for all conditions (based on the mean of the posterior parameter estimates) and qualitatively compared the predicted and observed RT distributions. This comparison again suggested a poorer fit of the model for slow errors.

3.2. Exploratory analyses

3.2.1. Distributional analyses

While we found a significant interaction between Functional State and Congruency in the repeated measure ANOVA on ER, we did not observe an interaction in the drift rate estimates of the winning model in the modeling analysis. However, a thorough assessment of model fit revealed a marked misfit of our modeling approach in slow errors. Arguably, these findings might be informative with respect to how concurrent SRs are maintained in WM. More specifically, they raise the intriguing possibility that the degree of interference of irrelevant SRs increases as more time passes between the retro-cue and the response to Probe 1. To explore further in this direction, we decided to perform distributional analyses to better understand the nature of interference in the current task (Burle, Spieser, Servant, & Hasbroucq, 2014; Dittrich et al., 2014, see methods for details). We conducted a Bayesian Linear Mixed Effects model predicting the accuracy delta (difference in errors between incongruent and congruent trials) by RT quintile (i.e. response speed) and Functional State. Parallel to the ANOVA results, we first observed a difference between Functional States, $b = -0.03$, 95% HDI = [−0.04, −0.01], $pd = 99.99\%$, showing overall more interference on the Prepare relative to the Maintain condition. Interestingly, we also found an effect of quintile, $b = 0.03$, 95% HDI = [0.02, 0.04], $pd = 99.99\%$,

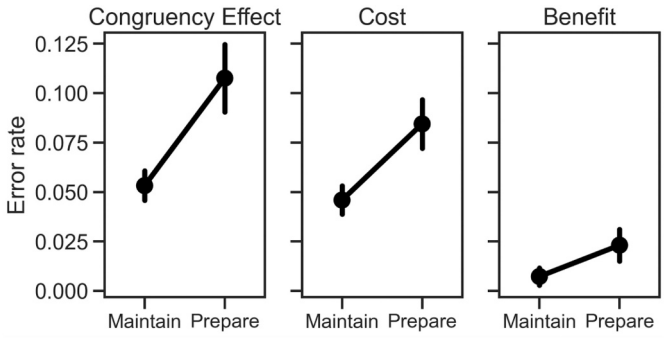

Fig. 3. Congruency effect, cost, and benefit on error rates across Maintain and Prepare Functional States.
Error bars represent s.e.m.

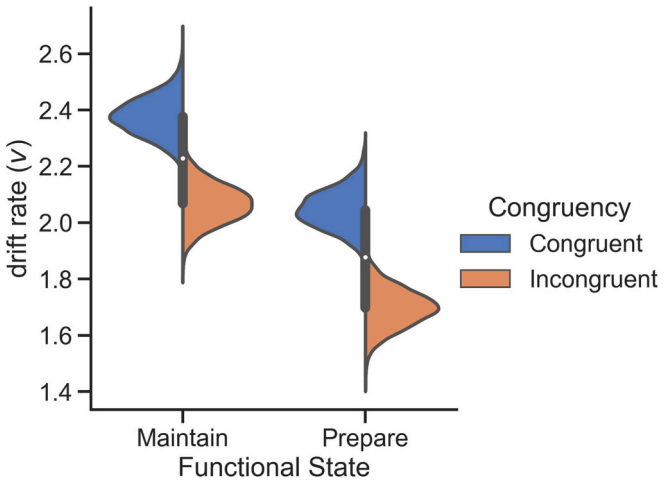

Fig. 4. Drift diffusion model results.
Thick black boxes show the quartiles of the dataset, separately for Maintain and Prepare conditions. The inner white dots denote the means, and the whiskers represent the 1.5 interquartile range. The distributions of values for each condition are depicted by means of density kernels.

revealing that the interference effect in accuracy increased with response latency (Figure 6, left panel). Last, we did not find an interaction between quintile and Functional State, $b = 0.003$, 95% HDI = [−0.02, 0.01], $pd = 69.31\%$. These analyses confirm the larger interference effect for the Prepare condition we observed with the traditional ANOVA and the congruency effect analyses (Figs. 2 and 3). Importantly, they further reveal that the interference effect becomes larger with later response times and in similar fashion for both functional states. Given that these analyses are performed during Probe 1, when the uncued SR is (at least, likely, in the Maintain condition) to become immediately relevant, we interpret this pattern as the result of the increasing

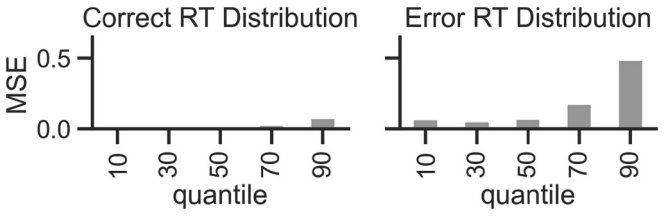

Fig. 5. HDDM model fit.
Misfit between observed and predicted (i.e., simulated based on retrieved model parameters) mean RTs for the error and the correct trials distributions, separately by quantiles. Larger MSE values indicate a larger discrepancy between the observed and predicted data.

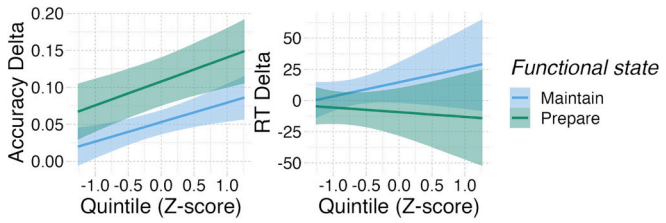

Fig. 6. Delta plots for accuracy and reaction times.
Predicted accuracy and RT delta (incongruent - congruent) as a function of quintile (i.e. response speed) for Maintain and Prepare conditions. Shaded areas indicate the 95% credibility intervals of the predicted values at each quintile.

relevance of the uncued SR as time passes during Probe 1 (van Ede, Deden, & Nobre, 2021).

When using RT delta as dependent measure (Fig. 6, right panel), we did not find an effect of quintile, $b = 3.53$, 95% HDI = [−8.44, 15.54], $pd = 71.38\%$, showing that the interference effect in RT was not modulated by response latency. Additionally, we did not find a difference between Functional States, $b = 11.91$, 95% HDI = [−1.17, 25.17], $pd = 96.28\%$, nor an interaction between quintile and Functional States, $b = 7.42$, 95% HDI = [−5.80, 20.10], $pd = 86.99\%$.

3.2.2. *Reaction times analysis*

For completeness, we carried out in an exploratory fashion the same 3 × 3 repeated-measures ANOVA we used with error rates, now with RT (Fig. 7). Similar to the error rates analysis, this analysis yielded significant differences across Functional States ($F_{1.84,\ 189.34} = 57.44$, $p < .001$, $\eta p2 = 0.36$, $BF_{10} > 150$), with fastest responses for Drop, followed by Maintain and Prepare conditions. Congruency level also significantly affected RTs ($F_{1.87,\ 192.45} = 8.19$, $p < .001$, $\eta p2 = 0.07$, $BF_{10} = 16.23$): Neutral probes yielded the fastest responses compared to Congruent ($t = 3.01$, $p = .005$, Hedges' $g = 0.14$, $BF = 7.58$) and Incongruent trials ($t = 3.11$, $p = .005$, Hedges' $g = 0.16$, $BF = 9.84$), probably reflecting a benefit in speed of having to process only one salient feature vs. two. Finally, the ANOVA did not yield a significant interaction between the factors Functional State and Congruency ($F < 1$, $BF_{10} = 0.02$). Descriptive statistics are reported in Table 3.

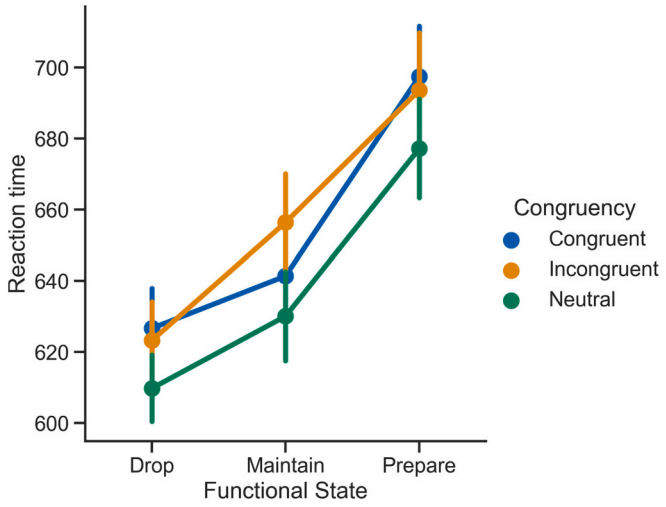

Fig. 7. Reaction times results.
Reaction times in Probe 1 as a function of the Functional State and the Congruency of the irrelevant feature. Error bars indicate s.e.m.

Table 3
Descriptive statistics on reaction times.

Functional State	Congruency	Mean (ms)	SD
Drop	Congruent	626.691	115.250
	Neutral	609.406	96.307
	Incongruent	623.503	111.932
Maintain	Congruent	641.812	130.409
	Neutral	630.148	129.001
	Incongruent	657.162	141.179
Prepare	Congruent	698.478	145.810
	Neutral	675.845	142.303
	Incongruent	693.615	164.562

4. General discussion

In the current study, we aimed at investigating, first, whether internal attention modulates the functional state of novel SRs in WM, and second, how the functional state of these impacts ongoing behavior. To this goal, we devised a paradigm in which participants had to encode 4 novel SRs on each trial and respond to two subsequent targets accordingly. Crucially, immediately after the encoding screen, a retro-cue indicated the prospective relevance of two sets of stimulus features, effectively tagging two of the SRs as completely irrelevant (i.e. Drop condition), likely relevant (i.e., Maintain condition), or certainly relevant (i.e. Prepare condition) for the second target. In line with our predictions, we observed a progressive impairment of ongoing performance as a function of the functional state of the latent SRs. Moreover, the increase of such impairment was even more pronounced when the relevant (i.e. active) and irrelevant (i.e. latent) features of the target lead to incompatible responses, suggesting that the representation of different SRs in WM becomes progressively overlapping as their functional relevance increases.

Previous research on visual WM has extensively reported on the relationship of active vs latent states (Christophel, Iamshchinina, Yan, Allefeld, & Haynes, 2018; Iamshchinina et al., 2021; Panichello & Buschman, 2021; van Loon et al., 2018; Yu, Teng, & Postle, 2020). In particular, retro-cues have been shown to modulate the relevance of cued and uncued items via internal attention (Griffin & Nobre, 2003; Souza & Oberauer, 2016). Whereas attention reconfigures cued items into an "active" state capable of driving behavior, items that must be kept in WM but are irrelevant for ongoing behavior can enter a dormant, "latent" state. Here, we show that, similarly, internal attention towards novel planned actions can modulate their functional state within WM. Specifically, we observed a marked effect of the future relevance of the latent SRs on all our measures of interest (error rates, RTs, and hDDM parameters). In other words, we extend on previous findings by highlighting how latent SRs exert a different degree of interference on ongoing behavior based on how likely they are to be executed in the future. Additionally, our results show that such strategic functional modulation can also take place retrospectively on maintained items, after an initial encoding stage in which participants are agnostic regarding which mappings will be relevant and when.

Previous studies with novel instructions have reported that novel SRs in WM are subjected to strategic control, so that the expected readout (i.e., recognition vs execution) defines whether the item is maintained declaratively or procedurally (Liefooghe et al., 2012; Meiran et al., 2015a). Additionally, the likelihood of newly encoded SRs being implemented, as opposed to recalled, impacts the format in which participants represent them, and thus induces a different degree of interference on ongoing behavior (Whitehead & Egner, 2018b). Declarative and procedural WM have been traditionally considered independent in terms of cognitive resources and functioning (Gade, Druey, Souza, & Oberauer, 2014; Oberauer, 2010), suggesting a strictly dichotomous scenario, in which only proceduralized and active items exert an influence on ongoing behavior. According to this perspective, in the current paradigm, one would predict no difference between Maintain and Drop conditions, as these are equivalent with respect to the number of proceduralized SRs. However, the pattern of results we observe suggests that maintaining latently a set of SRs for a second probe (i.e., Maintain condition) has a detrimental effect on the execution of the prioritized actions compared to entirely dropping them from WM. Relatedly, there is evidence that novel SRs maintained declaratively for subsequent recognition interfere with the execution of currently relevant proceduralized SRs (Formica et al., 2020). Therefore, the observed pattern of results could be accommodated within the traditional dichotomous representational framework, assuming that latent declarative items have the capacity of interfering with ongoing behavior. Alternatively, we propose that latent items are shifted towards a more proceduralized state on a declarative-to-procedural continuous representational space based on their assigned functional relevance, akin to recent proposals on prioritization of SRs in WM (Whitehead and Egner, 2018b). It is worth pointing out that when prioritized in WM, novel planned actions are assumed to enter an active, proceduralized, and behavior-guiding state (González-García et al., 2020; Myers et al., 2017). In this context, the labels of "active" and "latent" are to a certain extent overlapping with "procedural" and "declarative", as intended in the WM model proposed by Oberauer (2009). Further research should investigate deeper the structure of the representational space for actions in WM, to conclusively adjudicate between a continuous or dichotomous organization.

It might appear counterintuitive that the condition enforcing the highest degree of preparation (i.e., Prepare) is also the one showing a larger interfering effect of the currently irrelevant SRs. However, this is likely due to the fact that once tagged as certainly relevant for the second probe, the currently irrelevant task set enters a state of prioritization that has a cost in terms of cognitive load during the first probe. On the contrary, when the second set of SRs is simply maintained, its degree of prioritization is lower, resulting in more uncertainty about what will happen in the second probe, but also lower cognitive load during the first probe. From this perspective, interference is tied more to the degree of prioritization of the latent items, rather than the degree of uncertainty on what feature will be relevant for the second probe. Moreover, the current task was optimally designed to test the specific question of how concurrently held sets of SRs influence ongoing behavior depending on their degree of prioritization and prospective relevance, thus we focused our analyses on performance in response to Probe 1. We speculate that overall better preparation would result in overall better performance (i.e., across both Probes), but this assumption cannot be tested within the current experiment as Maintain and Prepare conditions crucially differ with respect to the cueing of Probe 2, rendering comparisons uninterpretable. This question warrants follow up studies to elucidate the benefits of thorough preparation.

Notably, our findings further inform the ongoing debate concerning the representational coding scheme of active and latent items in WM. We observed the hypothesized interaction between functional state and congruency of the two features presented in the first bivalent target. While an absence of interference from the latent items would have hinted at a perfect orthogonalization (i.e., independence and separation) of the two functional states, the results from the current study favor an alternative scenario. The observed interference, in contrast, points towards a non-orthogonal (i.e., non-independent) coding scheme, indicating that active and latent content is not fully segregated. This setup allows items to flexibly and rapidly change their state from active to latent, but such flexibility comes at the cost of effective shielding (Stokes et al., 2020). This hypothesized coding scheme, referred to as Attentional Gain Coding, assumes the distinction between active and latent functional states to be quantitative, rather than qualitative, thus implying that the degree of prioritization to which an item is subjected can raise its activation level to the extent of being tagged as active and relevant for ongoing behavior. Crucially, this coding scheme predicts that the degree of crosstalk between active and latent items, and thus the extent to which latent representations can affect ongoing WM processing, depends on the relative activation strength of these latent items, as

we observe in our results. This hypothesis proved to be true for error rates, suggesting that such attentional gain coding impacts primarily the quality of the WM representations.

To further support this interpretation, we fit our data with drift diffusion modeling. The goal of this analysis was to identify which decision parameters were predominantly affected by our manipulations, and thus adjudicate between alternative explanations for the cognitive mechanisms involved in SRs prioritization. On one hand, lower performance in the Prepare condition might be attributed to a delay in accessing the correct item for decision making, resulting in larger latencies for the onset of evidence accumulation. This account, known as the retrieval head-start hypothesis, implies that differences between functional states should be reflected by the non-decision time parameter. On the other hand, our Functional state manipulation might impact how effectively the available information is used as a decision variable. This refers to the so-called matched filter hypothesis and assumes the variance across conditions to be captured by the drift rate (Muhle-Karbe et al., 2021). In line with the latter alternative, we found the drift rate to be significantly reduced in the Prepare compared to Maintain condition. This finding indicates that the quality of the information entering the decision process is lower as the latent representations become more relevant, resulting in a reduced rate of evidence accumulation. Analogously, also the effect of Congruency was captured by variance in the drift rate, again suggesting that incongruent trials provide representations of lower quality for the evidence accumulation process.

Interestingly, a more detailed analysis of the distributions of errors suggested that the best DDM parameters to model our empirical data failed to successfully account for slow errors. Such distributional analysis confirmed, first, the higher error rate for Prepare than Maintain, but also crucially it revealed that the detrimental effect of the incongruent feature dimension significantly increased as a function of RT, leading to a larger number of errors in both the Maintain and Prepare conditions for slow responses. While fast errors can be attributed to automatic, bottom up processing of the irrelevant stimulus dimension (Ulrich et al., 2015), the nature of slow errors is open to more interpretations. Recent accounts propose that slow errors are caused primarily by the low quality of evidence entering the accumulation process (Damaso, Williams, & Heathcote, 2020). In this particular case, one intriguing possibility is that the heightened level of prioritization of currently relevant representations (i.e., needed to respond to the first probe) wanes over time, relaxing the degree of shielding (i.e., the separation between competing representations). Moreover, given that in our paradigm latent items during the first probe are likely to become relevant immediately after (that is, taking more time to answer to the first probe implies both longer maintenance of the active representations and increased relevance of the latent ones), it is possible that the decay of the active SRs comes with a simultaneous increasing prioritization of the latent items. Such dynamics could explain, first, the increased congruency effect in accuracy as a function of RTs, and second, the observed larger effect in Prepare trials, where latent items are 100% likely to become immediately relevant after answering to the first probe. Moreover, these dynamics fit well within the previously mentioned attentional gain coding scheme, where competing representations are separated in one dimension, but the approaching of the response deadline for Probe 1 might cause an increase of activation for the latent item, resulting in larger interference with longer RTs (Murphy, Boonstra, & Nieuwenhuis, 2016).

To what extent can the current results be explained by alternative mechanisms, such as memory swaps (i.e., the wrong item from the memory set is retrieved instead of the probed one; Bays, 2014, 2016) or action slips (i.e., the unintentional implementation of an action plan in WM in place of the intended one; Miller, Kiyonaga, Ivry, & D'Esposito, 2020)? Although it is hard to rule these out in our task given the dichotomous nature of the responses, the observed pattern of results, such as the fact that errors increase with time, is coherent with other reports of "slow errors" (Murphy et al., 2016; van Maanen, Katsimpokis, & van Campen, 2019) and can be rather interpreted as a consequence of noisy representations (Ma, Husain, & Bays, 2014) and control limitations (Braem et al., 2019). Whereas swaps or action slips could induce fast errors due to response capture of the incorrect feature, our results could reflect a consequence of increasing time pressure or urgency to respond as the response deadline and, therefore, the onset of the second probe, approaches. Accordingly, increasing the relevance of latent SRs would come at the expense of overall noisier representations. Such noisier representations would induce slower accumulation of evidence for the decision (as revealed by the DDM) and, in turn, more chances of approximating the response deadline. According to Murphy et al. (2016), the generation of urgency induced by an approaching deadline (plus upcoming Probe 2 in our task) could be induced by a global modulation of neural gain. Arguably, the increasing interference as a function of response times we report here could be explained by such boosting of the latent representations. In turn, this interpretation again suggests that active and latent SRs are maintained in non-orthogonal neural codes, so that increasing the activation of the latent representation induces higher interference on current behavior. Importantly, although the current results allow us to speculate about how planned actions are maintained in WM, the correspondence between their functional states and their neural coding scheme is largely unexplored. Studies that combine behavioral paradigms similar to the current one with neuroimaging recordings are thus needed to test such correspondence empirically.

Altogether, our results show that novel SRs can be retrospectively tagged with specific future relevance within WM, inducing differentiated functional states. The current results are in line with the idea of representations of novel SRs held in WM with a functional state ranging on a continuum of "activeness", that can be dynamically adjusted depending on goals and task demands during the course of a trial. However, further research should confirm this view and rule out the alternative of a dichotomous representational space. Furthermore, our results provide initial evidence that representations with different functional states are maintained via a non-orthogonal coding scheme. More specifically, our results are consistent with the idea of an attentional gain coding mechanism, where attentional prioritization of planned lines of action would boost the activation of the corresponding representations at the cost of greater impact on ongoing behavior.

Author contributions

S.F., C.G.-G., A.P., N.M., and M.B. designed the research; S.F., C.G.-G., and A.P. performed the research; S.F., C.G.-G., A.P., and L.V., analyzed the data; S.F., C.G.-G., and A.P. wrote a first draft of the manuscript, all authors edited and revised the final paper.

Funding

SF was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Germany's Excellence Strategy-EXC 2002/1, Science of Intelligence (Project Ref.: 390523135) and the Einstein Foundation Berlin. AP was funded by the Andalusian Autonomic Government (Grant Ref.: PAIDI 21_00207). LV was supported by the Research Foundation - Flanders (FWO-Vlaanderen, Project Ref: 11H5619N). MB is supported by an Einstein Strategic Professorship (Einstein Foundation Berlin). CGG was supported by Project PID2020-116342GA-I00 funded by MCIN/AEI/10.13039/501100011033, and Grant RYC2021-033536-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRT.

CRediT authorship contribution statement

Silvia Formica: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Ana F. Palenciano:** Conceptualization, Formal analysis, Methodology,

Writing – original draft, Writing – review & editing, Investigation. **Luc Vermeylen:** Formal analysis, Methodology, Writing – review & editing. **Nicholas E. Myers:** Conceptualization, Writing – review & editing. **Marcel Brass:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Carlos González-García:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Data availability

link is available in the manuscript

References

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience, 34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports, 6*(1), Art. 1. https://doi.org/10.1038/srep19203

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences, 23*(9), 769–783. https://doi.org/10.1016/j.tics.2019.07.002

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience and Biobehavioral Reviews, 81*(June), 16–28. https://doi.org/10.1016/j.neubiorev.2017.02.012

Burkner, P.-C. (2015). *brms: An R Package for Bayesian Generalized Linear Mixed Models using Stan* (p. 2013). Plummer.

Burle, B., Spieser, L., Servant, M., & Hasbroucq, T. (2014). Distributional reaction time properties in the Eriksen task: Marked differences or hidden similarities with the Simon task? *Psychonomic Bulletin & Review, 21*(4), 1003–1010. https://doi.org/10.3758/s13423-013-0561-6

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(1), 39–50. https://doi.org/10.1080/10705511.2018.1490648

Carlisle, N. B., & Woodman, G. F. (2019). Quantifying the attentional impact of working memory matching targets and distractors. *Visual Cognition, 27*(5–8), 452–466. https://doi.org/10.1080/13506285.2019.1634172

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J. D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience, 21*(4), 494–496. https://doi.org/10.1038/s41593-018-0094-4

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Damaso, K., Williams, P., & Heathcote, A. (2020). Evidence for different types of errors being associated with different types of post-error changes. *Psychonomic Bulletin & Review, 27*(3), 435–440. https://doi.org/10.3758/s13423-019-01675-w

Dittrich, K., Kellen, D., & Stahl, C. (2014). Analyzing distributional properties of interference effects across modalities: Chances and challenges. *Psychological Research, 78*(3), 387–399. https://doi.org/10.1007/s00426-014-0551-y

van Driel, J., Ort, E., Fahrenfort, J. J., & Olivers, C. N. L. (2019). Beta and theta oscillations differentially support free versus forced control over multiple-target search. *Journal of Neuroscience, 39*(9), 1733–1743. https://doi.org/10.1523/JNEUROSCI.2547-18.2018

van Ede, F., Deden, J., & Nobre, A. C. (2021). Looking ahead in working memory to guide sequential behaviour. *Current Biology, 31*(12), R779–R780. https://doi.org/10.1016/j.cub.2021.04.063

Formica, S., González-García, C., & Brass, M. (2020). The effects of declaratively maintaining and proactively proceduralizing novel stimulus-response mappings. *Cognition, 201*, Article 104295. https://doi.org/10.1016/j.cognition.2020.104295

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology, 20*(2), 177–182. https://doi.org/10.1016/j.conb.2010.03.005

Gade, M., Druey, M. D., Souza, A. S., & Oberauer, K. (2014). Interference within and between declarative and procedural representations in working memory. *Journal of Memory and Language, 76*, 174–194. https://doi.org/10.1016/j.jml.2014.07.002

Gamerman, D., & Lopes, H. F. (2006). Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, second edition. In *, Vol. 1. Texts in statistical science.* Taylor & Francis.

González-García, C., Formica, S., Liefooghe, B., & Brass, M. (2020). Attentional prioritization reconfigures novel instructions into action-oriented task sets. *Cognition, 194*, Article 104059. https://doi.org/10.1016/j.cognition.2019.104059

González-García, C., Formica, S., Wisniewski, D., & Brass, M. (2021). Frontoparietal action-oriented codes support novel instruction implementation. *NeuroImage, 226*, Article 117608. https://doi.org/10.1016/j.neuroimage.2020.117608

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience, 15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Henderson, M. M., Rademaker, R. L., & Serences, J. T. (2022). Flexible utilization of spatial- and motor-based codes for the storage of visuo-spatial information. *ELife, 11*, Article e75688. https://doi.org/10.7554/eLife.75688

Iamshchinina, P., Kaiser, D., Yakupov, R., Haenelt, D., Sciarra, A., Mattern, H., … Cichy, R. M. (2021). Perceived and mentally rotated contents are differentially represented in cortical depth of V1. *Communications Biology, 4*(1). https://doi.org/10.1038/s42003-021-02582-4. Art. 1.

Jasp Team. (2019). JASP (version 0.11.1)[computer software]. https://jasp-stats.org/.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software, 8*(85), 5351. https://doi.org/10.21105/joss.05351

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods, 49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

Li, A. Y., Liang, J. C., Lee, A. C. H., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *Journal of Experimental Psychology: General, 149*, 949–966. https://doi.org/10.1037/xge0000693

Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1325–1335. https://doi.org/10.1037/a0028148

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*, 1–25. https://doi.org/10.7554/elife.38677

van Loon, A. M., Olmos-Solis, K., & Olivers, C. N. L. (2017). Subtle eye movement metrics reveal task-relevant representations prior to visual search. *Journal of Vision, 17*(6), 13. https://doi.org/10.1167/17.6.13

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for featuresand conjunctions. *Nature, 390*(1996), 279–281.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356. https://doi.org/10.1038/nn.3655

van Maanen, L., Katsimpokis, D., & van Campen, A. D. (2019). Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships. *Behavior Research Methods, 51*(5), 2378–2389. https://doi.org/10.3758/s13428-018-1110-z

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology, 10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Mallett, R., & Lewis-Peacock, J. A. (2018). Behavioral decoding of working memory items inside and outside the focus of attention. *Annals of the New York Academy of Sciences, 1424*(1), 256–267. https://doi.org/10.1111/nyas.13647

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015a). Reflexive activation of newly instructed stimulus–response rules: Evidence from lateralized readiness potentials in no-go trials. *Cognitive, Affective, & Behavioral Neuroscience, 15*(2), 365–373. https://doi.org/10.3758/s13415-014-0321-8

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015b). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 768–786. https://doi.org/10.1037/xlm0000063

Miller, J. A., Kiyonaga, A., Ivry, R. B., & D'Esposito, M. (2020). Prioritized verbal working memory content biases ongoing action. *Journal of Experimental Psychology: Human Perception and Performance, 46*, 1443–1457. https://doi.org/10.1037/xhp0000868

Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2021). A hierarchy of functional states in working memory. *The Journal of Neuroscience, 41*(20), 4461–4475. https://doi.org/10.1523/jneurosci.3104-20.2021

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications, 7*(1), Art. 1. https://doi.org/10.1038/ncomms13526

Myers, N. E., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Benefits of flexible prioritization in working memory can arise without costs. *Journal of Experimental Psychology: Human Perception and Performance, 44*(3), 398–411. https://doi.org/10.1037/xhp0000449

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences, 21*(6), 449–461. https://doi.org/10.1016/j.tics.2017.03.010

Nobre, A. C., & Stokes, M. G. (2019). Premembering experience: A hierarchy of time-scales for proactive attention. *Neuron, 104*(1), 132–146. https://doi.org/10.1016/j.neuron.2019.08.030

Nobre, A. C., & Van Ede, F. (2018). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience, 19*(1), 34–48. https://doi.org/10.1038/nrn.2017.141

Oberauer, K. (2009). Design for a working memory. In *, Vol. 51. Psychology of learning and motivation* (pp. 45–100). Academic Press. https://doi.org/10.1016/S0079-7421(09)51002-X.

Oberauer, K. (2010). Declarative and procedural working memory: Common principles, common capacity limits? *Psychologica Belgica, 50(3 &, 4*, 277–308. https://doi.org/10.5334/pb-50-3-4-277

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

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356.

Ratcliff, R., Smith, P. L., Brown, S. D., & Mckoon, G. (2016). Feature review diffusion decision model: Current issues and history. *20*(4). https://doi.org/10.1016/j.tics.2016.01.007

Shepherdson, P., Oberauer, K., & Souza, A. S. (2018). Working memory load and the retro-cue effect: A diffusion model account. *Journal of Experimental Psychology: Human Perception and Performance, 44*(2), 286–310. https://doi.org/10.1037/xhp0000448

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics, 78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron, 91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

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

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology, 78*, 148–174. https://doi.org/10.1016/J.COGPSYCH.2015.02.005

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software, 3*(31), 1026. https://doi.org/10.21105/joss.01026

de Vries, I. E. J., Slagter, H. A., & Olivers, C. N. L. (2019). Oscillatory control over representational states in working memory. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.11.006

de Vries, I. E. J., van Driel, J., Karacaoglu, M., & Olivers, C. N. L. (2018). Priority switches in visual working memory are supported by Frontal Delta and posterior alpha interactions. *Cerebral Cortex, 28*(11), 4090–4104. https://doi.org/10.1093/cercor/bhy223

de Vries, I. E. J., van Driel, J., & Olivers, C. N. L. (2017). Posterior α EEG dynamics dissociate current from future goals in working memory-guided visual search. *The Journal of Neuroscience, 37*(6), 1591–1603. https://doi.org/10.1523/JNEUROSCI.2945-16.2016

White, C. N., Servant, M., & Logan, G. D. (2017). Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1271-2

Whitehead, P. S., & Egner, T. (2018a). Cognitive control over prospective task-set interference. *Journal of Experimental Psychology. Human Perception and Performance, 44*(5), 741–755. https://doi.org/10.1037/xhp0000493

Whitehead, P. S., & Egner, T. (2018b). Frequency of prospective use modulates instructed task-set interference. *Journal of Experimental Psychology: Human Perception and Performance, 44*(12), 1970–1980. https://doi.org/10.1037/xhp0000586

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics, 7*(August), 1–10. https://doi.org/10.3389/fninf.2013.00014

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS Biology, 18*(6), Article e3000769. https://doi.org/10.1371/journal.pbio.3000769