Understanding subprocesses of working memory through the lens of model-based cognitive neuroscience

Trutti, A.C.; Verschooren, S.; Forstmann, B.U.; Boag, R.J.

DOI

Publication date
2021

Document Version
Final published version

Published in
Current Opinion in Behavioral Sciences

License
CC BY

Citation for published version (APA):
Understanding subprocesses of working memory through the lens of model-based cognitive neuroscience
Anne C Trutti1,2, Sam Verschooren3, Birte U Forstmann1 and Russell J Boag1

Working memory (WM) refers to a set of processes that makes task-relevant information accessible to higher-level cognitive processes. Recent work suggests WM is supported by a variety of information gating, updating, and removal processes, which ensure only task-relevant information occupies WM. Current neurocomputational theory suggests WM gating is accomplished via ‘go/no-go’ signalling in basal ganglia-thalamus-prefrontal cortex pathways, but is less clear about other subprocesses and brain structures known to play a role in WM. We review recent efforts to identify the neural basis of WM subprocesses using the recently developed reference-back task as a benchmark measure of WM subprocesses. Targets for future research using the methods of model-based cognitive neuroscience and novel extensions to the reference-back task are suggested.

Addresses
1 Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
2 Institute of Psychology, Leiden University, Leiden, The Netherlands
3 Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

Corresponding author:
Trutti, Anne C (r.j.boag@uva.nl), Boag, Russell J (r.j.boag@uva.nl)

Prominent computational theories suggest that WM resolves the stability-flexibility trade-off by operating in two modes: An updating (gate-open) mode, which allows new information to enter WM, and a maintenance (gate-closed) mode, which prevents irrelevant and distracting information from interfering with the current contents of WM [15–22]. In the gate-open mode, updating is further supported by two main subprocesses: Item removal and item substitution, which together ensure that only relevant information is kept active in WM [23,24]. Together, these processes allow WM to alternate modes between flexible (when new information is encountered) and stable (when distractors are encountered). This enables successful performance in dynamic environments in which distractions are common and the relevance of information frequently changes.

To-date, the most detailed neurocomputational account of the gating mechanism controlling the trade-off between updating and maintenance is the prefrontal

---

4 We use the term 'item' to refer to an individual representation held in WM. ‘Item’ is thus synonymous with ‘chunk’ [97] and ‘cognitive object’ [98,7] which denote the same concept. There is ongoing debate about whether items in WM are represented in discrete slots (items held with high precision in a number of discrete memory locations), allocation of continuous resources (items allocated limited resources in inverse proportion to the total number of items in WM), or some hybrid of the two frameworks (e.g. Refs. [78,74,79,80]). The models and general approach that we discuss in this paper are not committed to either architecture but could be used to test between the competing accounts (see Section ‘Current directions’ below).
The components of the PBWM model have received broad support from functional magnetic resonance imaging (fMRI) studies ([19,28,31–33,22,34]). For example, activation in striatum and dorsolateral PFC has been widely reported in tasks broadly involving WM updating (e.g. Refs. [31–34]), while other work has localized activity specifically related to the updating and gating processes rather than other WM processes. Roth et al. [22] identified a frontoparietal network specifically involved in updating, while Murty et al. [19] found selective engagement of SN/ventral tegmental area (VTA), caudate, dorsolateral PFC, and some areas of parietal cortex related to the updating but not maintenance mode of WM. Striatal dopamine-receptor expressing neurons and dopamine-producing midbrain structures have also been implicated in WM updating [19,28,33], and dynamic causal modelling suggests that BG plays a central role in gating information to PFC [35]. Moreover, a number of cortical areas (e.g. dorsolateral PFC, medial PFC, posterior parietal cortex) have been linked to the maintenance mode of WM but not updating ([22,36–38]). This is consistent with the idea that tonic dopamine activity in PFC controls the stability of WM representations whereas phasic dopamine release in the striatum trains the BG when to open the gate (via disinhibition of thalamus and PFC) to allow information into WM [27].

Overall, these findings show that WM updating engages cortico-striatal circuitry involving BG, midbrain, and PFC structures broadly in line with the neurocomputational mechanisms of the PBWM model [39,26] and more general accounts of cognitive control (e.g. Ref. [40]). However, as will be discussed, recent work highlights that WM also depends on several important subprocesses not accounted for in the PBWM, and on neural substrates outside of the PBWM’s BG-thalamus-PFC pathways. Modelling these processes and their neural basis is necessary to achieve a complete neurocomputational understanding of WM.

This review discusses recent progress toward this goal. We focus on recent efforts to link brain measurements with behaviour on the reference-back task (Figure 2; [24*,21*]), a WM-based decision-making task that provides separate behavioural measures of gate opening and closing, as well as updating and substitution processes not accounted for in the PBWM. In doing so, we suggest that

---

3 The PBWM model assumes that WM sits in the ‘gate-closed’/maintenance mode by default. We note that this assumption is likely too strong, since it implies that gate opening must always accompany updating. Under this assumption the PBWM would fail to predict the different gating costs to WM updating that occur in behavioural data (e.g. Refs. [24*,21*]).
further progress can be made by applying the methods of *model-based cognitive neuroscience* [41,42*], which links brain activity to behaviour via detailed computational models of cognitive and neural processes [43–45]. Model-based cognitive neuroscience generates detailed quantitative theories that span multiple levels of abstraction (e.g. behavioural, cognitive, neural). This provides greater constraint on theory and leads to more robust and detailed inferences. In particular, combining model-based approaches with developments in ultra-high field fMRI enables testing neurocomputational theories of WM (such as the PBWM) with greater spatial and psychometric precision than has previously been possible. Applying these methods to the reference-back task promises a more detailed neurocomputational understanding of WM than is currently available.

**Measuring WM subprocesses with the reference-back paradigm**

Most laboratory tasks used to study WM (e.g. n-back, delayed-match-to-sample) are designed to investigate the capacity and temporal properties of WM but are unable to differentiate the contribution of WM subprocesses to observed behaviour ([24**,46,47**,48*,21*,22]). A recently developed exception is the *reference-back* task [24**,21*], which provides dissociable measures of core WM subprocesses (gate opening, gate closing, updating, substitution) from behavioural choice-response time (RT) data.

To perform the reference-back, participants hold one of two stimuli (e.g. an ‘X’ or ‘O’) in WM while deciding whether a series of probes match the current item in WM (Figure 2). On *reference trials* (indicated by a red frame around the stimulus), the participant must update WM with the currently displayed letter. On *comparison trials* (indicated by a blue frame), the participant simply compares the current stimulus to the one held in WM (the one appearing in the most recent red frame) without updating WM. Both reference and comparison trials require a same/different decision but only reference trials require updating. Comparing performance on reference and comparison trials thus provides a behavioural measure of the cost of *updating*. By similar logic, switching from comparison to reference trials requires opening the WM gate (to allow for updating), while switching from reference to comparison trials requires closing the WM gate (to maintain the current contents). *Gate opening* is measured by comparing trials on which participants switch towards a reference trial to those where reference trials are repeated. Likewise, *gate closing* is measured by comparing trials on which participants switch towards a comparison trial to those where comparison trials are repeated. Finally, *substitution* is measured via the interaction effect of trial type (reference/comparison) and match type (same/different) and represents the cost of updating a *new* item into WM.

The benchmark behavioural finding from the reference-back task is that trials requiring additional WM processes tend to have slower RTs and/or more frequent errors than trials that do not require such processes [24**,47**,49*,21*,50*,51*,52*]. These costs are typically interpreted as reflecting a combination of time required for additional subprocesses to run outside of the same/
different decision stage, and subprocesses interfering with the primary task (e.g. creating noisier WM representations due to drawing attention/capacity away from the decision process) [53]. However, distinguishing these accounts requires detailed choice-RT models of the latent cognitive processes underlying memory-based decision making (e.g. the highly successful evidence accumulation framework, [54,55]), which are yet to be applied to the reference-back paradigm. Before discussing approaches to modelling the reference-back task, we first review recent efforts to identify the neural substrates of WM subprocesses by correlating brain activity with behavioural measures derived from the reference-back.

**Neural correlates of the reference-back task**

As outlined above, there is broad consensus from neuro-imaging supporting the role of BG, thalamus, and PFC in WM gating as instantiated in the PBWM [27]. However, the neural basis of several core WM subprocesses (e.g. gate closing, updating, substitution) is less clear. Recent work has begun to address this gap by linking behavioural measures derived from the reference-back with neuro-physiological measures such as EEG and fMRI [47**,49**,51*,52*].

Two initial studies investigated EEG correlates of the reference-back task. Rac-Lubashovsky and Kessler [51*] found that gate closing was associated with increased theta power, a neural signature of cognitive control [56–58], while gate opening and updating were associated with increased delta power, a signature of reactive (event-driven) control and action selection processes that engage in response to reward prediction errors [59–61]. This suggests a functional role for delta and theta signals in the control of WM consistent with 'go/no-go' signalling in the PBWM model [25,39,26]. A follow-up study explored the role of the P3b EEG signal (a positive event-related potential that signals task-relevant events and peaks 300 ms after stimulus onset) in gating and updating [52*]. P3b amplitude spiked depending on whether the stimulus matched the WM reference item, implicating P3b in stimulus comparison/categorisation processes rather than updating per se. Greater negative activity (in an N2-like ERP component unrelated to the P3b) was found in anterior cortical regions on reference versus comparison trials. This signal has been implicated in controlled inhibition and action selection [62] and, in the context of the reference-back task, likely reflects a gate-opening or updating signal, consistent with the PBWM’s assumption that reference trials trigger an update or ‘go’ signal to allow new information into WM. This initial work demonstrates that neural signatures of specific updating and gating processes are detectable in EEG oscillatory signals that show activity broadly consistent with ‘go/no-go’ signalling in BG-thalamus-PFC pathways involved in WM gating [25,26]. However, the poor spatial resolution of EEG limits our ability to draw conclusions about the specific structures associated with each WM subprocess.

Extending this work, Nir-Cohen et al. [47**] used 3T fMRI to identify neural substrates of WM subprocesses using a modified reference-back with more complex face-morph stimuli. BG, frontoparietal cortex, and task-relevant sensory areas such as visual cortex were involved in gate opening, Gate closing activated parietal cortex and substitution elicited activation in left dorsolateral PFC and inferior parietal lobule. A whole-brain conjunction analysis revealed shared activity in the supplementary motor area for updating and substitution, while updating and gating both activated the posterior parietal cortex. These results broadly agree with the PBWM model [26] and support the role of BG and PFC in controlling the flow of information into WM and replacing old with new information. However, parietal cortex activation during gate closing is not predicted by the PBWM. This suggests that additional brain structures are involved in controlling WM subprocesses and points to an opportunity to extend the PBWM to explain the neural basis of WM subprocesses beyond gate opening.

Jongkees [49*] provided further evidence for the dopaminergic basis of WM gating and updating processes by administering dopamine precursor L-tyrosine to young adults and comparing reference-back performance to a placebo-control group. The L-tyrosine group had less variable gate opening times than placebo controls, suggesting that the drug improved WM performance for poor performers but impaired high performers. There was no effect on updating or gate closing, consistent with the role of striatal dopamine signals in opening the gate to WM in line with the PBWM [25,26]. Further indirect support for striatal dopamine involvement comes from a study linking event-based eye-blink rate (a proxy measure of striatal dopamine) to WM updating in the reference-back task [50*]. However, follow-up work combining this approach with ultra-high field fMRI is needed to identify how activity in small subcortical structures as well as layers in cortex (e.g. striatum, GP, thalamus, PFC) is modulated by dopamine.

**Current directions**

The work reviewed above has taken important first steps toward identifying the neural substrates of WM subprocesses beyond the BG-thalamus-PFC ‘go/no-go’ gating mechanism of the PBWM [39,26]. However, existing work has so far been limited to relating brain activity directly to the reference-back’s behavioural measures rather than the latent cognitive processes that give rise to behaviour. Model-based approaches that link brain and behaviour via computational cognitive models offer numerous advantages over
traditional statistical analyses of mean RT and error rate in understanding the cognitive and neural basis of WM. For example, applying evidence accumulation models of choice-RT (e.g. Refs. [54,55]) to reference-back data would reveal whether performance costs occur because WM subprocesses add time outside of the decision stage (longer nondecision time), interfere with the decision process itself (reduced or noisier processing rate; [53]), or induce strategic adjustments engaging top-down cognitive control (increased response caution). Decomposing behavioural effects (e.g. gating, updating costs) into a set of latent cognitive processes (e.g. accumulation rate, nondecision time, cognitive control of thresholds) rather than coarse behavioural-level summary statistics enables exploring the neural substrates of WM in greater detail than is possible with traditional methods [63,64]. This places stronger constraints on theory and ultimately produces more robust and detailed inferences about the latent processes that generate behaviour. Applying cognitive models to the reference-back holds great promise in this regard.

In its standard form, the reference-back paradigm ignores several important additional WM processes. These include mechanisms that operate on information already active in WM [65–67], such as object selection and retrieval [7], item-specific removal [23]; but see Ref. [68], for evidence of removal in the reference-back), and grouping and reorganization operations (e.g. sorting items into alphabetical or chronological order, chunking or grouping items together to form a single accessible representation, changing the serial position of items; [69–72]). These mechanisms support effective remembering by restructuring information into more memorable formats and ensuring only relevant information is maintained and retrieved from WM. The standard reference-back also ignores phenomena associated with WM’s limited capacity (e.g. WM load/set-size effects; [73–75,7]) and the temporal degradation (e.g. by decay or interference) of WM representations (for a review, see Ref. [76]). Analyses that do not account for these processes risk misattributing their effects to other processes, resulting in biased inferences.

Simple extensions to the reference-back task (e.g. using multiple-item WM sets, inserting delays between the update cue and stimulus presentation), however, enable testing such effects alongside the gating and updating processes of the standard reference-back. For example, Verschooren et al. [77] developed a modified reference-back paradigm where one among several items in long-term memory or perception is gated into WM. This allows for comparing gating dynamics for perceptual versus long-term memory information. Similar multiple-item modifications can be used to investigate some of the WM phenomena described above, including informing the ongoing debate about whether items in WM are held in a small number of discrete high-precision slots [74] or allocated capacity from a limited pool of continuous resources [78–80]. In discrete slots models, the fidelity of items in WM only degrades once all memory slots are full (e.g. when n > 4). In continuous resource models, an item’s fidelity is determined by its share of the available resources and thus should degrade in inverse proportion to the total number of items in WM*. Evidence accumulation models are well suited to test between these competing accounts (e.g. via accumulation rate parameters) as they can be used to assess the fidelity of WM representations and measure capacity-sharing effects; [74,81]). Varying set size in the reference-back and assessing the effects on decision-making and WM processes (as measured by cognitive models) could test between slots and resource architectures. Similarly, combining a multiple-item reference-back task with reinforcement learning (e.g. by reinforcing some items but not others) could shed light on the interplay between WM and learning (e.g. Refs. [73,75]) and the role of expected value in WM-based decisions. Overall, we believe that detailed choice-RT modelling will play an important role in resolving these important questions and in explaining additional WM phenomena captured by variants of the reference-back task.

Combining computational approaches with recent developments in ultra-high field fMRI (7T and higher) (e.g. increased resolution and better signal-contrast-to-noise ratios) holds great promise for identifying activity in small subcortical structures (e.g. GP, SN, subthalamic nucleus, VTA; [82,83]) and gaining a deeper understanding of their functional role in WM than is currently available. For example, this would enable a stronger test of the so-called ‘third phase’ response of the PBWM model [27], which evaluates the updating process via dopaminergic midbrain neurons that code reward prediction errors [84]. Under the PBWM, midbrain dopamine responses that train the BG when to update should no longer occur once updating-related task rules have been learned. This mechanism has proven difficult to verify with low field strength fMRI [85,86], however, imaging reference-back performance with ultra-high field fMRI and linking neural measurements to cognitive model parameters would enable identifying these anatomical and functional mechanisms in greater detail and provide additional constraint on cognitive models of WM. Specifically, when modelling two or more sources of data (e.g. fMRI and choice-RT) simultaneously, the power to detect joint effects (e.g. correlations between BOLD

---

Note, however, that continuous resource models can mimic discrete slots models. For example, if a resource pool has capacity to accommodate four items, then item fidelity may only begin to degrade once demands exceed capacity (i.e. when n > 4), thus producing similar predictions to a discrete slots model. Careful experimental design is needed in order to correctly attribute effects to capacity limitations [99].
Concluding remarks

This review discussed recent efforts to identify the neural basis of subprocesses that support WM in the recently developed reference-back task. Current empirical work supports the idea that WM gating is controlled by striatal ‘go/no-go’ signalling in BG-thalamus-PFC pathways. However, the neural substrates of several additional WM subprocesses are yet to be established, pointing to a need for ultra-high field functional imaging combined with detailed computational cognitive modelling. Targets for future research include extending the reference-back task to account for additional WM subprocesses (e.g. removal, selection, and reorganization) and effects of WM load and capacity (e.g. longer retrieval times, noisier WM representations), as ignoring such processes leads to mis-specified models and potentially biased inferences. Applying the methods of model-based cognitive neuroscience to the reference-back task would provide a major advance in understanding WM at neural, cognitive, and behavioural levels. A comprehensive understanding of WM subprocesses and their neural basis is within reach, with implications for both cognitive and clinical neuroscience.

Author contributions

RJB and ACT conducted the literature review and led the writing of the manuscript. All authors provided feedback at different stages, reviewed, edited, and revised the manuscript.

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank Dr Yoav Kessler for making several suggestions that improved an earlier version of this manuscript. This work was supported by a grant from the Netherlands Organisation for Scientific Research (NWO; grant number 016.Vici.185.052; BUF).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


47. Nir-Cohen G, Kessler Y, Egner T: Distinct neural substrates for opening and closing the gate from perception to working memory. bioRxiv 2019 http://dx.doi.org/10.1101/853630. Used 3T fMRI to identify neural substrates of WM subprocesses using the reference-back paradigm. This work identifies brain regions involved in gate opening, gate closing, updating, and substitution processes, broadly consistent with the PBWM model in which BG-thalamus-PFC pathways control the flow of information into and out of WM.


and comparing reference-back performance to a placebo-control group. Updating and gating were differentially affected by the dopaminergic manipulation, highlighting the importance of the brain’s dopamine systems for controlling WM.


77. Verschooren S, Kessler Y, Egner T: Evidence for a single mechanism gating perceptual and long-term memory information into working memory. PsyArXiv 2020:0-33 http://dx.doi.org/10.31234/osf.io/3m7up.


82. de Hollander G, Keuken MC, van der Zwaag W, Forstmann BU, Trampl R: Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7 T and 3 T. Hum Brain Mapp 2017, 38:3226-3248 http://dx.doi.org/10.1002/hbm.23586.


Subprocesses of working memory Trutti et al. 65


Reviews neuroimaging research on the links between brain and behaviour in episodic memory and value-based decision making, with a focus on interactions between hippocampus and ventromedial prefrontal cortex. Provides several interesting suggestions for obtaining a more detailed understanding of brain-behaviour relationships in memory-based decision making.


