Mapping visual working memory models to a theoretical framework

William Xiang Quan Ngiam

Department of Psychology, University of Chicago

Institute of Mind and Biology, University of Chicago

Correspondence: wngiam@uchicago.edu

Note: This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.3758/s13423-023-02356-5. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

Abstract (250 words)

The body of research on visual working memory (VWM) – the system often described as a limited memory store of visual information in service of ongoing tasks – is growing rapidly. The discovery of numerous related phenomena, and the many subtly different definitions of working memory, signify a challenge to maintain a coherent theoretical framework to discuss concepts, compare models and design studies. A lack of robust theory development has been a noteworthy concern in the psychological sciences, thought to be a precursor to the reproducibility crisis (Oberauer & Lewandowsky, 2019). I review the theoretical landscape of the VWM field by examining two prominent debates – whether VWM is *object-based* or *feature-based*, and whether *discrete-slots* or *variable-precision* best describe VWM limits. I share my concerns about the dualistic nature of these debates and the lack of clear model specification that prevents fully determined empirical tests. In hopes of promoting theory development, I provide a working *theory map* by using the broadly encompassing Memory for Latent Representations model (Hedayati et al., 2022) as a scaffold for relevant phenomena and current theories. I illustrate how opposing viewpoints can be brought into accordance, situating leading models of VWM to better identify their differences and improve their comparison. The hope is that the theory map will help VWM researchers get on the same page – clarifying hidden intuitions and aligning varying definitions – and become a useful device for meaningful discussions, development of models, and definitive empirical tests of theories.

Keywords

visual working memory; theory development; theory map; model comparison; definitions

Introduction

What *is* visual working memory (VWM)? A generic introduction favored by researchers is that VWM is the system responsible for maintaining visual information in a state of heightened accessibility for ongoing perception and cognition. However, the temporary storage of visual information is far more complex than is encapsulated by this core definition. The flexible and multifaceted nature of the VWM system is evident in the wealth of diverse but related empirical phenomena – attractive and repulsive biases in recall, both feature-based and object-based encoding effects, sustained and activity-silent neural representations, and more. Further, the VWM system is interconnected with perceptual and long-term memory systems (Atkinson & Shiffrin, 1968; Cowan, 1999; Schneider & Shiffrin, 1977; Teng & Kravitz, 2019). It is also not straightforward to disentangle the aforementioned generic definition from a broad definition of visual attention, which most would also define as heightening accessibility to visual information and is thought to be limited. This overall ambiguity may be one cause for the proliferation of subtly different definitions of working memory (Cowan, 2017), and various verbal theories and computational models of working memory (Logie et al., 2021; Oberauer et al., 2018).

Table 1

Chapter and Authors	Definition
A Multicomponent Model of Working Memory – *Alan Baddeley, Graham Hitch, and Richard Allen*	A limited capacity system for the temporary maintenance and processing of information in the support of cognition and action.
An Embedded-Processes	The ensemble of components of the mind that hold a

Approach to Working Memory - *Nelson Cowan, Candice C. Morey, and Moshe Naveh-Benjamin*	limited amount of information temporarily in a heightened state of availability for use in ongoing information processing.
The Time-Based Resource-Sharing Model of Working Memory – *Pierre Barrouillet and Valérie Camos*	WM is the structure where mental representations are built, maintained, and modified according to our goals.
Towards a Theory of Working Memory From Metaphors to Mechanisms – *Klaus Oberauer*	WM is a medium for building, holding, and manipulating temporary representations that control our current thoughts and actions.
Multicomponent Working Memory System with Distributed Executive Control – *André Vandierendonck*	WM is the part of the memory system used to support goal-directed activities. This support includes maintaining the task goal, the selected way to achieve this goal, and the constraints or limitations of this achievement. The WM system also maintains all interim results so as to enable continuation after task interruption.
Individual Differences in Attention Control Implications for the Relationship Between Working Memory Capacity and Fluid Intelligence – *Cody A. Mashburn, Jason S. Tsukahara, and Randall W. Engle*	We define working memory as the cognitive system that permits the maintenance of goal-relevant information. More structurally, working memory comprises domain-general executive attention coupled with domain-specific short-term memories. We regard short-term memory as those aspects of long-term memory residing above some activation threshold, making them available or potentially available to influence ongoing cognition, as well as those processes necessary to keep this activation above threshold (e.g. subvocal rehearsal).
Working Memory and Expertise An Ecological Perspective – *David Z. Hambrick, Alexander P. Burgoyne, and Duarte Araujo*	In the spirit of Boring (1923), we define working memory capacity (WMC) as whatever is measured by the psychological instruments that the field can agree to call working memory tasks. We are agnostic about which theory and definition of working memory is the 'right' one. Taking an ecological perspective, we view working memory performance in terms of the relationship between the person (including knowledge, skills, and abilities) and The environment (including objects and other affordances).

Domain-Specific Working Memory Perspectives from Cognitive Neuropsychology – *Randi C. Martin, Brenda Rapp, and Jeremy Purcell*	Storage systems dedicated to maintenance of specific types of information that are crucial for operation of the system.
Remembering Over the Short and Long Term Empirical Continuities and Theoretical Implications – *Patricia A. Reuter-Lorenz and Alexandru D. Iordan*	WM is a capacity-limited system for the short-term maintenance and manipulation of (domain-specific) information held actively in mind, and commensurate with the notion of the 'activated portion of long term memory (LTM)'. We advocate for better integration of psychologically and neurally informed construct development.
Manifold Visual Working Memory – *Nicole Hakim, Edward Awh, and Edward K. Vogel*	We endorse the embedded-processes model, which puts working memory (WM) in the context of other types of memory. However, we define WM as the processes that maintain a limited amount of information via active neural firing. Therefore, our view of WM closely aligns with the embedded-processes model's definition of the focus of attention.
Cognitive Neuroscience of Visual Working Memory – *Bradley R. Postle*	The ability to hold information in an accessible state—in the absence of relevant sensory input—to transform it when necessary, and to use it to guide behaviour in a flexible, context-dependent manner
A Dynamic Field Theory of Visual Working Memory – *Sobanawartiny Wijeakumar and John Spencer*	In dynamic field theory (DFT), WM is an attractor state where representations are self-sustained through strong recurrent interactions between excitation and inhibition.
Integrating Theories of Working Memory – *Robert H. Logie, Clément Belletier, and Jason M. Doherty*	Our hypothesis is that WM is a collection of domain-specific temporary memory stores and cognitive functions that work in concert to support task performance. Detailed definitions vary according to the research questions and the level of explanation being addressed rather than because of fundamental theoretical differences.

Table 1: Definitions of working memory provided by authors (in italics) for each chapter in the recently published Working Memory: State of the Science (Logie et al., 2020).

The expansive growth of VWM research poses not only the challenge of building an integrative theory that encapsulates all related phenomena, but to also maintain a coherent framework to discuss concepts and theories within. Subtle differences in definition and models are likely to have resulted from disparate research questions, using varying measures and levels of analysis, but may not actually reflect theoretical adversity (Logie et al., 2021). As such, the field risks its common theoretical core unraveling. An underscored precursor to the reproducibility crisis in the psychological sciences has been the lack of direct connection from experiments and tested hypotheses to the underlying theory (Guest & Martin, 2021; Klein, 2014; Meehl, 1978; Oberauer & Lewandowsky, 2019; Scheel et al., 2021). That is, an over-reliance on the hypothetico-deductive method and null-hypothesis significance testing without substantial groundwork to construct collective theories has hindered progress in the psychological sciences (Borsboom et al., 2021; Devezer & Buzbas, 2023; Meehl, 1978).

Just as the lack of rigorous theory development is being scrutinized across psychology, our field is beginning its own introspection. To curb the idiosyncratic nature of theory development, eminent researchers collectively set initial benchmarks for VWM models (Oberauer et al., 2018), and have begun to scrutinize the auxiliary assumptions of such models (Robinson et al., 2022; Williams et al., 2022). A recent thoughtful introspection of the VWM field by Vencislav Popov reveals concerns that the development and evaluation of formal models, while needed, will not be enough to produce a convincing theory of memory (Popov, 2023). It is clear the field could benefit from a concerted constructive effort in establishing a coordinated framework from which to structure well-determined specifications from theory and/or model to hypotheses about observed empirical phenomena (Borsboom et al., 2021; Maatman, 2021; Scheel et al., 2021).

As a step towards a refined theoretical framework, especially one that is productive for discussing and conducting research, I review current theoretical debates in the VWM field (*object-based versus feature-based* and *discrete-slots versus variable precision*) and revamp them in the form of a *theory map*. There are an immense number of related phenomena and models that the field needs to keep track of, and so the idea is to create something that helps us make sense of the theoretical landscape – a *map*. The *theory map* is an illustrative representation to help navigate thinking about how a set of VWM phenomena are linked to a set of functional mechanisms. I adopted the Memory for Latent Representation (MLR) model (Hedayati et al., 2022) as the scaffold, given that it includes a diverse set of ways that information can be stored and has multiple different capacities that can accommodate existing theoretical positions (see *Figure 1*). The goal of this review *is not* to be prescriptive about what is considered VWM and what is not, nor to identify which of our current models is most accurate, nor to advance MLR as the correct model. The goal *is* to provide a rich synthesis of the extant VWM field and push for an integrative, coherent foundation (see Nobre (2022) for a call to updating the standard paradigm and integrating working memory research from various domains). Like old world maps that were not perfectly accurate but still critical for navigation and exploration, the *theory map* is not to be taken as an exact attempt of a unifying theory but to establish a space within which the wider field can start to examine where various models may be reconcilable or incompatible. The hope is that it unifies the language and understanding of our field in a way that promotes clearer situation and specification of VWM models and phenomena – reducing disagreements that have resulted from subtle differences of definition (Cowan, 2017). Ultimately, I hope to help researchers better define their hypotheses and tests to enable study designs that achieve incisive inferences.

A brief summary of recent visual working memory research

The hallmark of visual working memory is its sharply limited capacity, in contrast to the long-term memory (LTM) system which is thought to be immeasurably vast in its capacity. The intense focus of much research on the limited capacity of VWM (Adam et al., 2017; Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Fukuda et al., 2010; Luck & Vogel, 1997; Ngiam et al., 2022; Olson & Jiang, 2002; Vogel et al., 2001, 2006; W. Zhang & Luck, 2008) has brought about the discovery of numerous phenomena and methods of measurement, as well as a plethora of theories and models to describe and explain them (Bays & Husain, 2008; Brady et al., 2011; Cowan, 1999; Hedayati et al., 2022; Oberauer & Lin, 2017; Rouder et al., 2008; Schurgin et al., 2020; W. Zhang & Luck, 2008). I will provide what can only be a narrow summary of this research to illustrate the theoretical landscape of the VWM field – note that substantial empirical research has occurred with aims beyond characterizing the capacity limit.

In the seminal study of Luck and Vogel (1997), this capacity limit was estimated to be approximately four items' worth, after they observed that accurate change-detection performance dropped only when the memory arrays exceeded four items. This pattern appeared to be unchanged by the addition of simple features to these items, such that the capacity limit seemed to be defined by feature-integrated objects – what is now well-known as the *slots* model. In opposition to this, others have since found change-detection performance varied with the complexity of the to-be-remembered stimuli (Alvarez & Cavanagh, 2004) – an indication that it was the number of features, and not the number of objects that determined the capacity limit – what is now well-known as the *resources* model. This gave rise to an initial framework motivating research of the VWM system – *slots versus resources*.

This initial framework still influences current research, which pit various forms of *slot* models and *resource* models against each other. This can be seen in two prominent debates in the field; one being comparisons of *object-based* versus *feature-based* models when examining memory for multi-featured stimuli (Fougnie et al., 2012; Fougnie & Alvarez, 2011; Hardman & Cowan, 2015; Markov et al., 2019; Sone et al., 2021), and the other being *discrete-slots* versus *variable-precision* models when examining change-detection or delayed recall (Ma et al., 2014; Nosofsky & Donkin, 2016; Nosofsky & Gold, 2018; Rouder et al., 2008; van den Berg et al., 2014; W. Zhang & Luck, 2008). There are various signs that revising the theoretical framework may be needed in both these debates.

Comparing object-based and feature-based models of VWM

The initial *slots versus resources* debate on how to characterize capacity limits was related to questions on how content was represented in VWM. A natural expectation is that to first understand capacity limits, we must first understand the unit of representation to measure VWM by. The initial *slots* model argued that the ceiling was determined by a fixed number of *slots* devoted to feature-integrated objects (Luck & Vogel, 1997). Many took this characterization of the capacity limit to suggest the rigid representation of information as *objects* (the *strong object* model). Proponents of *resource* models have argued that capacity limits are determined by allocation of resources to features, citing the diminishing capacity with increasingly complex items (Alvarez and Cavanagh, 2004). Most researchers hoping to uncover the unit of representation in VWM have since asked whether it is object-based or feature-based in a binary manner (Fougnie et al., 2010; Shin & Ma, 2017), despite these seminal papers (Alvarez & Cavanagh, 2004; Luck & Vogel, 1997) not exactly appealing for a purely *object-based* or purely *feature-based* account respectively.

It is likely that the answer is not one or the other, but both in some manner (Fougnie et al., 2010; Markov et al., 2019; Shin & Ma, 2017). Consider the phenomena that have been documented in the literature – recall is improved when information is stored as objects, known as the *object-based benefit* (Fougnie et al., 2012), but also features have been found to be forgotten independently (Fougnie & Alvarez, 2011; Hardman & Cowan, 2015; Markov et al., 2019). The recall for features of an object (e.g. its color and its orientation) has been observed to be independent (Bays et al., 2011; Shin & Ma, 2017), yet shown to also be strongly associated and integrated (Li et al., 2022; Sone et al., 2021). Indeed, these object-based and feature-based phenomena have been observed occurring in concert in a recent study using a novel experimental paradigm – whole-report with conjunction stimuli (Ngiam et al., 2023). If a researcher accepts that both object-based and feature-based phenomena co-occur in VWM, then the dualistic debate with object-based **versus** feature-based model comparisons is unhelpful. Instead, researchers will need to better motivate their studies, by specifying whether their model or theory aims to provide a mechanism which explains a given set of effects, or does aim to explain both sets of effects, and conduct empirical studies accordingly. Indeed, the more precise measurement and careful characterization of these phenomena will aid model development.

Comparing discrete-slots and variable-precision models of VWM

The main arena for *discrete-slots* versus *variable-precision* model comparisons has been on continuous report studies, where the precision of observers' working memory recall can be measured with a circular wheel (Bays & Husain, 2008; Wilken & Ma, 2004; W. Zhang & Luck, 2008). One direct mechanistic contrast between these classes of models is their explanation for guessing behaviors. Discrete-slots models suggest a zero-information state (having no representation of the to-be-remembered stimuli in memory) and this results in random responding. On the other hand, variable precision models typically deny a zero-information state, suggesting that all responses can be explained by variation in memory strength due to noise. This has left researchers to attempt to decipher whether responses are a result of uninformed behavior from a lack of representation or very imprecise responding from noisy representations. It has been raised that this distinction is impossible to make with standard continuous report studies because models may mimic each other (Adam et al., 2017). In simulations, Adam et al. (2017) reported that a million noise-free samples would be needed to successfully distinguish responses that did in fact result from a memory representation that was widely imprecise (a von Mises probability distribution with a standard deviation of 193 degrees) but nevertheless did exist, and not from random responding posited due to having no representation. Indeed, a factorial model comparison of *discrete-slots* and *variable-precision* models found it would end in a stalemate with standard continuous report tasks (van den Berg et al., 2014).

A critical argument for an upper bound on the number of discrete representations in VWM is the existence of responses based on zero-information states ('true guesses'). A recent study attempted to break the deadlock between *discrete-slots* and *variable-precision* models by reconfiguring the continuous report task to produce a signature of pure guessing that distinguishes it from imprecise memories (Ngiam et al., 2022). Despite finding evidence for guessing responses that cannot straightforwardly be explained by a noisy memory, the authors noted that the finding does not determine whether VWM resources are discrete or continuous in nature – the observed data could be explained by a version from both discrete and continuous classes of models if they are allowed some modifications. Indeed, this is a usual retort by proponents of pure resource-based or resource-rational models – that modern resource models can account for such zero-precision estimates without an additional mechanism (Bays et al., 2022; Schneegans et al., 2020; van den Berg & Ma, 2018). Then, it is unclear what data and what evidence would be needed to deterministically decide between existing models of VWM.

I believe both classes of models, as all working scientific theories do, have their shortcomings – the *discrete-slots* models do not explicitly provide a clear mechanism for how working memory representations can be variably noisy, and the *variable-precision* models do not provide a strict constraint on how resources are distributed across memoranda and may give rise to item limits or object-based benefits (Oberauer et al., 2016). Although hotly contested, much like the object-based versus feature-based debate discussed above, it is likely that the nature of VWM is both discrete (say object-based representations) and continuous (say noisy representations) in some way – perhaps by possessing mechanisms at both levels of representation and/or due to flexibility in the system. Improving instruments to better measure visual working memory phenomena will likely lead to better evaluation of models, but what is needed is also the better *specification* of these models in relation to observed phenomena to allow for more severe tests and then stronger inferences.

Issues conducting research without a well-specified theoretical framework

As an illustrative example of the difficulty in resolving scientific debates, take the challenge of researching the nature of VWM capacity limits by comparing *discrete* versus *continuous* theories. As defined above, *discrete* models propose a maximum number of items being represented in VWM. They do not typically specify a mechanism for the representation of features (say within a slot) or by which noise or uncertainty is introduced into the representation (but see the *slots-plus-averaging* model (W. Zhang & Luck, 2008)). As such researchers may assume the strict version of the *slot* model where memory for features exist only within integrated objects in WM, are noiseless (or are of fixed noise), and lost in an **all-or-none** fashion ("the *strong object* model"). One might conduct a formal model comparison between this specific strong object model against a *continuous resources* model which provides flexibility in its explanation and prediction in any of those regards. When the *strong object* model ultimately fits the observed data poorly, researchers may erroneously bundle assumptions and infer that no *discrete slot* explanation of the phenomena can be accurate, presenting their findings as evidence of "confirming" their alternative *continuous resource* model. Rather the most accurate inference is that the findings have falsified the specific strong object model but not *all* versions of the discrete-slot model (say for example, *weak object models* that are **some-or-none**). But note that the criticism that *discrete* models do not always provide an explicit explanation for noisy representations is a valid one. Unfortunately, the severity of the inference that can be made is limited by poor model specification – the model comparison does not specify mechanisms at a level of detail sufficient to choose a definitive victor.

Thus, a prevailing factor that impedes progress is that a lack of specification in VWM models has meant research is unable to produce evidence that strongly determines one theory over another (Maatman, 2021; Meehl, 1990), though researchers often make the declaration they have found evidence that favors one class of models and dismiss the other. A lack of specification can enable flexibility in a model such that *ad hoc* changes allow the model to account for any and all empirical results, skirting any strong test (Navon, 1984). One further danger of underdetermination from unspecific theories and obtuse studies is that it allows researchers to bundle assumptions of an opposing model as to create a straw-man for it to be outcompeted (as the above example hoped to illustrate).

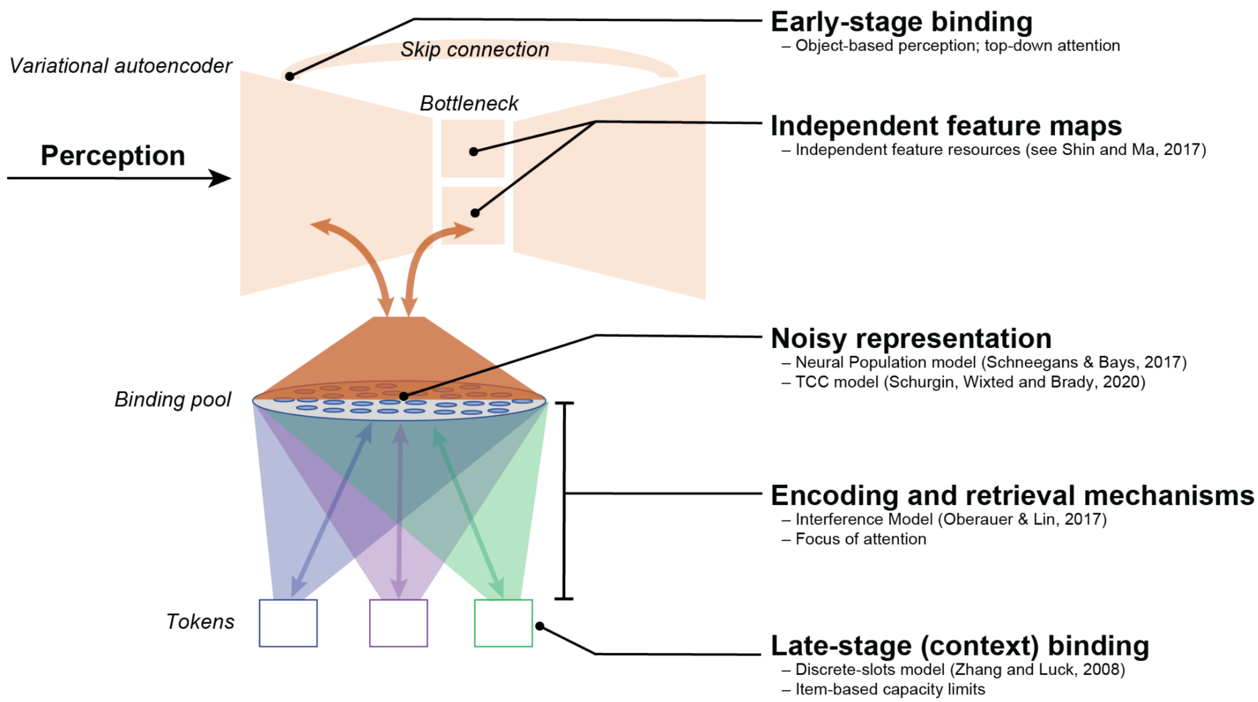

Figure 1. A simplified schematic of the Memory for Latent Representations (MLR) model architecture (Hedayati et al., 2022) with visual working memory phenomena and current models mapped on to its components: the variational autoencoder (VAE), the binding pool, and the tokens. This theory map aims to provide a coherent framework within which to organize visual working memory phenomena and discuss the relevant explanatory models. As such, the compatibility or inconsistencies between models can be better identified, and subsequently tested. For example, one could use a working definition for the noisy representation in VWM as the noise held in the pattern of neuron activity in the binding pool that follows a summation of information from various perceptual sources.

Initial steps to a working theory map of visual working memory

In my view, by and large, more progress could be achieved by rethinking our theoretical framework and broadly adopting a model-based approach (Devezer & Buzbas, 2023) so as to specify and examine how VWM actually *operates* and gives rise to its sharp capacity limit and other extant phenomena. A fixation on understanding the exact representative unit of VWM to characterize its capacity limit, a remnant of the *slots versus resources* comparisons, remains pervasive in shaping current research approaches. One might reasonably expect the VWM system to operate in a flexible manner to accommodate different demands (Boettcher et al., 2021; Nasrawi & van Ede, 2022; van Ede, 2020; van Ede & Nobre, 2023). If one accepts that to be the case, finding a highly specific format for working memory that explains all observed phenomena might be a hapless pursuit.

In an attempt to address the aforementioned challenges of promoting construction of a broader theoretical framework, I created a *theory map* by situating modern VWM models onto the Memory for Latent Representations (MLR) model (Hedayati et al., 2022). Using an existing model to scaffold the theory map may be puzzling to some. This choice was made in-part to avoid the theory map being interpreted as introducing an entirely new model altogether – something I believe would be largely redundant and not so helpful for the field. To reiterate, the goal is to compare and contrast the existing models in the literature, and show that those models have mechanisms that can somewhat overlap and integrate. And by demonstrating that is the case, it will emphasize that clear specification and thoughtful experimental design is needed to identify and subsequently test where existing models are in disagreement.

The MLR model provides a suitable basis for the theory map for many reasons. To briefly introduce the MLR model, it consists of multiple subsystems that encapsulates what many may broadly consider to be (or interact with) VWM. It is built with functional neural mechanisms that are computationally implemented, providing various specific pathways for how visual information may be represented and what limitations might exist. The specific implementation of various memory mechanisms makes it ideal for pinning down the theoretical landscape of VWM that may presently be too nebulous to effectively be mapped with language, boxes and arrows. Further, as the MLR contains multiple subsystems, one can visualize various notable theoretical proposals – for example, structured hierarchical representations (Brady et al., 2011; Cowan, 1999) and partial packaging of features into objects (Shin & Ma, 2017). I want to stress that the MLR model is used simply as a starting point for a map – and it does not cover the entire possible theoretical space, nor is it claimed that the MLR model is more accurate in comparison to other models. But here, by mapping existing VWM models and concepts to this scaffold, I hope the field better defines the VWM phenomena that are observed and better specifies what each existing VWM model attempts to and does successfully explain. In completing this exercise, I also hope to provide a taxonomy of existing WM models with an accompanying summary. The result is an initial schema of VWM models – a broad theoretical framework, detailing various phenomena and how they are explained by various theories.

The goal is that the *theory map* will help researchers better scrutinize where different models may be harmonious or at odds in their explanations of various phenomena. It may also provide a common language or platform for researchers to discuss their varying perspectives on how the VWM system operates. VWM models have vastly and rapidly evolved since their initial slot or resource conceptions, and the wider field may not have kept track of the key data and subsequent critical changes to the models (Bays et al., 2022). For example, despite the *strong object* model generally being disproven (Hardman & Cowan, 2015; Olson & Jiang, 2002) and not currently widely believed, its specification and ideas continue to feature in research in various ways (Robinson et al., 2022; Williams et al., 2022). The hope is to inspire researchers to consider what may require substantial theoretical construction and specification, and where research may perhaps be best targeted to advance our understanding of VWM.

A primer of the Memory for Latent Representations (MLR) model

The MLR model architecture consists of two main components: a variational autoencoder (VAE) and a token-based binding pool (BP) (*Figure 1*). The VAE contains layers of neurons arranged in a bowtie shape – the input layer contains many neurons, contracting to a very small number of neurons in the middle, before expanding back out to an output layer with many neurons. The first half of the bowtie, termed the ***encoder***, converts a visual stimulus into a highly compressed representation that is separated into distinct independent feature maps. The second half of the bowtie, termed the ***decoder***, reverses the process, producing a visualization of the compressed representation. The VAE resembles the ventral visual system hierarchy – the encoder corresponds to the feedforward connections from primary visual cortex (V1) to the inferior temporal cortex (IT), while the decoder corresponds to feedback projections in the opposite direction.

Any stimulus presented to the model will evoke a series of neural firing patterns in the encoder, where each layer is referred to as a ***latent*** representation. Note the skip connection, an additional component of the VAE, that links the first layer of the encoder to the last layer of the decoder, bypassing the feature maps. The skip connection allows the MLR to model differential effects of novel and familiar stimuli (Asp et al., 2021; Chung et al., 2023; Hedayati et al., 2022; Ngiam et al., 2018; Xie & Zhang, 2017).

The Binding Pool (BP) (Swan & Wyble, 2014) is a separate neuron layer for the storage of memories. The BP can *encode* any combination of the latent spaces in the encoder of the VAE (the first half of the bowtie). Importantly, the connections between the encoder and the BP are bidirectional, allowing the encoded memories to be *retrieved* – the BP reconstructs (noisily) what it had encoded back into the VAE. The BP itself is connected bidirectionally to a set of tokens – each token is connected to a subset of BP neurons, with overlap possible in which neurons connect to each token. The token does not itself represent any featural information, but rather can be activated to reproduce the stored activity in the subset of BP that it is connected to.

Notionally, tokenization is the indexing of BP activity that pertains to an *item* following the feature-binding process. Multiple tokens allow for multiple items to be individuated, stored and retrieved, even if the items share spatial locations or feature values. These tokens may support higher-level cognitive operations – say, the chunking of items through associative learning or mental rotation of an object.

It may be helpful to draw comparisons between the MLR model and the *embedded process* model of working memory (Cowan, 1999). The embedded process model views VWM as a hierarchy that comprises a long-term memory store, an activated subset of long-term memory (a short-term store), and a subset of the activated memory to be in conscious awareness (the focus of attention). The BP can be thought of like the *focus of attention,* in the sense that it holds the subset of VWM information that is actively selected (with that information becoming "in mind" once projected into the VAE). Then, the tokens are somewhat comparable to *activated long-term memory* – information held in highly accessible but latent states that can be quickly brought into the focus of attention – much like how the tokens can be reactivated in the BP (but remember that the tokens are not active representations of the information themselves). It could be theorized that these tokens link to content held in long-term memory (e.g. learned structural and semantic knowledge) that enable VWM resources to be freed up, and provide an interface by which information is eventually encoded into long-term memory (O'Reilly et al., 2022). This connection to long-term memory is not yet computationally implemented in the current version of the MLR model (but note that the MLR model can account for differential effects of novel and familiar stimuli as mentioned above).

In sum, visual information is represented by the 'perceptual brain' (the VAE) where each layer of the encoder projects into the BP. The resulting pattern of activity gets indexed as tokens to allow storage of multiple items. Any given token can be reactivated back into the BP, which projects back into the VAE and used to generate responses or translated through the decoder to generate a reconstruction of the original stimulus (akin to mental imagery). See Hedayati et al. (2022) for more information on the architecture of the MLR model, and a detailed implementation written in Python 3.7 using *pytorch* is hosted at https://osf.io/tpzqk/.

Bringing *slots* and *resources* into accordance

Earlier in this review, I provided a limited summary of *slots* and *resources* ideas in two subdomains (object-based versus feature-based models, and discrete-slots versus variable-precision models) and suggested that the likely answer is that VWM is unlikely to be specifically one or the other in each of those subdomains. This VWM theory map (*Figure 1*) allows us to visualize how these ideas may interact in accordance, rather than place these classes of models in direct opposition. Further, it brings a new perspective to capacity limits in the VWM system – that bottlenecks of visual information can occur at various levels rather than being treated simply as a singular limit that can only exist on features or on objects.

The core idea of *object-based* and *discrete-slots* models is that there exists a representation in working memory where the features of an object are bound and held in mind – this is akin to the **tokens** in the theory map. The tokens are grounded on the concepts of *object files* (Kahneman et al., 1992)*, fingers of instantiations* (Pylyshyn, 1989) and most recently in the VWM subdomain as *content-independent pointers* (Balaban et al., 2019; Thyer et al., 2022) – specific feature values (e.g. color, orientation, location) produce activity in the binding pool that is assigned to a token. One important distinction that might need clear definition is *identity content* (feature values) from the pointers themselves. A current idea of interest is whether VWM pointers are spatiotemporal in nature (Thyer et al., 2022), whereby time and location are critical and necessary components for the binding of features (Heuer & Rolfs, 2021; Schneegans et al., 2023; Schneegans & Bays, 2017). However, these pointers may be defined in an object-based manner from Gestalt processes, and not only in terms of its context (binding to time and/or location) (Balaban et al., 2019; Balaban & Luria, 2016; T. Gao et al., 2011; Z. Gao et al., 2022). This *token* mechanism may correspond to observed neural correlates for object-based pointers (Thyer et al., 2022), and to the explicit conjunctive coding of object features that has been observed in the perirhinal cortex (Erez et al., 2016; Liang et al., 2020). These *context-bound* tokens are supposed to be critical for sustaining and updating an object-based representation, and a recent review suggests pointers as a plausible attention-based neural mechanism connecting representations of content to representations of structure in the human brain (O'Reilly et al., 2022). Note that the MLR model does not have a set limit on the number of tokens, but object-based models typically assert that there is (or perhaps *can be*) an item-based capacity limit, typically referred to as *K* (Cowan et al., 2005) – a maximum number of tokens that one can actively maintain in working memory (Adam et al., 2017; Ngiam et al., 2022). To be precise, VWM capacity may not be limited by the number of objects exactly, but by the number of objects that can be bound to its spatiotemporal context and actively maintained (see also Huang, 2020 for a Boolean map account of VWM).

A core tenet of *resource* models is that working memory representations vary in strength in a continuous manner. I have straightforwardly mapped this on to the **binding pool nodes** to encapsulate that facets of VWM like variable precision has been predominantly modeled as signal and noise in neural populations (Bays, 2014; Bays et al., 2022; Schneegans et al., 2020; Schneegans & Bays, 2017). Population coding accounts model the representation of VWM as an encoding-decoding process – a variable number of samples generated from a neural population tuned for feature values is read-out to produce a signal that informs response behavior (Bays, 2014; Schneegans et al., 2020). Note that there exists multiple 'neuron' layers in the MLR model that will contain various degrees of noise from which a signal may be drawn (see Hedayati et al. (2022) for how the different layers are differentially noisy for novel and familiar to-be-remembered items). Another relevant facet of the theory map are the independent feature layers that project into the binding pool, in a correspondence to models proposing independent resources for separate features (Fougnie et al., 2010; Fougnie & Alvarez, 2011; Markov et al., 2019; Shin & Ma, 2017). Feature-based phenomena, such as effects of stimulus complexity (Alvarez & Cavanagh, 2004; Hardman & Cowan, 2015; Olson & Jiang, 2002) or the independent loss of features (Fougnie & Alvarez, 2011), can be related to mechanisms involving these independent feature layers.

The critical point then, as demonstrated by the VWM theory map, is that these two prominent classes of models and the ideas they represent are not necessarily mutually exclusive – object-based representations (in the form of *tokens*) can co-exist with noisy representations (in the form of neural populations). If this theory map is taken to be plausible, it is then a substantial challenge for VWM researchers to define an experimental design that would fully determine currently observed working memory phenomena or claim that representations take one form or the other. That is, a researcher is warned against basing their inferences on a dualistic framework (or at least claiming their results reflect a true nature of VWM broadly) without determining that it is indeed truly dualistic – it is not slots *or* resources, object-based *or* feature-based, discrete *or* variable-precision.

Then why do we see papers with such "clear-cut evidence" in favor of one idea or the other? The theory map described above illustrates that the complex VWM system can embody both ideas, which means that an experiment can be tailored to reveal the constraints imposed by one functional theory (e.g. a *slots* account or *resources* account). A set of empirical results may show capacity limits indeed being constrained by the number of objects as per a *slots* account in one paradigm but this does not deny that capacity limits can *also* be constrained at the feature-level in other experimental conditions (and the vice versa). If researchers do not deny that the VWM system may flexibly adjust in various experimental conditions, each with potentially varying factors setting capacity limits, then researchers should not readily make the grand claim that their select theory truly characterizes VWM. Researchers should instead carefully identify and report the boundaries where their theory or model applies (Donkin et al., 2016).

Further, key disagreements between current VWM researchers may be more specifically defined by using the above theory map, and then perhaps more meaningfully discussed and contested. Consider the current debate between proponents for an object-based item limit and resource-based accounts of VWM capacity on the existence of 'true guesses' (Adam et al., 2017; Bays et al., 2022; Ngiam et al., 2022; Schurgin et al., 2020) that was reviewed earlier. Here, 'true guesses' as defined by item-limit theorists may be best characterized as responses with no available token to inform the response, rather than defined as having 'no working memory representation' with which to inform behavior. This opens the possibility that activity may still be sustained in other layers of the VWM hierarchy from which a response may be produced – perhaps weights in the architecture that define an existing prior, or residual activity from previously encoded and represented information. As such, the burden of proof for evidencing an object-based limit on the VWM system should not rest on showing the existence of purely zero-information states, nor does providing a continuous resource account that explains the gamut of low-precision or 'guess responses' refute the possibility of a discrete token-based representation

(Ngiam et al., 2022). Therein lies a substantial challenge to design reliable measures or experiments that undoubtedly capture how information is represented and processed in working memory.

Comparing and contrasting leading models of VWM capacity limits

To supplement the above overview of *discrete-slots* and *variable-precision* models, I will provide a brief summary of three other leading accounts of capacity limits in VWM – the *neural population* model (Bays, 2014; Schneegans & Bays, 2017), the *target confusability competition* (TCC) model (Schurgin et al., 2020) and the *interference model* (Oberauer & Lin, 2017). These formal models have a large degree of correspondence in the phenomena they try to capture, and as such I try to situate these models within the theory map in an attempt to clearly specify which phenomena they may or may not connect with. This may provide a means to better compare, contrast and benchmark these leading models – to clearly delineate where they compete, or where they may provide unifying accounts for VWM phenomena (Oberauer et al., 2018). Note that all these models appear to have varying architectures, but they can be placed and compared within the same space for formal comparisons (Oberauer, 2023). The present goal is not to directly comment on the model architectures as to which is more accurate but to *situate* these models in a way that better understand their relation so that we can better produce tests for them (Popov, 2023).

The Neural Resource model

The Neural Resource model (sometimes referred to as the Neural Population model), first published by Bays (2014) and extended by Schneegans and Bays (2017), is a population coding account, defining working memory in terms of the spiking activity in a population of neurons tuned to encode stimulus features. The model applies an encoding-decoding process – during encoding, each neuron stochastically spikes based on its tuning (preference for the presented stimulus feature value) and the spiking activity from all neurons is then decoded to estimate the most likely stimulus feature. This process effectively captures the error distributions of single- and whole-report continuous VWM recall tasks (Bays, 2014; Schneegans et al., 2020; van den Berg et al., 2012), and is mathematically equivalent to the *variable-precision* model when a very large population of neurons is assumed (Schneegans et al., 2020).

The Neural Resource model was recently updated to incorporate a temporal dimension, accounting for dynamics in the neural activity of sensory areas (iconic memory) that project into WM (the Dynamic Neural Resource model) (Tomić & Bays, 2023). The VWM neuron population accumulates activity from a sensory signal that rapidly decays following stimulus offset. With the decay of the sensory trace, the VWM population accumulates noisier signals, leading to diffusion of the represented value and the eventual output becoming noisy. With multiple to-be-remembered items, the Dynamic Neural Resource model assumes that when an item is cued for recall, any signal for non-target items is dropped, releasing resources for the signal of the target item to be scaled up. It is not yet agreed upon how (or even whether) reallocation of mnemonic resources occurs with retro-cueing or orienting of internal attention (Gunseli et al., 2015; Y. Lin & Fougnie, 2022; Myers et al., 2017; Souza & Oberauer, 2016). Tomić and Bays (2023) provide empirical validation that the Dynamic Neural Resource model can accurately model aggregate error distributions across memory arrays with various set sizes and stimulus durations.

An important point is that the Neural Resource model, *variable-precision* model and *slots+averaging* model can be expressed within the unifying framework of *stochastic sampling*

(Schneegans et al., 2020). In this framework, VWM performance – the quality of the memory representations and its capacity limit – is determined by the total number of samples of the neuron populations, and their distribution among the to-be-remembered items. For example, consider a typical whole-report continuous recall task. In that task, an item might be represented more precisely in a given trial because there was a higher number of samples overall on that trial and/or because it captured more samples at the cost of other memory items. Thus, despite the underlying resource of samples being continuous in nature, the capacity limit can appear to be discrete because a subset of to-be-remembered items may typically capture a number of samples to reach an effective threshold of report. This appears to echo the stalemate between *variable-precision* and *discrete-slots* models in explaining error distributions, shown in a factorial comparison of these VWM models (Adam et al., 2017; van den Berg et al., 2014). Again, this illustrates the challenge in distinguishing such models with current empirical evidence and methods and encourages the development of precise measures of phenomena.

Here, I have localized the Neural Resource model to **the binding pool**, because like the MLR model, it models a pattern of activity being represented in a population of neurons specific to VWM. The new Dynamic Neural Resource model, by modeling the connection between iconic memory and VWM, can be likened to the connections between the VAE and the BP in the MLR. Of note, the Neural Resource model proposes feature binding occurs via spatial location (Schneegans & Bays, 2017) – given that neuron populations are likewise tuned to locations, the decoding process identifies the most likely target location to read out the associated feature values (e.g. a stimulus' color and orientation) (see also recent work on the role of time in feature binding (Schneegans et al., 2023) and a Boolean map account of VWM for a different perspective on the role of spatial location (Huang, 2020)). As such, the Neural Resource model does not invoke a token-based or item-based mechanism *per se* (as presented in the *theory map*), though it has parallels with the notion of spatiotemporal pointers that I outlined above when describing modern 'slots' accounts. This highlights a potential use case of the theory map – clarifying the proposed mechanisms involving space and time of each of the current VWM models, and specifying how they explain feature-binding or object-based effects (Fougnie et al., 2012) and other extant VWM phenomena.

The Target Confusability Competition model

The recently developed target confusability competition (TCC) model (Schurgin et al., 2020) proposes applying a signal detection framework unilaterally, but within a psychological similarity space rather than the stimulus defined space where VWM is typically modeled. It is well-known that the discriminability of the stimulus set impacts performance on standard VWM tasks. For example, change-detection accuracy is influenced by the similarity between the target and foil – changes are better detected when they are cross-category (e.g. shaded cube to a Chinese character) compared to when they are within-category (e.g. shaded cube to another shaded cube) (Awh et al., 2007). The TCC model formalizes the discriminability of the stimulus set by estimating the psychological similarity space using an empirical psychophysical measure, such as through a quad perceptual matching task (independent to the VWM task at hand). With the underlying discriminability of the stimulus set accounted for, VWM performance can then perhaps be more exactly compared across various stimulus sets and task conditions.

Consider the continuous report task that is often employed to probe VWM (Ma et al., 2014; W. Zhang & Luck, 2008). In this paradigm, subjects are briefly presented with an array of items (e.g., a set of colors) from a continuous stimulus space (e.g., a color wheel). Subjects are cued to recall a single item, and respond by precisely clicking within an annulus (i.e., selecting a color on the color wheel). Responses are typically modeled on the circular space itself – VWM precision is operationalized as the standard deviation in circular degrees from the target value. However, this assumes that the degree of error is uniform between all values in the stimulus set. Now, consider the effect of categories (e.g. color categories like 'red' or 'green') that has been shown to impact performance on a VWM continuous report task (Hardman et al., 2017; Pratte et al., 2017; Ricker et al., 2023; Souza et al., 2021). By definition, categories define stimulus values that are *similar* (within-category) and those that are *dissimilar* (out-of-category). So, when subjects are required to recall a red target, they are more likely to err with a shade of color in the 'red' category because they are similar and confusable, but not likely to err with a shade of color in the 'green' category because it is easily discerned as different. The TCC model explicitly accounts for the non-uniformity of any kinds of effects that influence the *psychological similarity* of the stimulus values, and by doing so, it can then better predict the error distributions on standard continuous report tasks (Schurgin et al., 2020; Williams et al., 2022). A helpful interactive primer for the TCC model can be found at https://bradylab.ucsd.edu/tcc/.

With the TCC model, the authors propose that visual working memories only differ according to their strength (the *d*' parameter), doing away with separate concepts for *number* and *precision* (Schurgin et al., 2020). In my opinion, it is important to still theorize about fluctuations to the underlying *psychological similarity* space that may occur with shifts in attention or learning and experience (for e.g., changes in perceptual fluency with statistical learning (Perfors & Kidd, 2022)). Further, the TCC model, like the variable-precision and Neural Population models, denies the existence of a guessing state (and assuming memories can be defined in terms of being remembered or not) – all to-be-remembered items are encoded with some variation in memory strength producing a familiarity signal. However, the nascent TCC model does not yet formally define *how* the memory strengths may vary across items within an array, though it is capable of representing that variability. In my view, this is where the TCC model and an object-based pointer model can perhaps be compatible –a signal-detection account can accurately account for the precision of recall at the individual item level, but the *distribution* of resources across the items may be best modeled with an object-based capacity limit and variation in achieving that maximal capacity (Hakim et al., 2020; Ngiam et al., 2022, 2023).

Mapping the TCC model to the *theory map* is not straightforward because the origin of the *psychological similarity* function is left undefined, though it is measured by an independent psychophysical task. Of note, the TCC model and the Neural Resource model (Bays, 2014) appear to have largely different accounts for VWM performance, the models share a large correspondence in architecture – both define a distributed pattern of activity produced by a preference to a feature value, that is corrupted by noise and subsequently decoded to output a feature value (Bays, 2019; Tomić & Bays, 2022). Recent work points to this large similarity in model architecture as a possible reason for why the TCC model can produce accurate fits (see *Figure 2*) – Tomić & Bays (2022) failed to find correlations between psychophysical perceptual similarity measures and VWM error distributions in four separate stimulus dimensions, despite it being a core rationale of the TCC model. Here, I have placed the TCC model alongside the Neural Resource model, connecting it to the binding pool of the theory map, to highlight their similarity in model architecture but their very different explanations of phenomena. It is perhaps disagreeable to do so, because that is at odds with a signal detection framework whereby a familiarity signal is computed across a distributed population of neurons (Bays, 2019). Nevertheless, mapping them together within the theory map emphasizes the need for careful theory-driven experimental design to separate and definitively test the TCC model and the Neural Resource models – or perhaps to integrate their ideas (*Figure 2*).

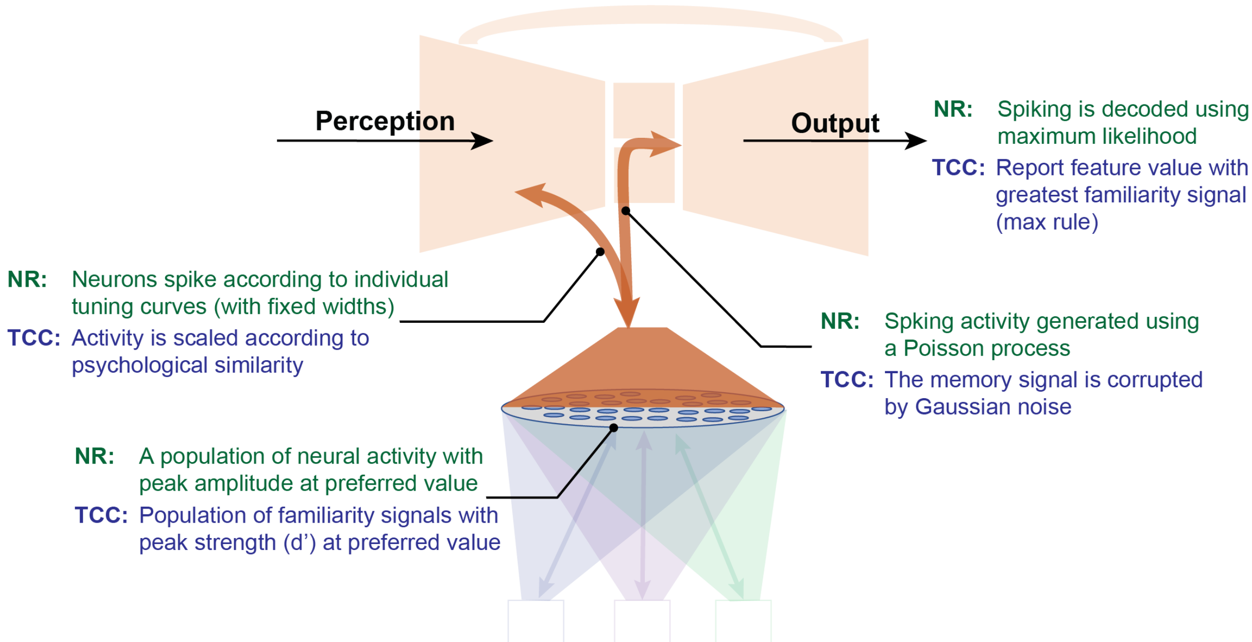

Figure 2. Using the theory map to compare the Neural Resource (NR) model (Bays, 2014) and the Target Confusability Competition (TCC) model (Schurgin et al., 2020). The computational implementation of both models are mapped on to different components of the theory map, showing their large degree of correspondence (see Figure 1 in Tomić & Bays, 2022). This theoretical backdrop may enable an incisive design that contrasts these two models.

The Interference Model

The Interference Model (Oberauer & Lin, 2017) proposes the VWM capacity limit (the decline in precision of recall with increasing memory load) is a result of interference between the representations in working memory. According to this model, encoding into working memory occurs when an item's content (e.g. the feature value of the item) is temporarily bound to its context (e.g. the location of the item within an array). During retrieval, there are three sources of activation – context-based activation (the retrieval of content information is activated according to the amount it is bound to the context cue), context-independent activation (persistent activation from maintaining the items on each trial) and uniform background noise across all response candidates. The probability of retrieval is based on the relative activation of each target item. It is also assumed that the focus of attention holds only one item, and as the context representations are limited in precision, items in nearby contexts (e.g. spatially neighboring items) will be activated and compete with the target item to be held in the focus of attention. With increasing set size of the memory array, more items are likely to be in the nearby context, thereby increasing interference and giving rise to capacity limits (*Figure 3*).

Oberauer and Lin (2017) demonstrated the useful distinction of *context* and *content* in their original paper introducing the Interference Model. Firstly, the Interference Model accurately predicts similarity in *content* (e.g. target and non-target are similar in color) produces a *benefit* to recall performance. This phenomena is commonly attributed to perceptual grouping or ensemble representation (Brady et al., 2011; Brady & Tenenbaum, 2013; see also Chunharas & Brady, 2023). In the Interference Model framework, when a target and non-target share similar feature values, their summed activation produces a peak close to the target feature value. When target and non-target are dissimilar, their summed activation distorts away from the target feature value (i.e. non-target intrusions have some likelihood). Secondly, the Interference Model predicts similarity in the *context* of target and non-targets (e.g. target and non-target share similar spatial locations, when cued by a location probe) produces a *cost* to recall performance. The Interference Model accurately predicts that non-target confusions are more likely to occur with greater similarity in the cue dimension. In brief, the Interference Model captures that *similarity* has differential effects in the *content* and *context* dimensions.

Given its architecture rests somewhat on *context* and *content* dimensions, it felt appropriate to map the Interference Model to the encoding and retrieval operations between the binding pool and tokens. To flesh out the analogy, tokens may reflect the result of binding visual information to a spatial location in time (*context*-binding), and that interference arises from the competition of retrieval from tokens to the limited workspace (here, the binding pool) – but note that the architecture of the Interference Model does not invoke a token mechanism (*Figure* 3). As such, the Interference Model is compatible with an additional discrete capacity limit and the possibility of guessing states (Oberauer & Lin, 2017). The Interference Model was recently applied to capture performance on change-detection tasks, outperforming the variable precision model, slots-plus-averaging model and Neural Resource model in predicting change-detection performance – specifically in estimating the set-size effect and intrusions from non-target item probes (H.-Y. Lin & Oberauer, 2022). The Interference Model also satisfies benchmarks across both visual and verbal working memory domains (Oberauer & Lin, 2023) and may provide a unifying account between the two.

It is interesting to consider how interference (in the broader sense) may vary as a function of psychological similarity (say, subjective distinctiveness) *and/or* perceptual similarity (say, objective distinctiveness). Stimuli that observers have learned to discriminate might be predicted to produce smaller inter-item interference effects in both *content* and *context* dimensions. That is, an observer that has learned to make fine discriminations between shades of colors might be less susceptible to interference both when color is the retrieval cue (the *context*) or when color is the to-be-retrieved feature (the *content*) (but see McMaster et al. (2022) where cue-feature variability accounts for the prevalence of swap errors). Curiously, observers that were trained to identify letters of foreign alphabets as proficiently as fluent readers showed no improvement to memory span for those trained letters (Pelli et al., 2006), but encoding rate and capacity is distinctly increased for letters from familiar alphabets compared to unfamiliar alphabets, even when matched on stimulus complexity (Ngiam et al., 2018). That is to suggest that interference could be a confluence of the physical similarity of the stimulus and/or the individual's familiarity or experience with the stimuli. The theory map serves a helpful reminder that various factors, like perceptual and psychological similarity, may shape encoding of the VWM representation (say, acting on the various layers of the VAE), as well as also variably act on retrieval (say, which token is tapped to be retrieved into the binding pool) before the eventual output of a response on a VWM task.

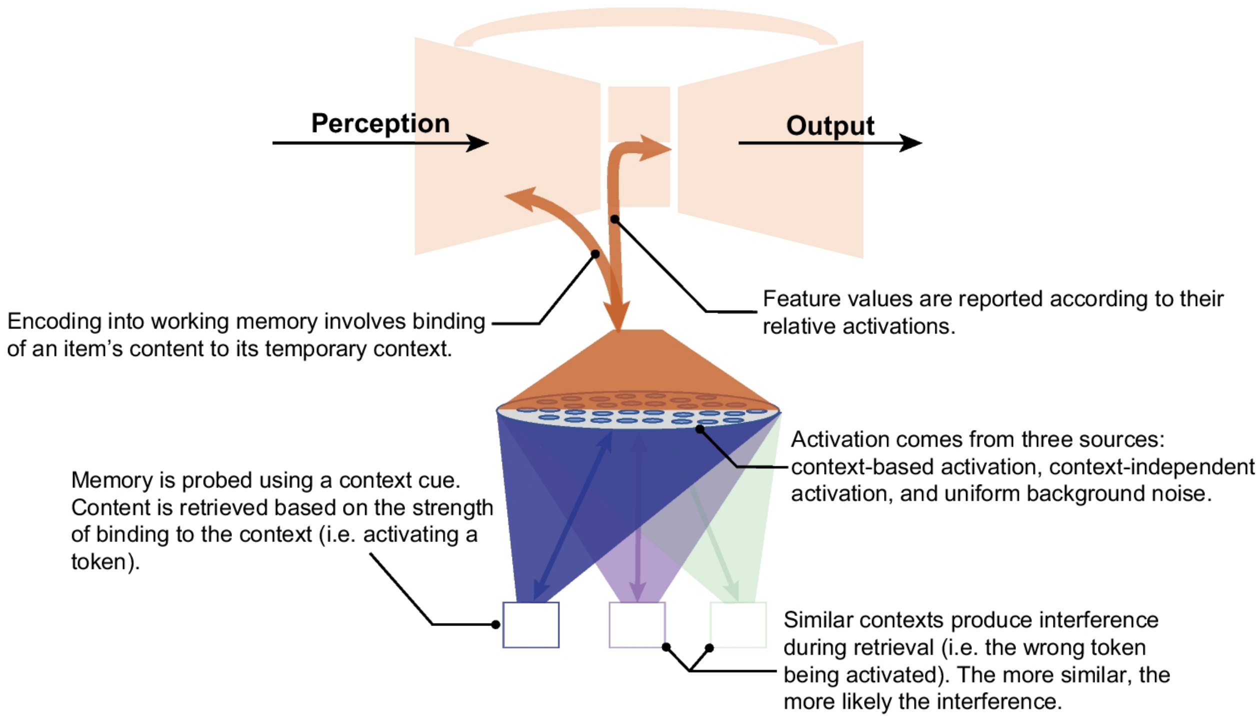

Figure 3. Using the theory map to visualize the Interference Model (Oberauer & Lin, 2017). A probe cues a specific context, prompting activation of a token (the blue token on the left). Tokens similar in context (the purple token in the middle moreso than the green token on the right) produces interference. This shapes the activity in the binding pool and what feature value is ultimately reported. Note that the Interference Model does not appeal to this token-based mechanism in its computation.

Using the theory map to discuss and develop visual working memory theory and phenomena

I hope that in trying to map the various current models above, I have demonstrated there are potential compatibilities between the models in explaining extant VWM phenomena. I hope to have also guided readers to the specific areas where the models may fundamentally disagree – good starting points to design more incisive experimental studies. The *theory map* provides a framework that visual working memory researchers can align discussions about various phenomena or models – through revealing hidden intuitions or clarifying imprecisions in verbal descriptions. Having a common framework may lead to more fruitful discussions about the mechanisms and models for visual working memory phenomena, preventing misunderstandings caused by having differing definitions (Cowan, 2017) or different measures and analysis methods (Logie et al., 2021).

A practical guide to theory development is the *Theory Construction Methodology* (TCM) (Borsboom et al., 2021). It aims to connect theory to phenomena to data in 5 steps – identify empirical phenomena, develop prototheory, formalize theory and phenomena, check explanatory adequacy, and evaluate theory. I believe the theory map will be useful for at least the first two steps – the theory map encourages clearly defining which empirical phenomena one hopes to explain at the outset, *before* generating a new prototheory or applying existing models. Oberauer et al. (2018) has provided a detailed list of relevant empirical phenomena that a theory of working memory should explain and be benchmarked on. The hope is that this theory map will encourage researchers to avoid taking a dualistic approach to defining theories and testing them (theory A versus theory B), and to carefully situate how their empirical studies or data may relate to and shape existing theories. Following the TCM should also promote careful consideration of what empirical data would be definitive evidence supporting or undermining the different existing models.

To demonstrate how the *theory map* can be a useful device in discussion about VWM concepts, let us consider how the *creation* of the VWM representation may be achieved (refer back to Figure 1). From the beginning, initial *encoding* into VWM can be influenced by numerous factors like top-down attentional modulation (Gazzaley & Nobre, 2012; Teng et al., 2022; Teng & Kravitz, 2019) or learned knowledge (Asp et al., 2021; Brady et al., 2016; Hedayati et al., 2022; Ngiam et al., 2018, 2019; Xie & Zhang, 2017) – I point to the *skip connection* as the place where these effects occur, like in the MLR model. However, these factors that influence the initial apprehension of information may not exactly define the *tokenization* or creation of *pointers* – the binding of featural information to the spatiotemporal context – in the same way. Hence, the theory map readily differentiates between these two levels as *early-stage* and *late-stage* feature-binding, with the latter as critical for gating into VWM. Thus, when describing or discussing the potentially obscure concepts of *encoding* or *representation* in VWM, relevant phenomena or mechanisms can be situated on the theory map to pinpoint what exactly is being considered.

The *theory map* can be a helpful starting point for the description of VWM phenomena in terms of specific mechanisms. As an example, take the retro-cue effect – a key focus of current research of VWM that has not yet been discussed in this review so far. The retro-cue effect is the enhanced memory for an item following a retroactive cue to the spatial location of that item (Griffin & Nobre, 2003; Landman et al., 2003). It has been a fruitful empirical effect to explore the workings of internal attention (see reviews by Myers et al., 2017; Souza & Oberauer, 2016). Many potential explanations have been offered for the retro-cue effect – protection from time-based decay, prioritization for comparison, removal of non-cued information, attentional strengthening or refreshing, retrieval head start, and protection from perceptual interference as categorized by Souza and Oberauer (2016). Here, the *theory map* may help researchers detail the specific mechanisms with each of these potential explanations, and thereby perform more definitive empirical tests or provide better formal models. For example, one could define the removal of non-cued information mechanistically as the complete loss of the token indexing that content, preventing its retrieval into the focus of attention – similar to *distraction* (Z. Zhang & Lewis-Peacock, 2023). However, this may not always occur – in other empirical conditions, the non-cued items may instead be held in latent states (not presently represented in the binding pool or *focus of attention*) but still indexed as tokens, and thus, can be retrieved but perhaps with some cost – similar to *distortion* (Fukuda et al., 2022). Of note, the *theory map* reminds the researcher that the *focus of attention* is one component of the entire VWM system – one that is an interface for many possible VWM phenomena. Thus, it is an important consideration in defining potential boundary conditions for when different retro-cue effects might occur.

It is expected that there will be disagreement with various aspects of the *theory map* as I have presented it here – perhaps with the connection of specific mechanisms to VWM phenomena that I have laid out in this review, or with the possible agreement of existing models that I have implied with the map. I believe these are seeds for impactful discussions that will inform empirical research that more incisively tests existing theories. I hope the map provides a common starting point for discussions across our field, getting researchers who focus on disparate but related phenomena, or use entirely different methods and approaches, on the same page. This may be a catalyst for our field to take a model-oriented approach to empirical research that is grounded in theory, clearly specifying the connection between phenomena and the mechanisms of tested models (Borsboom et al., 2021; Devezer & Buzbas, 2023; Oberauer & Lewandowsky, 2019). From these discussions, one could imagine that the VWM field identifies that it is not yet ready to conduct certain empirical tests – that innovation of measures and methods, or formalization or computational implementation of models are needed – perhaps spurring collaboration across labs on common goals. In this way, perhaps the field can progress on the challenge to shape and determine a complete model of VWM (Popov, 2023).

I would like to encourage researchers to practice *counterinduction* – we should seek to strengthen all competing models, rather than promoting a pet theory or model (Feyerabend, 2020). We should avoid the *toothbrush problem* (coined by Watkins, 1984): “Psychologists treat other people’s theories like toothbrushes – no self-respecting person wants to use anyone else’s.” (Mischel, 2008). One way that this may be achieved is through so-called *adversarial collaborations,* where researchers with differing theoretical views commit to collaborating on an empirical test (Cowan et al., 2020). This sort of collaboration is supposed to foster an understanding of opposing viewpoints, and give rise to a new theoretical position that unifies these views. Whatever the form of coordinated discussion on VWM, my theory map here may be a useful device to recognize where viewpoints specifically differ, and/or potential ways they may be in accordance, like I have demonstrated in this review with discrete-slots and variable precision models, and object-based and feature-based accounts. Our field would do well to readily develop and apply the best versions of various existing models, in the hopes of a truly consequential test of theories.

Conclusion

The aim of this review was to encourage development of a theoretical framework on which to ground research of visual working memory. I created a theory map using the broadly encompassing MLR model (Hedayati et al., 2022) as a scaffold to describe and compare current VWM phenomena and models. The hope is for the wider field to use the map as a helpful device to situate and promote further development of theories and models of VWM. By providing the map as a starting common point, more fruitful discussions and definitive experiment designs are enabled. I believe the map will help clarify the necessarily imprecise verbal definitions, reveal hidden intuitions, enable more specific descriptions of current models of VWM and the mechanisms through which they connect to empirical phenomena. Differences in intuitions or models about VWM phenomena may then be more specifically identified, and this may lead to more definitive studies from which we can more accurately determine the workings of the VWM system.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Alvarez, G. A., & Cavanagh, P. (2004). The Capacity of Visual Short-Term Memory Is Set Both by Visual Information Load and by Number of Objects. *Psychological Science*, *15*, 106–111.

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful. *Journal of Cognitive Neuroscience*, *33*(5), 902–918.

Atkinson, R. C., & Shiffrin, R. M. (1968). *Human Memory: A Proposed System and its Control Processes* (Vol. 2, pp. 89–195). Elsevier. https://doi.org/10.1016/s0079-7421(08)60422-3

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. *Psychological Science*, *18*(7), 622–628. https://doi.org/10.1111/j.1467-9280.2007.01949.x

Balaban, H., Drew, T., & Luria, R. (2019). Neural evidence for an object-based pointer system underlying working memory. *Cortex*, *119*, 362–372. https://doi.org/10.1016/j.cortex.2019.05.008

Balaban, H., & Luria, R. (2016). Integration of Distinct Objects in Visual Working Memory Depends on Strong Objecthood Cues Even for Different-Dimension Conjunctions. *Cerebral Cortex*, *26*, 2093–2104.

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. (2019). *Correspondence between population coding and psychophysical scaling models of working memory* (p. 699884). bioRxiv. https://doi.org/10.1101/699884

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.

Bays, P. M., Schneegans, S., Ma, W. J., & Brady, T. (2022). *Representation and computation in working memory*. PsyArXiv. https://doi.org/10.31234/osf.io/kubr9

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–1631. https://doi.org/10.1016/j.neuropsychologia.2010.12.023

Boettcher, S. E., Gresch, D., Nobre, A. C., & van Ede, F. (2021). Output planning at the input stage in visual working memory. *Science Advances*, *7*(13), eabe8212.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4–4. https://doi.org/10.1167/11.5.4

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464. https://doi.org/10.1073/pnas.1520027113

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

Chung, Y. H., Brady, T. F., & Störmer, V. S. (2023). No Fixed Limit for Storing Simple Visual Features: Realistic Objects Provide an Efficient Scaffold for Holding Features in Mind. *Psychological Science*, 09567976231171339. https://doi.org/10.1177/09567976231171339

Chunharas, C., & Brady, T. (2023). *Chunking, attraction, repulsion and ensemble effects are ubiquitous in visual working memory*. PsyArXiv. https://doi.org/10.31234/osf.io/es3b8

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., Naveh-Benjamin, M., Barrouillet, P., Camos, V., & Logie, R. H. (2020). How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration. *Perspectives on Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Devezer, B., & Buzbas, E. O. (2023). *Rigorous exploration in a model-centric science via epistemic iteration*. MetaArXiv. https://doi.org/10.31222/osf.io/qe46u

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology*, *85*, 30–42. https://doi.org/10.1016/j.cogpsych.2016.01.002

Erez, J., Cusack, R., Kendall, W., & Barense, M. D. (2016). Conjunctive Coding of Complex Object Features. *Cerebral Cortex*, *26*(5), 2271–2282. https://doi.org/10.1093/cercor/bhv081

Feyerabend, P. (2020). *Against method: Outline of an anarchistic theory of knowledge*. Verso Books.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 1–12. https://doi.org/10.1167/11.12.3

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 27–27. https://doi.org/10.1167/10.12.27

Fougnie, D., Cormiea, S. M., & Alvarez, G. A. (2012). Object-Based Benefits Without Object-Based Representations. *Journal of Experimental Psychology: General*, *142*(3), 621–626. https://doi.org/10.1037/a0030300

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

Fukuda, K., Pereira, A. E., Saito, J. M., Tang, T. Y., Tsubomi, H., & Bae, G.-Y. (2022). Working memory content is distorted by its use in perceptual comparisons. *Psychological Science*, *33*(5), 816–829.

Gao, T., Gao, Z., Li, J., Sun, Z., & Shen, M. (2011). The perceptual root of object-based storage: An interactive model of perception and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1803.

Gao, Z., Li, J., Wu, J., Dai, A., Liao, H., & Shen, M. (2022). Diverting the focus of attention in working memory through a perceptual task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *48*(6), 876–905. https://doi.org/10.1037/xlm0001112

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

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

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Gunseli, E., van Moorselaar, D., Meeter, M., & Olivers, C. N. L. (2015). The reliability of retro-cues determines the fate of noncued visual working memory representations. *Psychonomic Bulletin & Review*, *22*(5), 1334–1341. https://doi.org/10.3758/s13423-014-0796-x

Hakim, N., deBettencourt, M. T., Awh, E., & Vogel, E. K. (2020). Attention fluctuations impact ongoing maintenance of information in working memory. *Psychonomic Bulletin & Review*, *27*(6), 1269–1278. https://doi.org/10.3758/s13423-020-01790-z

Hardman, K. O., & Cowan, N. (2015). Remembering complex objects in visual working memory: Do capacity limits restrict objects or features? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 325–347. https://doi.org/10.1037/xlm0000031

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical Working Memory Representations are used in Delayed Estimation of Continuous Colors. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(1), 30–54. https://doi.org/10.1037/xhp0000290

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, *6*(5), Article 5. https://doi.org/10.1038/s41562-021-01264-9

Heuer, A., & Rolfs, M. (2021). Incidental encoding of visual information in temporal reference frames in working memory. *Cognition*, *207*, 104526. https://doi.org/10.1016/j.cognition.2020.104526

Huang, L. (2020). Unit of visual working memory: A Boolean map provides a better account than an object does. *Journal of Experimental Psychology: General*, *149*, 1–30. https://doi.org/10.1037/xge0000616

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. https://doi.org/10.1016/0010-0285(92)90007-O

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, *24*(3), 326–338.

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164. https://doi.org/10.1016/S0042-6989(02)00402-9

Li, A. Y., Fukuda, K., & Barense, M. D. (2022). Independent features form integrated objects: Using a novel shape-color “conjunction task” to reconstruct memory resolution for multiple object features simultaneously. *Cognition*, *223*, 105024. https://doi.org/10.1016/j.cognition.2022.105024

Liang, J. C., Erez, J., Zhang, F., Cusack, R., & Barense, M. D. (2020). Experience Transforms Conjunctive Object Representations: Neural Evidence for Unitization After Visual Expertise. *Cerebral Cortex*, *30*(5), 2721–2739. https://doi.org/10.1093/cercor/bhz250

Lin, H.-Y., & Oberauer, K. (2022). An interference model for visual working memory: Applications to the change detection task. *Cognitive Psychology*, *133*, 101463. https://doi.org/10.1016/j.cogpsych.2022.101463

Lin, Y., & Fougnie, D. (2022). No evidence that the retro-cue benefit requires reallocation of memory resources. *Cognition*, *229*, 105230. https://doi.org/10.1016/j.cognition.2022.105230

Logie, R. H., Belletier, C., & Doherty, J. M. (2021). Integrating theories of working memory. In *Working memory: State of the science* (pp. 389–429). Oxford University Press.

Logie, R. H., Camos, V., & Cowan, N. (2020). *Working Memory: The state of the science*. Oxford University Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

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

Maatman, F. O. (2021). *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It*. PsyArXiv. https://doi.org/10.31234/osf.io/puqvs

Markov, Y. A., Tiurina, N. A., & Utochkin, I. S. (2019). Different features are stored independently in visual working memory but mediated by object-based representations. *Acta Psychologica*, *197*, 52–63.

McMaster, J. M. V., Tomić, I., Schneegans, S., & Bays, P. M. (2022). Swap errors in visual working memory are fully explained by cue-feature variability. *Cognitive Psychology*, *137*, 101493. https://doi.org/10.1016/j.cogpsych.2022.101493

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006x.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Mischel, W. (2008). The Toothbrush Problem. *APS Observer*, *21*. https://www.psychologicalscience.org/observer/the-toothbrush-problem

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

Nasrawi, R., & van Ede, F. (2022). Planning the potential future during multi-item visual working memory. *Journal of Cognitive Neuroscience*, *34*(8), 1534–1546.

Navon, D. (1984). Resources—A theoretical soup stone? *Psychological Review*, *91*, 216–234. https://doi.org/10.1037/0033-295X.91.2.216

Ngiam, W. X. Q., Brissenden, J. A., & Awh, E. (2019). “Memory compression” effects in visual working memory are contingent on explicit long-term memory. *Journal of Experimental Psychology: General*, *148*(8), 1373. https://doi.org/10.1037/xge0000649

Ngiam, W. X. Q., Foster, J. J., Adam, K. C. S., & Awh, E. (2022). Distinguishing guesses from fuzzy memories: Further evidence for item limits in visual working memory. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-022-02631-y

Ngiam, W. X. Q., Khaw, K. L. C., Holcombe, A. O., & Goodbourn, P. T. (2018). Visual working memory for letters varies with familiarity but not complexity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000682

Ngiam, W. X. Q., Loetscher, K., & Awh, E. (2023). *Object-based encoding constrains storage in visual working memory*. PsyArXiv. https://doi.org/10.31234/osf.io/mc5p9

Nobre, A. C. (2022). Opening Questions in Visual Working Memory. *Journal of Cognitive Neuroscience*, *35*(1), 49–59. https://doi.org/10.1162/jocn_a_01920

Nosofsky, R. M., & Donkin, C. (2016). Qualitative contrast between knowledge-limited mixed-state and variable-resources models of visual change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1507.

Nosofsky, R. M., & Gold, J. M. (2018). Biased guessing in a complete-identification visual-working-memory task: Further evidence for mixed-state models. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(4), 603.

Oberauer, K. (2023). Measurement models for visual working memory—A factorial model comparison. *Psychological Review*, *130*, 841–852. https://doi.org/10.1037/rev0000328

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*, 758–799. https://doi.org/10.1037/bul0000046

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*, 885–958. https://doi.org/10.1037/bul0000153

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

Oberauer, K., & Lin, H.-Y. (2023). *An Interference Model for Visual and Verbal Working Memory*.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the “strong-object” hypothesis. *Perception & Psychophysics*, *64*(7), 1055–1067. https://doi.org/10.3758/BF03194756

O’Reilly, R. C., Ranganath, C., & Russin, J. L. (2022). The Structure of Systematicity in the Brain. *Current Directions in Psychological Science*, *31*(2), 124–130. https://doi.org/10.1177/09637214211049233

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research*, *46*(28), 4646–4674.

Perfors, A., & Kidd, E. (2022). The Role of Stimulus-Specific Perceptual Fluency in Statistical Learning. *Cognitive Science*, *46*(2), e13100. https://doi.org/10.1111/cogs.13100

Popov, V. (2023). *If God Handed Us the Ground-Truth Theory of Memory, How Would We Recognize It?* PsyArXiv. https://doi.org/10.31234/osf.io/ay5cm

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. https://doi.org/10.1016/0010-0277(89)90014-0

Ricker, T. J., Souza, A. S., & Vergauwe, E. (2023). Feature identity determines representation structure in working memory. *Journal of Experimental Psychology: General*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xge0001427

Robinson, M. M., Williams, J. R., & Brady, T. (2022). *What does it take to falsify a psychological theory? A case study on recognition models of visual working-memory*. PsyArXiv. https://doi.org/10.31234/osf.io/7an3x

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

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *The Journal of Neuroscience*, *37*, 3913–3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., McMaster, J. M. V., & Bays, P. M. (2023). Role of time in binding features in visual working memory. *Psychological Review*, *130*, 137–154. https://doi.org/10.1037/rev0000331

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, *117*(34), 20959–20968. https://doi.org/10.1073/pnas.2004306117

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172.

Shin, H., & Ma, W. J. (2017). Visual short-term memory for oriented, colored objects. *Journal of Vision*, *17*(9), 12–12. https://doi.org/10.1167/17.9.12

Sone, H., Kang, M.-S., Li, A. Y., Tsubomi, H., & Fukuda, K. (2021). Simultaneous estimation procedure reveals the object-based, but not space-based, dependence of visual working memory representations. *Cognition*, *209*, 104579. https://doi.org/10.1016/j.cognition.2020.104579

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*, 1839–1860.

Souza, A. S., Overkott, C., & Matyja, M. (2021). Categorical distinctiveness constrains the labeling benefit in visual working memory. *Journal of Memory and Language*, *119*, 104242. https://doi.org/10.1016/j.jml.2021.104242

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

Teng, C., Fulvio, J. M., Jiang, J., & Postle, B. R. (2022). Flexible top-down control in the interaction between working memory and perception. *Journal of Vision*, *22*(11), 3. https://doi.org/10.1167/jov.22.11.3

Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. *Nature Human Behaviour*, *3*(8), 827–836. https://doi.org/10.1038/s41562-019-0640-4

Thyer, W., Adam, K. C. S., Diaz, G. K., Velázquez Sánchez, I. N., Vogel, E. K., & Awh, E. (2022). Storage in Visual Working Memory Recruits a Content-Independent Pointer System. *Psychological Science*, *33*(10), 1680–1694. https://doi.org/10.1177/09567976221090923

Tomić, I., & Bays, P. M. (2022). Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001172

Tomić, I., & Bays, P. M. (2023). *A dynamic neural resource model bridges sensory and working memory* (p. 2023.03.27.534406). bioRxiv. https://doi.org/10.1101/2023.03.27.534406

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial Comparison of Working Memory Models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963. https://doi.org/10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

van Ede, F. (2020). Visual working memory and action: Functional links and bi-directional influences. *Visual Cognition*, *28*(5–8), 401–413.

van Ede, F., & Nobre, A. C. (2023). Turning attention inside out: How working memory serves behavior. *Annual Review of Psychology*, *74*.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436

Watkins, M. J. (1984). Models as toothbrushes. *Behavioral and Brain Sciences*, *7*(1), 86–86. https://doi.org/10.1017/S0140525X00026303

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11–11.

Williams, J. R., Robinson, M. M., Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2022). You cannot “count” how many items people remember in visual working memory: The importance of signal detection–based measures for understanding change detection performance. *Journal of Experimental Psychology: Human Perception and Performance*, *48*, 1390–1409. https://doi.org/10.1037/xhp0001055

Xie, W., & Zhang, W. (2017). Familiarity Speeds Up Visual Short-term Memory Consolidation. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1207–1221. https://doi.org/10.1037/xhp0000355

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

Zhang, Z., & Lewis-Peacock, J. A. (2023). Prioritization sharpens working memories but does not protect them from distraction. *Journal of Experimental Psychology: General*, *152*, 1158–1174. https://doi.org/10.1037/xge0001309

Acknowledgements

My employment is funded by the National Institutes of Health R01 MH087214 research grant awarded to Edward Awh and Edward K Vogel. I would like to especially thank Brad Wyble for allowing the MLR model to be repurposed as a theoretical device and for continued conversations about aspects of this theoretical review. I would like to thank Piotr Styrkowiec, Vencislav Popov, Igor Utochkin and Philipp Musfeld for helpful discussions and their encouragement to write this paper.