WEP Warwick Evans Publishing

International Journal of Social Sciences and Public Administration

ISSN: 3005-9836 (Print), ISSN: 3005-9585 (Online) | Volume 9, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijsspa.v9n2.11
Journal homepage: https://ijsspa.org

A Review of Research on Working Memory

Yibo Zhang

Lancaster Mennonite School, Lancaster, PA 17602, USA

ABSTRACT

Working memory is a core cognitive system responsible for the temporary maintenance and manipulation of information essential for reasoning, learning, and problem solving. Research has developed multiple theoretical models, including Baddeley and Hitch's multi-component model, Cowan's embedded-processes framework, and Miller's capacity limits, to explain its structure and constraints. Behavioral paradigms, neuroscience techniques such as fMRI, EEG, and TMS, and animal models have advanced understanding of the neural mechanisms underlying working memory. Working memory impairments are closely linked to neurological and psychiatric disorders, and interventions such as cognitive training and brain stimulation show promise for improving working memory function.

KEYWORDS

Working Memory; Cognitive Models; Neural Mechanisms.

1. INTRODUCTION

The study of memory can be traced back to ancient Greek philosophers such as Aristotle and Plato, who developed early theories about memory. Plato compared memory to a wax tablet capable of receiving "imprints" from experience, while Aristotle emphasized the role of sensory experience, suggesting that memory depends on our senses. In the 20th century, psychologists began to adopt more systematic experimental methods and gradually formed modern theories of memory. In particular, the concept of working memory emerged as an important breakthrough in understanding human cognitive ability. Working memory (WM) refers to the ability to maintain and manipulate a limited amount of information over a short period of time. It has been described as "a limited amount of information that can be temporarily maintained in an accessible state, making it useful for many cognitive tasks" [1]. In contrast to long-term memory, working memory deals with information that is immediately relevant to ongoing cognitive activities. The term working memory is one of the most commonly used in psychology; it is often associated with constructs such as intelligence, information processing, executive function, comprehension, problem solving, and learning. The concept is ubiquitous in the field and has been defined in varying ways[1]. Evolutionary perspectives also consider why WM capacities might be constrained and how such constraints could have adaptive value[2]. WM is often contrasted with short-term memory (STM) in clinical and educational contexts; in practice, discussions of ADHD and learning frequently distinguish how WM goes beyond mere storage to active manipulation [3]. Working memory is not only essential in everyday life, but also plays a central role in complex cognitive tasks like learning, communication, decision making, and problem solving. Moreover, working memory research has become important in the context of brain disorders. For example, patients with Alzheimer's disease (AD), Parkinson's disease, and ADHD show significantly impaired working memory function, and some patients can improve cognitive performance through working memory training or pharmacological interventions. Modern

neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) have further advanced this field, providing new possibilities for early diagnosis and personalized treatment of neurological conditions.

Despite significant progress in working memory research, many mysteries and challenges remain. For example, it is still unclear how to accurately model individual differences in working memory capacity, how to optimize intervention strategies to restore memory function in patients with cognitive impairment, and how to effectively integrate new technologies (such as brain–machine interfaces) to study or augment working memory. With the advent of artificial intelligence and big data, researchers expect to gain a deeper understanding of the neural mechanisms of working memory and to develop novel applications. These advances may open new prospects for human cognitive enhancement and the treatment of memory disorders in the future.

2. THEORETICAL MODELS OF WORKING MEMORY

2.1. Baddeley and Hitch's Working Memory Model

2.1.1. Model Structure:

Baddeley and Hitch's original (1974) model of working memory consists of three main components: the central executive, the phonological loop, and the visuospatial sketchpad. The central executive is an attention-control module (analogous to a "watchdog") responsible for deploying attentional resources and coordinating the operation of the two subsystems. It determines which subsystem (phonological or visuospatial) will store incoming information and can retrieve information from each, essentially governing the whole working memory system. The phonological loop and the visuospatial sketchpad function as two "slave" subsystems under the central executive. As the core control unit, the central executive has limited capacity but flexible functionality. Baddeley likened the central executive to the "CEO" of the brain – it allocates attention and resources between tasks, and updates or integrates information from various sources. The central executive itself does not store specific information; rather, it manages the use of the two storage subsystems. For example, when we need to multitask, the central executive decides how to quickly switch between tasks, how to prioritize them, and when it is necessary to store or retrieve intermediate results using the phonological loop. (Notably, the original model did not elaborate in depth on the detailed operation of the central executive, so its definition remained relatively broad.) The phonological loop is responsible for the temporary storage and rehearsal of verbal or auditory information – essentially the brain's "inner ear" and "inner voice." The phonological loop consists of two parts: a phonological storage component, which briefly retains the sound-based representation of material, and an articulatory rehearsal mechanism, which refreshes these traces through silent repetition (an internal "rehearsal"). The phonological loop can encode information that is heard, or even seen, into a phonetic code via subvocal articulation. It plays an important role in language acquisition (for example, learning new vocabulary in one's native or a foreign language). The visuospatial sketchpad (also called the visuospatial notepad or visuospatial scratchpad) is used to store and manipulate visual and spatial information, such as images, colors, shapes, and the locations of objects in space. It is the brain's equivalent of a "mental sketch pad" that allows us to temporarily hold and manipulate visual scenes (for example, visualizing a route on a map or mentally writing down numbers during mental arithmetic). The visuospatial sketchpad is important for tasks such as mental rotation, navigation, and visual imagery. Some researchers have subdivided it into separate visual and spatial components, although Baddeley's 1974 model treated it as a single system. Although the 1974 model successfully explained many findings about short-term memory, some cognitive functions could not be fully accounted for by only three components. To address these limitations, Baddeley expanded the model in 2000 by adding a fourth component, the episodic buffer. The episodic buffer is a limited-capacity temporary storage system that integrates information from different sources into a unified "episode" or multimodal representation. It is termed a "buffer"

because it acts as an intermediary between other memory systems, and "episodic" because it can store integrated episodes or scenes (for example, the combined sight, sound, and meaning of a story, rather than isolated bits). The episodic buffer binds information from the two subsystems (phonological loop and visuospatial sketchpad) together with relevant information from long-term memory into a coherent episode. For instance, when reading a novel, we use the episodic buffer to integrate the words we read (phonological information) with a mental visualization of the characters and settings (visuospatial information) and our prior knowledge from long-term memory, thereby forming a holistic understanding of the story. The introduction of the episodic buffer considerably increased the explanatory scope of the model, allowing it to account for the integration of information across modalities and interfaces with long-term memory. Recent Perspectives: Baddeley's multi-component model has had remarkable longevity and influence. Even fifty years on, it remains a dominant framework for understanding working memory. Subsequent developments of the model have increased its scope, depth, and applications while retaining its core features [4]. Comparisons with a growing number of alternative models suggest that all models must explain a common set of basic phenomena, and in fact there are more similarities than differences between many models [4]. Differences between models tend to attract attention, but they may not reflect the most critical issues for future research – notably, the nature of executive control [4]. The longevity of the multi-component model reflects not only the central importance of working memory in cognition but also the usefulness of a simple, robust framework for further theoretical development and applications [4]. Furthermore, the multi-component model has proven useful in educational settings. For example, understanding the distinct roles of the phonological loop and visuospatial sketchpad has helped educators identify cognitive differences in students. Children with learning disabilities often have working memory deficits – for instance, dyslexia and specific language impairments are associated with poor phonological working memory, leading to difficulty following multi-step instructions in class. By breaking complex instructions into smaller steps and providing external memory aids (charts, checklists, etc.), teachers can reduce working memory load and improve learning outcomes. In short, working memory models like Baddeley's provide a framework for understanding and intervening in cognitive differences among learners[5].

2.2. Cowan's Embedded-Processes Model

American psychologist Nelson Cowan proposed an alternative but equally important working memory framework known as the embedded-processes model. This model emphasizes the dynamic relationship between working memory and long-term memory, suggesting that working memory is not a completely separate short-term storage system but rather the activated portion of long-term memory, combined with a limited-capacity focus of attention. According to Cowan, when the brain needs to process information, it activates relevant knowledge in long-term memory. All activated information (in transient accessible state) constitutes a broad working memory set. However, only a small subset of this activated information enters the focus of attention, which is the content we are consciously aware of and actively processing. Because attentional resources are limited, this focus can hold only about 3–5 distinct items at once – a refinement of the traditional "7±2" rule, emphasizing that the true capacity may be closer to 3–5 units under controlled conditions. Thus, Cowan's model underscores that working memory capacity is fundamentally constrained by attentional limits. Cowan also emphasized the role of executive control in managing working memory. An individual's ability to deliberately control their focus – for example, to resist distractions and retain task-relevant information – determines which information enters and remains in focal attention. This idea overlaps with Baddeley's notion of a central executive, though Cowan's model does not separate working memory into distinct subsystems (verbal, visuospatial, etc.) as Baddeley's model does. Instead, Cowan conceptualizes working memory and long-term memory as differing only in the level of activation and focus, not in absolute separation. There is no rigid structural boundary between working memory and long-term memory in this view; working memory is simply the currently activated subset of long-term memory, especially the part within the focus of attention. In

sum, Cowan's model highlights the central role of attentional focus in retaining and processing information, and it blurs the distinction between short-term and long-term memory by viewing them along a continuum of activation [1].

2.3. Miller's "Magic Number 7±2" and Chunking

2.3.1. The Birth of the Theory:

In 1956, cognitive psychologist George A. Miller published his famous paper "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" in Psychological Review. This work provided the first quantitative description of short-term memory capacity. Miller noted humorously that he had been "haunted by a certain number" – the number 7 (give or take 2) – which kept appearing in various experiments, sometimes a bit larger or smaller, but never by much. This "magical number 7±2" appeared to be a general limit on people's ability to process or remember information in the short term. Miller's goal was to point out that humans have a severe capacity bottleneck for information processing. The 7±2 figure came from multiple sources he compiled, including studies of immediate memory span and absolute judgment of unidimensional stimuli. By the mid-1950s, experiments had shown that the typical adult could immediately recall about seven random digits (the digit span task). Miller systematically generalized these observations and framed them in information-theoretic terms. Psychologists were influenced by information theory at the time, measuring information in "bits." Despite large differences in the amount of information per item, the number of items people could store in short-term memory tended to fall in the same range of 5–9. Miller's 7±2 rule emerged from this pattern, suggesting that short-term memory is limited by number of items rather than total information content [1].

2.3.2. Experimental Evidence:

Miller's claim was not made lightly; it was based on a variety of experimental findings. In his article, he discussed capacity limits in two types of tasks: (1) absolute judgment of a single stimulus dimension, and (2) short-term memory span tasks. He also noted observations on instantaneous memory (subitizing). In absolute judgment tasks, a person must classify a stimulus (e.g., a tone or color shade) into one of several categories. Performance drops as the number of possible categories increases. Miller observed that people could reliably categorize only about 5–9 levels on a single sensory continuum (like distinguishing about five to nine different loudness levels of a tone). Beyond that, errors rose markedly, indicating a capacity limit in differentiating stimulus values. The short-term memory span task provides another piece of evidence. In the digit span test (a classic memory span measure), a person hears or sees a list of random digits and must immediately recall them in order. Numerous studies showed that adults can repeat back roughly seven digits reliably. Interestingly, this span of 7±2 items holds across different types of material, even when the amount of information per item varies. For example, people can remember about seven letters or words, even though words carry much more information than digits. They might recall ~9 binary digits (bits) but only ~5 English words, yet in each case the item count is in the same ballpark. Miller reasoned that short-term memory is limited by number of items (chunks), not by the total information content. He introduced the term chunk to describe the basic units of short-term memory – meaningful groupings of information. Regardless of how much information each chunk contains, people can only hold about 7±2 chunks at a time.

2.3.3. The Concept and Significance of "Chunks":

One of Miller's key contributions was the concept of chunking. A chunk is a collection of elements that are strongly associated with one another but only weakly associated with elements in other chunks. By organizing information into larger units or chunks, people can overcome some of the capacity limitations of working memory. For example, remembering a seven-digit phone number is easier if grouped into chunks (e.g., "area code – three-digit – four-digit") than as a raw sequence of 7 individual digits. Chunking effectively reduces the number of items to remember by increasing the

information contained in each item. However, even with chunking, the capacity limit of about 7±2 chunks still holds; training and expertise mainly help one create bigger or more efficient chunks. Miller's discovery of the chunk limit was foundational - it alerted researchers to inherent limits in human information processing and spurred further research into how memory capacity can be measured, expanded, or optimized. Over time, researchers have refined Miller's estimate (for instance, Cowan argued that the true capacity when avoiding rehearsal is closer to 4±1 items). Nonetheless, the idea of a fixed capacity limit and the mechanism of chunking remain central in working memory theory. Principles based on the "magic number" are even applied in practical settings; for instance, instructional designers often follow the rule of not presenting more than about 5–7 new items at once to avoid overloading learners [1,3].

3. RESEARCH METHODS AND PARADIGMS OF WORKING MEMORY

3.1. Behavioral Experimental Methods

3.1.1. Dual-Task Paradigm:

The dual-task paradigm is a classic behavioral design in working memory research, in which participants are asked to perform two tasks simultaneously. The goal is to investigate how doing two tasks at once leads to sharing or interference of cognitive resources. In a famous demonstration, Baddeley and Hitch had subjects perform a reasoning task (e.g. verifying logical sentences) while simultaneously remembering a string of digits. The results showed that if the two tasks involve different modalities or domains (for example, one is a verbal task and the other is visuospatial), people can perform them together almost as well as separately. However, if both tasks rely on the same modality (e.g. both require verbal processing), then performance deteriorates significantly under dual-task conditions. This pattern of interference supports the idea that working memory has multiple specialized subsystems: tasks compete for resources only if they use the same subsystem. In experimental designs, researchers typically control the difficulty of each single task and include a single-task control condition to measure the performance cost of doing both tasks at once. Since its introduction in the 1970s, the dual-task paradigm has become a standard method in working memory research. Using this approach, Baddeley and Hitch were able to challenge the idea of a unitary short-term memory store and instead provide evidence for their multi-component model. Many subsequent studies have used dual-task interference to explore the functions of the central executive and the independence of different working memory buffers. For example, patients with Alzheimer's disease show relatively normal performance on single tasks but have great difficulty when attempting dual tasks, suggesting an impairment in their central executive system [6].

3.1.2. Auditory Sequential Tasks:

A typical example of an auditory sequential working memory task is the digit span test. In this task, participants listen to a string of random numbers and then immediately repeat them back, which tests the capacity of working memory for sequential verbal information. The digit span task comes in two forms: forward digit span (repeating the sequence in the original order) and backward digit span (repeating in reverse order). Forward digit span primarily measures short-term storage capacity, whereas backward digit span additionally requires manipulation of information (thus engaging executive control). Classic findings show that the average adult has a forward digit span of about 7±2 items – reflecting Miller's magic number. This capacity limit is thought to reflect the decay and rehearsal limits of the phonological loop (short-term auditory memory). The digit span task has a long history, dating back to 19th-century memory research (e.g., Jacobs, 1887), and it was later incorporated into standard IQ tests and neuropsychological assessments. It provides a quick index of working memory capacity, though it has limitations. One limitation is that simple forward span involves only passive retention and repetition, so it measures short-term memory more so than active "working" memory. Another is that participants can use mnemonic strategies (like grouping digits

into chunks or using semantic associations) to improve their span, so the task does not purely reflect raw capacity. Additionally, cultural and linguistic factors can affect performance – for instance, the word length effect (digit names that are longer or take longer to pronounce lead to lower span because of slower rehearsal). Despite these caveats, digit span remains a foundational measure. Using span tasks, researchers have discovered important phenomena such as the vulnerability of verbal short-term memory to phonological interference (e.g., background speech sounds can disrupt digit recall). Such findings support the existence of a phonological storage system that can be selectively interfered with by irrelevant phonetic input.

3.1.3. Visuospatial Tasks:

A classic test of visuospatial working memory is the Corsi block tapping task. In this task, an experimenter taps out a sequence on an array of spatially separated blocks, and the participant must reproduce the sequence by tapping the blocks in order (forward Corsi span) or in reverse order (backward Corsi span). The forward Corsi span gauges the capacity of visuospatial short-term memory, with average spans around 5 items for adults – slightly lower than typical digit span, reflecting limits in the visuospatial sketchpad. The backward version requires additional mental manipulation, engaging the central executive. The Corsi task was first introduced by researcher Philip Corsi in 1972 to assess spatial memory in patients, and it has since become a standard tool for measuring visuospatial working memory in both clinical and research contexts. By comparing an individual's performance on digit span (verbal) and Corsi span (spatial), one can assess working memory capacity for different types of material. Often, these capacities are somewhat independent – for example, some patients with brain damage show severely impaired digit span but near-normal Corsi span, suggesting separate verbal and spatial short-term memory stores. Such double dissociations were crucial evidence for the multicomponent structure of working memory. However, the Corsi task also has its limitations: performance can be affected by perceptual-motor factors (like eyesight or motor coordination), and like digit span it mainly provides a single capacity measure without insight into strategic processes or complex manipulation. Nevertheless, the accumulation of behavioral findings from simple span tasks laid the groundwork for establishing and validating working memory models. It demonstrated, for instance, that interference between simultaneous tasks is modality-specific, aligning with Baddeley's proposal that the phonological loop and visuospatial sketchpad are independent.

3.2. Neuroscience Techniques

3.2.1. Functional MRI (fMRI)

fMRI has been widely used in working memory research to identify which brain regions are active during working memory tasks. This technique measures blood-oxygen-level-dependent (BOLD) signals as an indirect indicator of neural activity. A typical experiment involves having participants perform a working memory task (such as retaining shapes or locations) while in the MRI scanner, and then analyzing which brain areas show increased activation. Numerous studies have found that working memory tasks consistently engage a fronto-parietal network in the brain – particularly regions in the dorsolateral prefrontal cortex (PFC) and the parietal cortex. For example, visuospatial working memory tasks tend to produce stronger activation in the right parietal cortex (consistent with the right hemisphere's role in spatial processing), whereas verbal working memory tasks elicit more activation in left-hemisphere frontoparietal areas, including language-related regions. These neuroimaging findings align with neuropsychological evidence: damage to the right parietal lobe impairs spatial short-term memory, while damage to the left parietal (especially near the temporoparietal junction) impairs verbal short-term memory. Critically, fMRI studies have shown that increasing the working memory load (the number of items to maintain) leads to increased activity in frontal and parietal regions, up to a point. When the number of items approaches an individual's capacity limit, the BOLD activity plateaus or peaks. For instance, if a person can hold 3–4 items,

parietal cortex activation will rise as the memory set size increases from 1 to 3 items, and then level off or even drop when a 4th or 5th item is added that exceeds capacity. This neural saturation effect mirrors behavioral capacity limits and supports the view that working memory maintenance depends on sustained neural activity for a limited number of items. Furthermore, high-resolution fMRI has allowed researchers to distinguish the subcomponents of working memory maintenance: for example, maintaining spatial locations preferentially activates superior parietal and dorsal frontal areas, whereas maintaining visual details (like faces or complex objects) engages more ventral frontal and temporal areas. In summary, fMRI has mapped out the network of brain regions that support working memory and has confirmed that multiple regions (especially in PFC and parietal cortex) jointly enable the short-term holding and processing of information. The strength of fMRI is its high spatial resolution, which helps identify specific brain areas involved in different aspects of working memory (encoding, maintenance, retrieval, etc.). Its main limitation is low temporal resolution (on the order of seconds), which makes it hard to capture rapid neural dynamics. Also, the BOLD signal is an indirect measure of neural firing. Researchers address these limitations by combining fMRI with other techniques (such as EEG for timing, or faster imaging sequences) to get a more complete picture [6].

3.2.2. Electroencephalography (EEG)

EEG records the brain's electrical activity with millisecond temporal resolution, making it ideal for tracking the time-course of cognitive processes in working memory. EEG studies of working memory often use event-related potentials (ERPs) to identify specific brain wave patterns associated with different stages of a memory task (stimulus encoding, maintenance delay, retrieval, etc.). For example, during the maintenance period of a visual working memory task, a slow-wave negative potential known as the contralateral delay activity (CDA) can be observed at posterior electrode sites. The amplitude of the CDA increases as more items are held in memory, and it plateaus when the number of items reaches the individual's capacity limit. Vogel and Machizawa (2004) first demonstrated a clear relationship between CDA amplitude and memory capacity: as subjects tried to remember more items, CDA grew but then stopped increasing at about 3–4 items, suggesting that no additional items could be effectively maintained beyond that point. This provided neural evidence for the item limit in visual working memory. EEG has also revealed characteristic changes in neural oscillations during working memory tasks. For instance, theta-band oscillations (~4–8 Hz), especially in frontal regions, tend to increase in power when people are maintaining multiple items. This has been linked to focused attention and perhaps an internal rehearsal or sequencing mechanism during maintenance. Gamma-band activity (>30 Hz) shows brief bursts of synchrony during memory encoding and maintenance, which may reflect the binding and communication of information across brain regions. Some theories propose an interaction between theta and gamma oscillations – for example, the theta–gamma coupling model suggests that the brain could use lower-frequency theta cycles to segment high-frequency gamma activity into discrete time "slots," effectively allowing multiple items to be active in sequence within a single theta cycle (as hypothesized by Lisman & Idiart). This is one way the brain might represent several items at once without them interfering. Indeed, EEG data have shown phase coupling between theta and gamma rhythms during working memory tasks, supporting this idea. In addition to studies of healthy individuals, EEG is used to examine working memory in clinical populations. Differences in certain ERP components or oscillatory patterns during working memory tasks have been observed in disorders such as schizophrenia, ADHD, and anxiety, suggesting underlying neural mechanism differences in how these brains handle working memory. The major advantage of EEG is its ability to sensitively track the real-time neural dynamics of working memory – revealing, for example, how quickly the brain updates working memory content (reflected in components like the P300 when new information must replace old). The main drawback is its limited spatial resolution: scalp electrodes cannot pinpoint deep or localized brain sources with high precision. Thus, EEG findings are often complemented by source localization algorithms or by combining EEG with fMRI to map the spatial generators of the observed signals. Notably, EEG is not just a passive measure; it can be paired with brain stimulation. For example, researchers have combined EEG recording with transcranial alternating current stimulation (tACS) to experimentally modulate brain

oscillations. In a recent study, slowing the brain's theta oscillation frequency via tACS was found to enhance working memory capacity. By synchronizing stimulation with the brain's rhythms, this approach demonstrated a causal link between theta cycle dynamics and memory performance (slower theta cycles may allow more gamma cycles – hence more items – to be represented within each theta cycle) [7]. Such findings open the door to new interventions for boosting working memory by targeting its underlying neural oscillatory mechanisms.

3.2.3. Transcranial Magnetic Stimulation (TMS)

TMS is a non-invasive brain stimulation technique that uses magnetic pulses to transiently stimulate or disrupt activity in specific brain regions. In working memory research, TMS is often applied to regions like the prefrontal or parietal cortex to test their causal importance for the task at hand. For instance, transiently disrupting the right parietal lobe or left inferior frontal gyrus via TMS has been shown to selectively impair memory for combined color–orientation visual patterns, while having a lesser effect on memory for single features. This suggests those regions are specifically important for binding multiple features in working memory. Similarly, TMS over the dorsolateral prefrontal cortex (DLPFC) often reduces working memory span or slows responses, supporting the DLPFC's critical role in actively maintaining and manipulating information. TMS can also be combined with EEG or fMRI to examine how stimulation affects brain networks. In TMS-EEG studies, the EEG response following a TMS pulse can reveal changes in functional connectivity – for example, a TMS pulse to the PFC might induce EEG signals reflecting its interaction with parietal cortex, shedding light on how the working memory network dynamically reorganizes. In a recent study combining TMS with fMRI, researchers found that TMS to frontal cortex changed the functional connectivity in a dose-dependent manner: the degree of connectivity change was proportional to the strength of the electric field induced by the TMS pulse. This indicates that stronger stimulation leads to greater network modulation, and it highlights how TMS can be used to probe the dynamics of the working memory network in a controlled way [8]. The advantage of TMS is that it allows causal inferences – it can show that a certain brain area is necessary for a specific working memory function by observing deficits when that area is temporarily offline. For example, applying repetitive TMS during the delay period of a memory task can test whether sustained neural activity in that interval is essential for maintenance: if TMS during the delay impairs performance, it supports the idea that ongoing activity (like persistent firing in PFC) is needed to keep information in mind. Indeed, such experiments have found that interrupting PFC activity during the maintenance phase causes forgetting of the remembered item, reinforcing the "persistent activity" mechanism of working memory. On the other hand, if some working memory content survives TMS disruptions, it raises the possibility of alternate storage mechanisms (e.g., transient synaptic changes that hold information silently). As discussed later, emerging theories consider that working memory might sometimes be supported by short-term synaptic plasticity rather than continuous firing. TMS studies contribute to this debate by probing how and when interfering with neural activity affects memory.

3.3. Animal Models

3.3.1. Experimental Methods in Animals

Because non-human primates and even rodents can transiently hold information, researchers have developed animal models to explore the biological mechanisms of working memory under controlled conditions. One classic task is the delayed response paradigm. Here, an animal (often a monkey) is first shown a target stimulus or cue (for example, food hidden in one of several locations). After a delay period of several seconds during which the stimulus is removed, the animal is allowed to make a response (such as choosing a location) to obtain a reward. The animal must "remember" the location of the food during the blank delay interval to succeed. As early as the 1930s, experiments by Jacobsen and colleagues found that removing or damaging the prefrontal cortex in monkeys severely impaired

their performance on delayed response tasks. This provided some of the first evidence that the prefrontal cortex is critical for short-term memory.

Another related paradigm is delayed matching-to-sample (and its variant, delayed non-matching-to-sample). In this task, the animal is shown a sample stimulus, then after a short delay, two choice stimuli are presented. The animal must choose the one that matches the sample (or the one that is different, in the non-match version) to get a reward. This tests the animal's working memory for the sample's identity across the delay. For example, a monkey might see a particular picture, then after a delay must select the same picture out of two options. If the delay is too long or distracting stimuli are introduced, accuracy declines, indicating the time course and fragility of the memory trace. Such tasks also show that the probability of error increases as the delay lengthens, often in an exponential fashion, consistent with a time-based decay of the working memory trace. Experimental manipulations like temporarily cooling or pharmacologically inactivating the prefrontal cortex cause accelerated forgetting in these tasks, again underscoring the role of that region in maintaining information over delays.

In rodents, one common model is the spatial working memory maze task. For instance, in the radial arm maze, a rat is placed in the center of multiple arms radiating outwards, each arm potentially baited with food. The rat must remember which arms it has already visited (and depleted of food) so that it does not waste time revisiting them. This requires short-term memory for recently visited locations. Normal rats are quite efficient at avoiding revisits, whereas rats with hippocampal or prefrontal cortex lesions tend to make more repeat visits to empty arms, indicating a deficit in working memory for spatial locations. This highlights the importance of these brain regions (particularly the hippocampus in rodents, which plays a role somewhat akin to prefrontal cortex involvement in primates) for spatial working memory.

Animal studies have also allowed researchers to record directly from neurons during working memory tasks. Seminal work by Fuster & Alexander (1971) and by Funahashi, Bruce, & Goldman-Rakic (1989) involved training monkeys on tasks like oculomotor delayed response (where the monkey must remember a spatial location over a short delay). Neurons in the prefrontal cortex were found that not only respond when a stimulus is presented, but continue to fire throughout the delay period – even after the stimulus is gone – until the monkey is allowed to respond. This persistent firing is often selective: a given neuron might only sustain activity if the monkey is remembering a specific location or feature (its "preferred" direction or item), and remain quiet for other locations. Funahashi et al. coined the term "mnemonic scotoma" for the phenomenon that if a small region of PFC is lesioned, the monkey specifically loses short-term memory for the location corresponding to the preference of neurons in that region. In other words, each neuron or group of neurons has a kind of memory field, and damage causes an inability to remember items in that field. Approximately 80% of the delay-active neurons in these studies showed such spatial selectivity (often preferring contralateral space). These neurophysiological findings provided direct support for a neural correlate of working memory: when an animal holds something in mind, certain neurons maintain elevated activity to temporarily represent that information. This sustained delay period activity is regarded as a neural substrate of "active memory." Subsequent animal studies extended this concept beyond spatial memory – for example, finding neurons that can maintain information about an object identity, a color, or even an abstract rule during delays. Thus, the principle of persistent activity as a mechanism for holding information seems general [9].

3.3.2. Strengths and Challenges of Animal Research

Animal models offer insights into working memory that are difficult to obtain in humans, largely because of the ability to exert experimental control and use invasive techniques. In animal experiments, one can tightly control stimuli, administer thousands of trials to obtain stable performance measures, and even train animals to very high levels of expertise on specific tasks. More importantly, techniques like single-neuron recording allow researchers to observe the neural code of

working memory directly, and interventions like localized brain lesions, optogenetic manipulations (in rodents), or pharmacological injections can reveal the contributions of specific brain areas, circuits, or neurotransmitter systems. For example, by using pharmacological agents in monkeys, researchers discovered that blocking certain dopamine receptors in PFC impairs working memory, linking neurotransmitter modulators to memory maintenance [9]. Genetic tools in mice have enabled researchers to knock out particular receptors and observe the impact on working memory tasks (e.g., NMDA receptor knockdown affecting persistent firing). These approaches have traced working memory mechanisms down to specific receptor and molecular levels. Such findings have in turn guided human research – for instance, by identifying gene polymorphisms that might influence working memory capacity in the human population, or by suggesting targets for drug development to mitigate working memory deficits [10]. However, there are also challenges and caveats when using animal models for working memory. First, different species have different brain structures and cognitive repertoires. Rodents, for example, lack a well-developed lateral prefrontal cortex; their working memory functions are distributed across medial frontal areas and the hippocampus, which is somewhat different from primates (including humans) where the dorsolateral PFC is key. This means results from rodent studies must be extrapolated to humans with caution, especially for tasks that in humans rely on lateral PFC. Non-human primates (monkeys) are closer models for human frontal lobe function, but such studies are expensive and raise ethical considerations, limiting sample sizes. Second, certain forms of working memory that humans use (like verbal rehearsal or complex strategy use) cannot be easily replicated in animals. Animal tasks are often limited to spatial or simple feature memory, which cover important aspects but not the full spectrum of human working memory (e.g., we cannot test a monkey's working memory for a logical proposition or a sentence). Third, training animals to perform working memory tasks can be time-consuming, and their motivation or strategies may differ from humans'. Sometimes animals might solve tasks with unintended strategies (like using odor traces or subtle cues) if not carefully controlled. Lastly, while invasive methods provide a wealth of data, the interpretations can be complex: for example, a neuron's persistent firing might reflect working memory maintenance, or it might be encoding a planned movement, or an expectation of reward, etc. Researchers must design controls to dissociate these factors. Despite these challenges, animal research has profoundly shaped our understanding of working memory. Primate studies confirmed the existence of sustained neural firing in PFC as a correlate of active memory, a phenomenon that aligned beautifully with theoretical concepts of an "executive" holding information online. Rodent studies have extended this to a molecular level, showing how specific receptors (like NMDA receptors facilitating recurrent activity) are necessary for maintaining memory traces. These animal findings have motivated new theoretical models of working memory that incorporate persistent spiking, rhythmic oscillations, and even "activity-silent" mechanisms (short-term synaptic changes that do not require continuous firing). For instance, recent evidence suggests that after initial persistent firing, some information in working memory can be stored in latent synaptic states that do not require neurons to keep firing – a concept supported by studies demonstrating an interplay between ongoing activity and silent synaptic mechanisms in PFC. Such insights from animals are pushing the field to consider multiple mechanisms (both persistent and activity-silent) operating together to support working memory [10,11]. Overall, animal research provides a critical bridge between observable behavior and underlying neural processes, informing a more unified neuroscientific theory of working memory.

4. WORKING MEMORY IN BRAIN HEALTH AND DISEASE

Working memory research has significant applications in the diagnosis and treatment of brain disorders and in promoting brain health. Many neurological and psychiatric conditions – including Alzheimer's disease (AD), ADHD, depression, and schizophrenia) are associated with marked working memory impairments. These disorders often involve dysfunction in brain regions central to working memory, such as the prefrontal cortex, hippocampus, and parietal lobes. For example,

patients with Alzheimer's disease show severe deficits in short-term memory, reflecting neural damage that compromises the maintenance and manipulation of information. Patients with ADHD typically have reduced working memory capacity and impaired attention control, likely related to developmental delays or functional abnormalities in the prefrontal cortex. By using cognitive assessments and neuroimaging, researchers can characterize the specific working memory impairments in these conditions. Patterns of impairment – for instance, an ADHD patient failing to hold information in mind when distractions are present – can serve as early diagnostic markers. Indeed, behavioral tasks like the N-back (which taxes continuous working memory updating) or classic span tasks, as well as EEG and fMRI measures during those tasks, have been used to identify characteristic signatures of working memory dysfunction for conditions like ADHD. Such signatures can aid in early diagnosis or in monitoring disease progression and treatment effects. For ADHD, for example, a deficit in performing dual tasks or in suppressing irrelevant information on a memory task could be used as part of an assessment battery to detect executive function weaknesses early on [6].

Beyond diagnosis, an encouraging line of research is the use of targeted working memory training programs as interventions. There is growing evidence that specific training regimens (such as intensive practice on N-back tasks or complex span tasks, or strategy training to chunk and rehearse information) can lead to improvements in working memory performance. These improvements have been noted particularly in populations with mild cognitive impairment (MCI) and ADHD. For instance, multiple studies report that children with ADHD show gains in working memory and attentional tasks after several weeks of computerized cognitive training, though the degree of generalization to broader cognitive skills or academic performance can vary [12]. In a recent comprehensive review, Al-Saad et al. (2021) concluded that evidence-based interventions like computerized cognitive training (CCT) can serve as a useful adjunct to medication in managing ADHD symptoms, especially when medication side effects are a concern [12]. These authors emphasize that while training is not a cure-all, it consistently produces improvements in certain aspects of working memory and executive function in ADHD, which could translate to better day-to-day functioning [12].

Likewise, in older adults with MCI – a precursor to dementia – cognitive training and brain stimulation have been explored as ways to bolster memory function. A recent meta-analysis of nearly three dozen randomized controlled trials involving about 1,489 participants with MCI found that computerized cognitive training yielded significant improvements in multiple memory domains, including working memory, with moderate effect sizes [13]. Notably, supervised training (where a coach or therapist guided the sessions) had larger benefits than unsupervised home training, suggesting that adherence and strategy coaching might enhance the efficacy of training [13]. Even so, the fact that unsupervised at-home training produced some improvement in verbal memory demonstrates the potential for accessible interventions [13]. In addition to cognitive exercises, non-invasive brain stimulation techniques (like transcranial direct current stimulation or repetitive TMS) are being tested; combining rTMS with cognitive training has shown additional improvements in small-scale studies [8]. Such approaches aim to directly enhance the neural plasticity underlying working memory, thereby slowing cognitive decline.

Apart from direct cognitive training, a growing number of studies show that a healthy lifestyle benefits working memory function. Regular physical exercise, meditation and mindfulness practices, and engagement in cognitively and socially stimulating activities have all been linked to better maintenance of working memory, especially in aging populations. For example, aerobic exercise programs have been found to increase prefrontal cortex activity and improve connectivity within frontoparietal networks, correlating with improvements in working memory tasks. Mindfulness meditation training has been associated with enhanced attention and working memory, potentially by reducing stress and improving attentional control. These interventions not only help healthy individuals stave off age-related decline in working memory, but might also slow down progression in those already experiencing cognitive impairment. Indeed, longitudinal studies indicate that older

adults who maintain higher levels of physical and mental activity tend to show less decline in working memory over time. Therefore, lifestyle modifications are recommended as part of comprehensive brain health maintenance strategies.

In clinical practice, working memory measures are increasingly incorporated into neuropsychological testing for disorders like AD, Parkinson's disease, ADHD, and schizophrenia. They serve as important cognitive endpoints in treatment trials as well. For instance, if a new drug is intended to improve cognitive function in schizophrenia, researchers will measure changes in working memory performance to gauge efficacy. In neurorehabilitation, strategies to improve patients' working memory (either via training or compensatory strategies like using memory aids) can significantly impact day-to-day independence – for example, enabling a stroke patient with frontal damage to better remember and carry out multi-step tasks.

In summary, working memory is a critical cognitive domain that is compromised in many brain disorders. By characterizing these impairments and developing interventions to address them, working memory research provides valuable tools for early diagnosis, cognitive remediation, and improving quality of life in clinical populations. For ADHD and MCI in particular, recent studies give hope that targeted cognitive training can yield tangible benefits in working memory performance [12,13]. These findings bridge the gap between basic research and practical applications, illustrating how understanding the mechanisms of working memory can guide therapies for cognitive disorders. Beyond training, non-invasive rhythmic stimulation (e.g., 40-Hz visual flicker) is being explored for disease modification in AD models [14].

5. SUMMARY AND FUTURE PROSPECTS

Working memory refers to the ability to retain and manipulate information over short periods, and it involves the coordinated activity of multiple brain regions (especially prefrontal and parietal areas). Research has shown that working memory is essential for a wide range of everyday cognitive tasks and is central to learning, decision making, and problem solving. Several influential theoretical models have been developed to characterize working memory. Baddeley and Hitch's multi-component model emphasizes separate subsystems for storing verbal and visuospatial information, coordinated by a central executive. Cowan's embedded-processes model, in contrast, describes working memory as the activated subset of long-term memory, with a tight focus on a limited attentional spotlight of about 3–5 items. Despite differences, these models agree on core phenomena such as capacity limitations and the crucial role of attention.

A variety of research methods have advanced our understanding of working memory. Behavioral experiments (e.g., dual-task paradigms, span tasks) provided initial evidence for the structure and limits of working memory. Neuroscientific techniques (EEG, fMRI, TMS) have mapped the neural substrates and dynamics of working memory, revealing how oscillatory brain activity and specific neural circuits support the short-term retention of information. Computational and animal models have further clarified the mechanisms, suggesting that working memory may be maintained by persistent neural firing as well as short-term synaptic changes ("activity-silent" states) . Key factors that affect working memory performance include inherent capacity limits, interference from distracting information, and influences of mental states such as stress or fatigue. Indeed, a wide range of variables – at least 21 factors by one review – can cause working memory performance to fluctuate, which helps explain why an individual's working memory can vary considerably across contexts and conditions.

Neurological and psychiatric disorders often involve working memory deficits. For example, Alzheimer's disease and related dementias feature early and severe impairments in working memory, reflecting damage to neural circuits needed for active maintenance. ADHD is characterized by limited working memory capacity and poor focus, which contribute to difficulties in academic and daily tasks.

These links make working memory a valuable target for clinical assessment and intervention. Cognitive training, pharmacological treatments, and lifestyle interventions (exercise, meditation) have all been explored for their potential to improve working memory or mitigate its decline. Meta-analytic evidence indicates that computerized cognitive training can produce moderate improvements in working memory in populations with mild cognitive impairment, and structured working memory training can be a useful adjunct therapy for ADHD. Furthermore, emerging techniques like rhythmic brain stimulation (e.g., gamma frequency light flicker or tACS) show promise in boosting working memory function or reducing pathology.

As we look to the future, the continued development of cutting-edge technologies such as artificial intelligence, brain–computer interfaces (BCI), and gene editing may revolutionize working memory research and its applications. For instance, AI-based analysis of large neuroimaging datasets could uncover subtle patterns in brain activity that predict working memory capacity or decline, enabling earlier interventions. BCI technology might one day allow real-time augmentation of working memory, perhaps by offloading some memory demands to an external device or by providing neurofeedback that helps users optimize their cognitive strategies. Advances in genetics might identify key genes that influence working memory, opening the door to gene therapies for congenital memory deficits. High-resolution and multimodal imaging (such as combined EEG-fMRI or next-generation optical imaging) will provide even more detailed views of the neural mechanisms, potentially capturing the interaction of persistent activity and activity-silent storage that current models propose. These technological breakthroughs, combined with interdisciplinary research efforts, will deepen our understanding of working memory and help translate that knowledge into practical solutions to enhance human cognition and combat memory-related problems. In short, working memory research sits at the crossroads of basic cognitive science and real-world applications, and its continued development will guide new directions in brain science, artificial intelligence integration, and human health.

REFERENCES

[1] Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. Language, Speech, and Hearing Services in Schools, 49(3), 340–355. https://doi.org/10.1044/2018_LSHSS-17-0114

[2] Carruthers, P. (2014). Evolution of Working Memory. In In the Light of Evolution: Volume VII: The Human Mental Machinery. National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK231620/

[3] Saline, S. (2022). Working Memory vs. Short-Term Memory: Differences, Similarities with ADHD. ADDitude. https://www.additudemag.com/adhd-working-memory-vs-short-term/

[4] Hitch, G. J., Allen, R. J., & Baddeley, A. D. (2025). The multicomponent model of working memory fifty years on. Quarterly Journal of Experimental Psychology, 78(2), 222–239. https://doi.org/10.1177/17470218241290909

[5] McLeod, S. (2023, Nov 20). Working Memory Model (Baddeley & Hitch). Simply Psychology. https://www.simplypsychology.org/working-memory.html

[6] Li, S., Voznak, J., & Xu, L. (2023). Alterations of neural activity in the prefrontal cortex associated with deficits in working memory performance. Frontiers in Behavioral Neuroscience, 17, 1213435. https://doi.org/10.3389/fnbeh.2023.1213435

[7] Aktürk, T., de Graaf, T. A., Güntekin, B., et al. (2022). Enhancing memory capacity by experimentally slowing theta frequency oscillations using combined EEG-tACS. Scientific Reports, 12, 14199. https://doi.org/10.1038/s41598-022-18665-z

[8] Balderston, N. L., et al. (2024). Neuromodulatory transcranial magnetic stimulation changes functional connectivity proportional to the electric-field induced by the TMS pulse. Clinical Neurophysiology, 165, 16–25. https://doi.org/10.1016/j.clinph.2024.06.007

[9] Funahashi, S. (2015). Functions of delay-period activity in the prefrontal cortex and mnemonic scotomas revisited. Frontiers in Systems Neuroscience, 9, 2. https://doi.org/10.3389/fnsys.2015.00002

[10] Sanderson, D. J., Good, M. A., Skelton, K., Sprengel, R., Seeburg, P. H., Rawlins, J. N. P., & Bannerman, D. M. (2010). Spatial working memory deficits in GluA1 AMPA receptor subunit knockout mice reflect impaired short-

term habituation: Evidence for Wagner's dual-process memory model. Neuropsychologia, 48(8), 2303–2315. https://doi.org/10.1016/j.neuropsychologia.2010.03.018

[11] Barbosa, J., Stein, H., Martinez, R. L., Galan-Gadea, A., Li, S., Dalmau, J., et al. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature Neuroscience, 23(8), 1016–1024. https://doi.org/10.1038/s41593-020-0644-4

[12] Al-Saad, M. S. H., Al-Jabri, B., & Almarzouki, A. F. (2021). A review of working memory training in the management of attention-deficit/hyperactivity disorder. Frontiers in Behavioral Neuroscience, 15, 686873. https://doi.org/10.3389/fnbeh.2021.686873

[13] Chan, A. T. C., Ip, R. T. F., Tran, J. Y. S., Chan, J. Y. C., & Tsoi, K. K. F. (2024). Computerized cognitive training for memory functions in mild cognitive impairment or dementia: A systematic review and meta-analysis. npj Digital Medicine, 7, 1. https://doi.org/10.1038/s41746-023-00987-5

[14] Singer, A. C., Martorell, A. J., Douglas, J. M., et al. (2018). Noninvasive 40-Hz light flicker to recruit microglia and reduce amyloid-β load. Nature Protocols, 13, 1850–1868. https://doi.org/10.1038/s41596-018-0021-x