2

Theoretical Models of Working Memory

Working memory can be defined as a cognitive function that is responsible for storing, holding, manipulating, and retrieving novel information. All theorists agree on this definition, despite disagreement regarding such aspects as limits, capacity, structure, and function. Currently, the most accepted model for explaining working memory is Alan Baddeley's model (Baddeley & Hitch, 1974; Baddeley, 2000; Baddeley, Allen, & Hitch, 2011). It is a complex model that can be tested under several conditions and using different experimental paradigms, which is in contrast to other models, such as Cowan's (1999, 2010). However, it does not explain certain phenomena, such as the enhancement of working memory by familiarity with stimuli (Cowan, 2010), the verbal encoding of olfactory, visual, and tactile stimuli (Jönsson, Moller, & Olsson, 2011), and the influence of mood in working memory tasks (Chan, Shum, Toulopoulou, & Chen, 2008).

Other theoretical models have been proposed to explain empirical data. In 1999, Akira Miyake and Priti Shah co-edited a book that gathered the main researchers in the working memory field at the time, including Alan Baddeley, Nelson Cowan, Randall Engle, Stephen Tuholski, Michael Kane, and Richard Lewis. Among these authors, Cowan's model is the second most well-known in the literature and the first option for explaining the effects of familiarity and attention (Cowan, 2010). However, Cowan’s model (1999, 2010) lacks precision in explaining different types of encoding and strong empirical evidence of individual and group differences in phonological, olfactory, and visuospatial working memory.

Working memory is thought to be executed, like other executive functions, in the prefrontal cortex. Engle, Kane, and Tuholski (1999) provided evidence of how working memory can be explained as an integrative part of fluid intelligence. Their evidence does not necessarily exclude either Cowan’s or Baddeley’s model, but it suggests that fluid intelligence performance and prefrontal cortex activation are associated with complex working memory tasks.

Working memory is also considered an executive function. Executive functions comprise a set of superior mental processes that are needed for concentration and attention when behaving automatically or relying on instinct or intuition would be ill-advised, insufficient, or impossible (Diamond, 2009a, b, 2013; Diamond, Lee, & Hayden, 2003). They include three low-order functions (inhibitory control, working memory, and cognitive flexibility) and three high-order functions (fluid intelligence, rational reasoning, and logical reasoning). Based on this perspective, working memory is limited to storing, holding, and retrieving novel information, whereas manipulating, controlling, updating, and inhibiting predisposed responses and self-regulation are part of executive functions but not responsible for working memory itself.

Baddeley’s model (Baddeley & Hitch, 1974; Baddeley, 2000; Baddeley, Allen, & Hitch, 2011) is used when researchers seek to study only the effects of a modality (e.g., phonological, visual, or any other sensorial input) or when they try to separate working memory into its four hypothesized components: phonological loop, visuospatial sketchpad, episodic buffer, and central executive. When researchers try to explain the effects of familiarity and working memory capacity, they tend to cite Cowan’s work. Diamond’s model of executive function is also widely used to explain the role of the prefrontal cortex in tasks that require novel solutions, self-regulation, decision making, and the inhibition of predisposed responses.

These different theoretical models represent different views of the same phenomenon. This means that we have the opportunity to study them in-depth and test them using empirical evidence. The objective of this chapter is to review behavioral evidence and further understand the crucial differences between these models so we can test their hypotheses using functional magnetic resonance (fMRI) data.

2.1

Baddeley’s working memory model

Information processing theory is one of the most frequently used psychological hypotheses to explain behavior that arises from psychological processes. It was first proposed during the cognitive revolution in the 1950s by important names in the history of psychology, such as Donald Broadbent, George Miller, and Noam Chomsky (Mills, 2000). The cognitive revolution emerged as a counterpart to the behaviorism movement that was concerned with only the product of the process, without caring about how behavior is generated in the mind.

One of the main landmarks of the cognitive revolution was the celebrated work of George Miller (1956) entitled, “The Magical Number Seven, Plus or Minus Two.” Miller proposed that one of the main mechanisms of human cognition, a memory subtype that was initially called short-term memory, was limited by the number of stimuli that could be retained at the same time (seven plus or minus two, from five to nine items). The experimental paradigm that was used and successfully replicated in numerous publications since then (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002) was the digit span task. The participant listens to a random sequence of numbers and is asked to repeat them orally. Miller showed that other cognitive processes were associated with this limit in the capacity of short-term memory.

With advances in psychological sciences, Baddeley and Hitch (1974) developed the first theoretical model that allowed an explanation of various empirical data that were generated by various digit span methods that were developed based on Miller's work in 1956 (Figure 1). The theoretical foundation on which the model of Baddeley and Hitch was created was information processing theory. This theory posits that the human mind works like a computer that processes stimuli as inputs and generates outputs. At this point, no one thought in terms of a behavioral product but rather in terms of the process that generates it. These authors suggested that not only the span of digits was limited; span limitations also exist for other types of information, such as words, colors, and shapes and the ability to recall them in reverse order of presentation, a task known as reverse or backward span. They also realized that the stimulus modality also mattered. Some people could perform better when the stimuli were auditory and worse when the stimuli were visual, and *vice versa*. Finally, they found that this entire process of retaining and manipulating information in the mind demanded a sort of general cognition that manages the underlying processes, such as an executive in a company, and its overall processing was intrinsically linked to the participant’s limit of attention. Baddeley and Hitch (1974)suggested that these processes reflected Miller’s working memory model, which can be defined as the ability to retain and manipulate new information and provide the most appropriate response that is dictated by the environment.

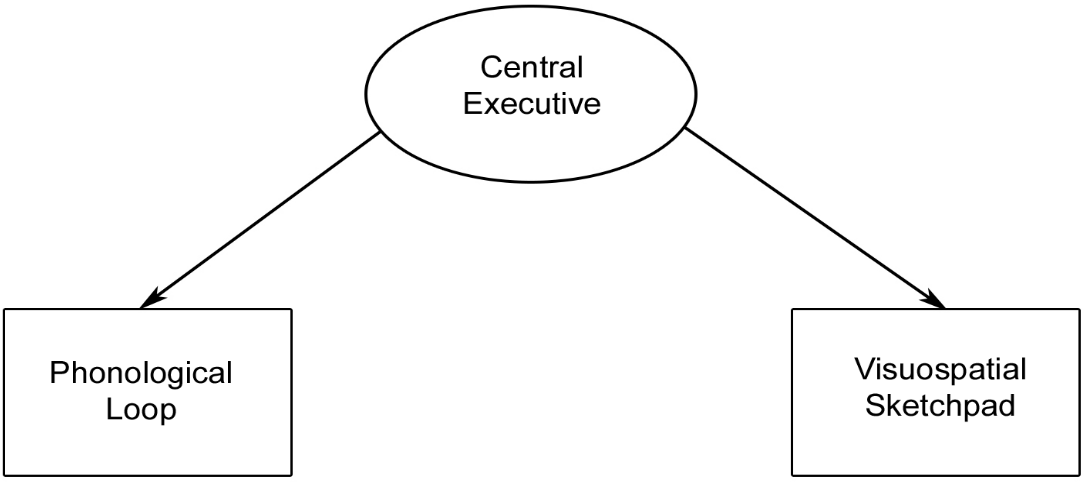

Fig 1. First working memory model adapted from the proposal of Baddeley and Hitch (1974).

The components of the theoretical model of Baddeley and Hitch are the phonological loop (responsible for retaining auditory and verbal information, such as words, letters, or sounds, while the central executive handles them), visuospatial sketchpad (which performs the same function as the phonological loop but with visual stimuli), and central executive (which serves to guide executive attention to the most relevant part of information at a time and manages the capacity of working memory according to task demands; Baddeley and Hitch, 1974).

When a person must remember a phone number, pick up the phone, and dial the number while repeating the numbers sequentially, he is relying on his own working memory. This also happens in other day-to-day activities, such as preparing a new recipe that was seen on television or trying to mimic a yoga teacher’s movements. These are all working memory tasks.

Since the model of Baddeley and Hitch in 1974, new hypotheses have emerged to explain the underlying processes and individual differences in working memory tasks. However, the most consistent model with much evidence to support it is Baddeley’s new model, revised in 2000. The independence of the modalities in domain-specific systems and presence of a general domain system comprise a more robust theoretical model to explain the processing of information using working memory (Baddeley, 2012). Baddeley included in the original model from 1974 a new component called the episodic buffer. The episodic buffer is a system that is responsible for integrating information from different modalities and sources into one, so an underlying component of the central executive serves as an interface between domain-specific systems and long-term memory to generate knowledge (Figure 2).

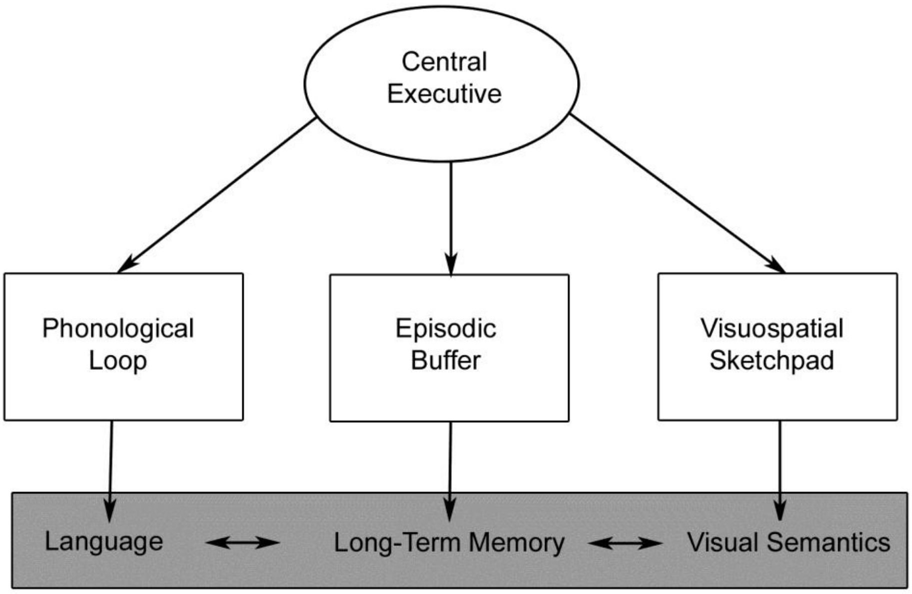

Fig 2. Second working memory model adapted from Baddeley (2000). Processes associated with fluid intelligence are in the white rectangles. The gray rectangle encompasses the processes associated with crystallized intelligence.

Since the new model was proposed, much evidence has emerged to support the hypothesis of an independent system of working memory and the importance of the episodic buffer as an integrative component. However, other empirical data indicated other subcomponents within slave-specific-domain working memory systems rather than just the phonological and visuospatial domains. After a series of experiments, Baddeley, Allen, and Hitch (2011) developed the latest version of the model, which includes the previously missing sensory modalities and importance of the episodic buffer in integrating modalities. In this latest model, the central executive is a general domain component that coordinates the episodic buffer only, so there are no connections between the central executive and other subsystems as previously thought. The episodic buffer integrates information and coordinates directly with the slave subsystems to execute whatever the central executive commands.

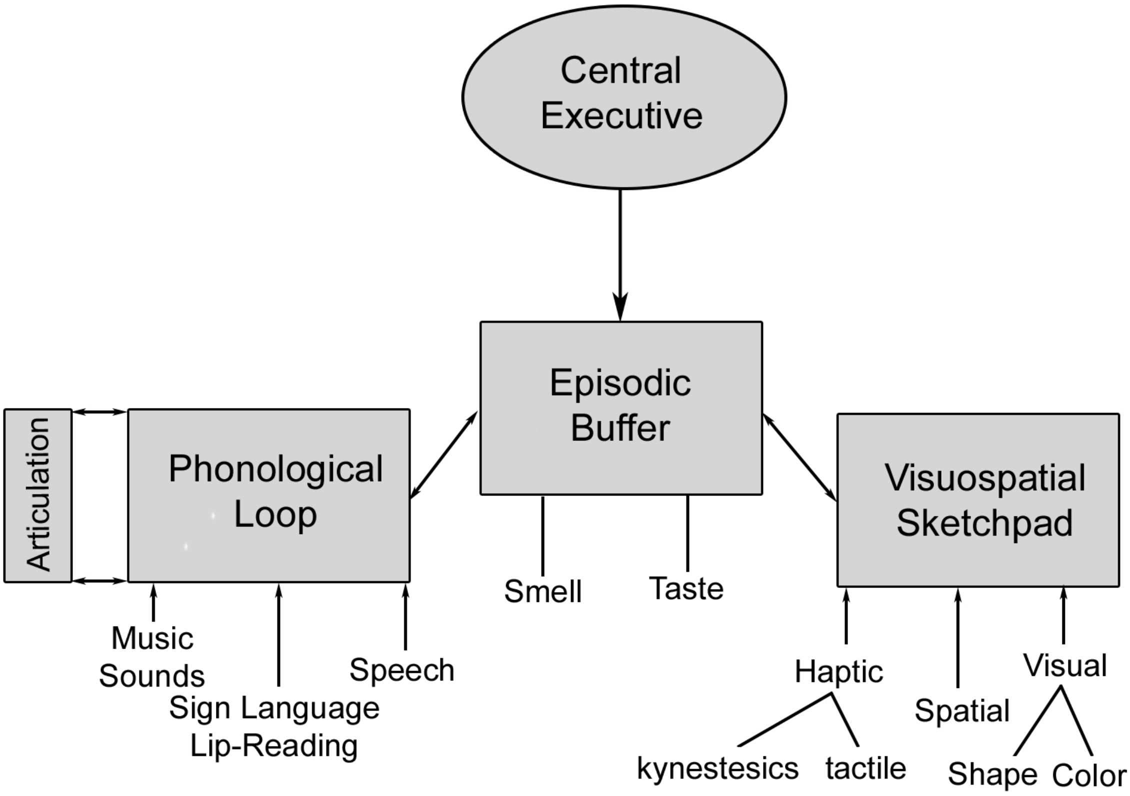

Fig 3. Latest working memory model adapted from Baddeley, Allen, and Hitch (2011). The episodic buffer became the only subcomponent that is directly controlled by central executive resources.

Recently, Allen, Baddeley, and Hitch (2014) suggested the presence of a double attentional component within the central executive to explain individual differences in visuospatial task performance. In their experiment, participants differed in the first three items from a serial working memory task in the presence of matched information in the same stimulus, such as color and shape at the same time. When the participant had to record that the square is always red, he could retain that duplicated information in the first three stimuli (e.g., red square, yellow rectangle, and green circle). The performance of the participants changed after the fourth stimulus, demonstrating the possible involvement of an attentional system for the first items but recruitment of a new system that is responsible for the latest information after the fourth item. This means that the central executive provides resources for the episodic buffer to work differently. The initial three items in short-term memory are privileged, but for the items that are subsequently presented, executive attention attempts to maintain their representation only when recall is needed. This new discovery by Allen, Baddeley, and Hitch (2014) allowed the development of new hypotheses of the functioning of the central executive.

2.1.1

Phonological working memory according to Baddeley’s model

The part of Baddeley’s working memory model that accounts for sounds, voices, language, and any kind of auditory input is phonological working memory. It is empirically defined as the integration between the central executive, episodic buffer, and phonological loop. Earlier in this chapter, we discussed the roles of the central executive: a domain-general system that manages the amount of attentional resources and coordinates the demand for the integration of modalities and the episodic buffer (a domain-specific slave subsystem that is responsible for integrating different modalities into one manageable piece of information, such as sound + speech, shape + color, and sound + color).

However, the phonological loop is another important slave subsystem. Understanding how it works can shed light on phonological working memory from Baddeley's point-of-view. The distinction between short-term memory and working memory is crucial. Baddeley (2003a) clarified that short-term memory is a system that is involved only in storing information, whereas working memory manipulates and retrieves stored information through mental processing. This means that the slave subsystems (phonological loop and visuospatial sketchpad) are in fact domain-specific short-term memories that are specialized with regard to their respective sensorial modality or input.

The structure and cognitive strategies that are used to store and maintain auditory and language information that is to be used in working memory tasks were explained by Baddeley (2003a). The first pivotal point regarding the phonological loop is that it is divided into two activities: temporary storage and rehearsal. It involves a subvocal rehearsal system that not only maintains information within the store but also records visual information within the store, provided a visual item can be named. What appears to happen is that sound similarity impairs immediate recall, likely because of sound discrimination. Although subjects can readily recall a sequence of letters (e.g., B,W,Y,K,R,X), they are likely to have considerable difficulty retaining sequences of letters with similarly sounding names (e.g., T,C,V,D,B,G; Conrad & Hull, 1964). A similar phenomenon occurs when words are used. A word sequence such as *man, cat, map,* and *cab* can be correctly recalled less than 20% of the time, whereas subjects will score above 80% with a dissimilar sequence, such as *pit*, *day*, *cow*, *sup*, *pen* (Baddeley, 1966a). The fact that this is a characteristic of short-term memory rather than long-term memory systems was demonstrated in a further study in which subjects were presented with lists of 10 words from each set and required to learn the sequence across a series of trials. Under these conditions, the similarity of meaning was important, and phonological similarity lost its effect (Baddeley, 1966b). This evidence indicates that familiarity does not help phonological working memory in simple span tasks.

Evidence of a rehearsal system is provided by the word-length effect, which involves presenting subjects with a sequence of items and requiring immediate serial recall. The memory of a five-word sequence drops from 90% when the sequence consists of monosyllables to ~50% when five-syllable words are used, such as *university, opportunity, international, constitutional, auditorium* (Baddeley, Gathercole, & Papagno, 1998). The word-length effect can be abolished by simply requiring the subject to utter a sequence of irrelevant sounds, such as repeating the word *the*. It impairs performance because it both blocks the maintenance of the memory trace through rehearsal and prevents the subject from using subvocalization to record the items in the phonological store when visual presentation is used. The episodic buffer appears to play an important role in trying to concentrate attentional effort in one modality of information rather than integrating the whole set of stimuli. Much evidence has shown that verbal encoding actually improves phonological working memory performance (Cowan, 2010; Jönsson, Moller, & Olsson, 2011); thus, Baddeley's (2003a) assumption of impairment has been faced with contradictory empirical evidence.

Some of this effect undoubtedly occurs because long words take longer to recall, leading to more forgetting (Cowan, 1999). However, the fact that a word-length effect occurs when the output delay is held constant, either by using a probe procedure or by recognition (Baddeley, 2003a), indicates that the effect operates at both the ongoing rehearsal level and through forgetting during responding.

Another important point regarding phonological short-term memory is that rehearsal relies on overall speech-motor programming and not articulation. The process of subvocal rehearsal does not appear to depend on the capacity for overt articulation. Baddeley (1966b)showed that dysarthric patients who lost the ability to articulate can show clear evidence of subvocal rehearsal, reflected by the word-length effect or an effect of acoustic similarity with visually presented items. In contrast, dyspraxic patients whose problems stem from a loss of the ability to assemble speech-motor control programs show no sign of rehearsal. This implies that the capacity to set up speech-motor programs underpins rehearsal rather than overt articulation.

Evidence supports the notion that the phonological loop is influenced by conceptual knowledge. This probably means that the working memory system is not dissociated from long-term memory. Mutual influences likely exist, depending on the task. Baddeley, Papagno, and Vallar (1988) tested the ability of one patient, who had a very pure phonological short-term memory deficit, to acquire the vocabulary of an unfamiliar foreign language: Russian. The experiment required her to learn eight items from the Russian vocabulary (e.g., *svieti*[rose]), and comparisons were made with her ability to learn to associate pairs of unrelated words in her native language (e.g., horse-castle). They found that such native language pairs were learned as rapidly by the patient as by normal control subjects, whereas she failed to learn any of the eight Russian items (Baddeley, Papagno, & Vallar, 1988). The phonological loop appears to be a useful aid in learning new words. In another study, they found that requiring subjects to suppress rehearsal by uttering an irrelevant sound disrupted foreign but not native language learning and that phonological similarity among the items to be learned also disrupted the acquisition of novel vocabulary, as did increasing the length of the novel items (Papagno & Vallar, 1992). Both of these variables impaired phonological loop performance.

Two alternative views can also explain the role of phonological short-term memory (clearly synonymous to the phonological loop) and language. Other authors suggested that phonological storage itself is merely a reflection of deeper phonological processing problems. This model by Brown and Hulme (1996) differed from our own by emphasizing the role of existing language habits in facilitating vocabulary learning. Gathercole and Baddeley (1993) found that sequences that were closer to English (e.g., stirple, blonterstaping) were indeed consistently easier than less familiar phoneme sequences (e.g., kipser, perplisteronk). This strongly suggests the influence of existing language habits on current nonword repetition performance, exactly as the Brown and Hulme (1996) model would predict. One way of explaining this pattern of results is by considering the division of the phonological loop into separate storage and articulatory components. The nonword repetition task might demand both of these, whereas only the articulatory output system might depend on earlier language habits, leaving the phonological store relatively language-independent. Baddeley (2003a) suggested that existing language habits have a major effect on performance in tasks that resemble the acquisition of vocabulary through their impact on output and rehearsal, rather than by directly influencing phonological storage.

The other alternative explanation is that language acquisition relies on general phonological processing and not on the phonological loop. Furthermore, whereas both the nonword repetition and phonological awareness models are capable of predicting reading performance, they appear to account for separable variance (Baddeley et al., 1998). Therefore, it can be argued that the greater specificity of the phonological loop hypothesis has a clear advantage over a general phonological processing interpretation.

In the case of short-term memory patients, their language deficits appear to be limited to the major disruption of short-term phonological storage while other phonological and linguistic skills appear to be preserved (Vallar & Shallice, 1990).

Phonological short-term memory clearly plays a pivotal role in language acquisition, regardless of other deeper or higher processes. Baddeley (2003a) suggested a neurobiological basis of the phonological loop that can be tested using fMRI meta-analysis, in which the temporary storage system is centered in Brodmann area 44 (predominantly in the left hemisphere), and the rehearsal system is centered in Brodmann area 40 (Broca's area, predominantly in the left hemisphere). These are the proposed structures for the phonological loop. Auditory information is analyzed and fed into a short-term store. Information from this system can pass into a phonological output system and result in spoken output or rehearsal. This, in turn, may recycle information, both subvocally into the short-term store and into the ears when rehearsal is overt. Visually presented material may be transferred from an orthographic code to a phonological code and thereby recorded within the phonological output buffer (Vallar & Papagno, 2002).

To test Baddeley's model (Baddeley & Hitch, 1974; Baddeley, 2000; Baddeley, Allen, & Hitch, 2011), activation in the following areas would be expected (Baddeley, 2003b): phonological loop (Brodmann areas 6, 40, and 44, predominantly in the left hemisphere), episodic buffer (Brodmann area 7), and central executive (Brodmannareas 9 and 46).

2.2

Cowan’s embedded-processes model

Working memory was defined by Nelson Cowan (1999, 2010) as a cognitive process that retains old and novel information in an accessible state that is suitable for manipulating and carrying out tasks with mental components. Nonetheless, working memory does not exist as a separate entity; it constitutes merely a practical and task-oriented label so researchers can discuss it. His hypothesis was that working memory is, in fact, a set of embedded-processes from both attention and long-term memory. It also means that if an entire process is invoked without facilitating a task, then it is still considered working memory (e.g., the verbal encoding of meaningless shapes). Cowan argued that his model does not deny the definition of processes that are found in other models, but he attempts to explain a single way of functioning, regardless of the type of stimulus or input.

The stimulus is stored for a brief moment (hundreds of milliseconds) in a sensory store to be further driven to either an activated portion of long-term memory or the focus of attention. An unchanged stimulus tends to go to the activated long-term memory, whereas a novel stimulus and voluntarily attended stimulus stay within the focus of attention. The activated portion of long-term memory is also known as short-term storage or short-term memory, which keeps the information that is needed to complete a task activated. The focus of attention is the enhancement of processing of one piece of information to the detriment of another. Finally, the process that is responsible for gathering those mental processes together in a way that can follow or be modified by instructions and incentives is called the central executive.

Four processes are used during working memory tasks: encoding, representation, maintenance, and retrieval. The processing of information is based on this set of mechanisms and relies on both long-term memory and attention to ease further processing. Individual differences in working memory tasks can be explained by limitations in both attention and long-term memory. The activated portion of long-term memory appears to present a time decay effect, whereas attention is limited by the amount of information that can be held in the focus of attention at a given time. If Cowan's hypothesis is correct, then there should be evidence that activated long-term memory (or short-term memory) gradually diminishes over time in tasks with a delayed response. Attention should be limited to a critical number of items or chunks in complex span tasks. He cites several experiments that showed that activated long-term memory indeed decays over time (10-20 s) in delayed-to-sample tasks when distractors are presented during maintenance. In different sets of experiments that use a stimulus that cannot be chunked or rehearsed, participants tend to show performance of 4±1 items in complex span tasks. Altogether, this evidence suggests that Cowan's model is indeed one of the closest ones that can explain working memory. Perhaps this indeed reflects an overlap of long-term memory and attention rather than a singular cognitive entity.

The subset of memory that is represented in long-term storage must be activated to be accessible to the focus of attention (e.g., in a number span, the long-term memory that is associated with all known numbers is activated and remains this way throughout processing). Only activated information may enter into awareness, but the opposite is not true, in which it is possible to access information from outside conscious awareness (e.g., when you are doing a number span task and someone calls your name).Cowan's model emphasizes the relationship between memory and attention. There are different processing limits for each cognitive domain: memory and attention. The focus of attention is controlled by two systems (voluntary and involuntary), and conscious awareness can be influenced during processing (Cowan, 1999, 2010).

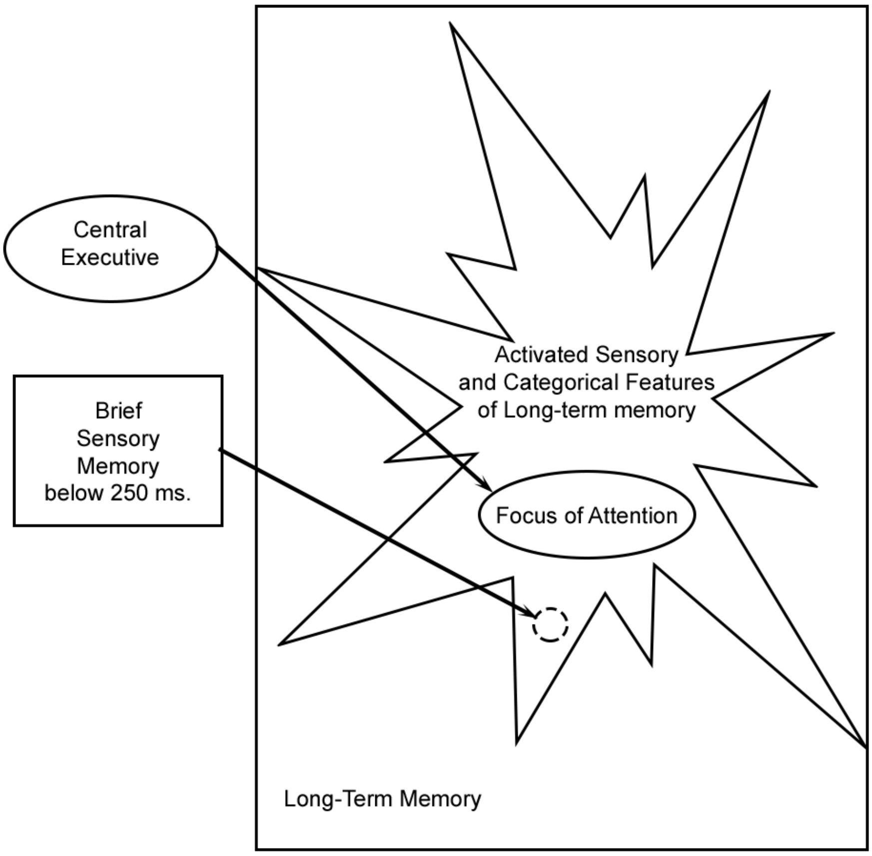

Fig 4. Cowan's working memory model adapted from Cowan (2010). The central executive plays the role of a supervisory attention system with two components: one that is automatic and another that is voluntary. Long-term memory is constantly activated and remains this way while it is needed. Finally, one small portion of the activated memory is actually a brief sensory memory that lasts <250 ms and serves only to orient the focus of attention to particularly dangerous or predisposed stimuli.

Attention and awareness are coextensive. Attention is the enhancement of the processing of some information and exclusion of other currently available information, and awareness is the ability to be consciously aware of information. Involuntary attention is the automatic recruitment of attention (e.g., a fire alarm, the sound of a car horn, or someone calling your name) to detect physical changes in the environment or changes in an habituated stimulus. Voluntary attention is an effort-demanding process (e.g., searching for a stimulus within a set of items or saying a word list backward) that is controlled by the central executive. The central executive is "the collection of mental processes that can be modified by instructions and incentives" (Cowan, 1999, p.65). Cowan (2010) stressed that his model is not intended as a description of processing but rather as a simple summary and organization of pivotal features of attention and memory as embedded processes; without the coordination of both, processing is not possible.

The long-term memory portion of working memory is then described as a set of features in long-term memory that are used to encode the stimulus to make it more familiar, thus enhancing further memory representation. Encoding can be abstract or sensorial. Abstract codes include phonological codes (ba, bo, da, etc.), semantic codes (the meaning of a word or sign), spatial orientation codes (left, right, up, down, etc.), and so on. Sensory codes are the modality of the input, including visual codes (shape, color, size, luminosity, etc.), hearing codes (tones), tactile codes (textures), olfactory codes (smells), and gustatory codes (taste). Cowan (1999) argued that executive control circulates information that is currently within the focus of attention using rehearsal, but it is possible to use long-term memory if relevant information is available to deal with the task, such as using chunking as a strategy. The focus of attention is important to enhance encoding. In attention-shifting tasks (e.g., reading a text and responding to the sound of a specific syllable), when participants pay attention to one thing at a time (i.e., they stop reading to pay attention to a sound), they tend to make fewer mistakes than when they are immersed in reading. This suggests that phonological encoding demands the focus of attention at least at categorical levels. Cowan (2010) suggested that semantic encoding is limited if there is not an important part of attention and conscious awareness involved.

Working memory is also a valuable tool to represent a set of stimuli in long-term memory. According to Cowan (1999), representation appears to depend on the form in which an item is represented. Phonological short-term memory is influenced by auditory tasks but not by visuospatial tasks. Cowan argued that Baddeley’s model neglects other types of representation, thus limiting his model to only phonological and visuospatial short-term memory. According to Cowan (2010), other modalities of representation appear to differ from auditory and visuospatial stimuli, such as tactile stimuli and nonverbal sounds. The properties of representation may vary accordingly to encoding properties. The verbal encoding of visual items is more suitable than serial recall if the items’ names are known, whereas visuospatial encoding is more suitable when items are organized according to direction or position. This phenomenon can be clearly viewed in studies that used the olfactory modality with high and low demands of verbal encoding (Jönsson, Moller, & Olsson, 2011).

The maintenance of information in the focus of attention is the most important feature of the embedded-processes theory. Maintaining a set of items in activated memory requires strategies to keep the stimulus circulating in the focus of attention. Rehearsing is the most common strategy, but other strategies may apply, such as recirculating items in a search task. If a similar persistence of information is spread among all items, then individual differences between children could be explained by the rate of pronunciation rather than interword pauses. However, as lists of words increase in a word span task, silent periods between words also increase. Cowan (1999) suggested that once a child retrieves a particular item from activated memory, the focus of attention changes quickly to the next item. Thus, it is not only maintaining active information in short-term storage but also circulating this information in the focus of attention.

Finally, retrieving information accurately is pivotal in working memory tasks. This is defined as entering the correct items into the focus of attention. Retrieval from long-term memory is limited only by practical reasons, but retrieval from activated memory (i.e., short-term memory + attentional control) needs to be fast because information will disappear through memory decay. If sufficient episodic memory is represented and stored in long-term memory, then it is possible to retrieve items even after deactivation—loss of the novel information.

Cowan (2010) based his model on Anne Treisman's attenuation-filter theory of attention (1996). He extends Treisman's theory by adding the concepts of attended and unattended information. Thus, information activates certain portions of memory when the stimulus is relevant. An irrelevant stimulus does not fade away; it remains unattended but available in memory for the person to use if needed or demanded. Evidence suggests that unattended information is still able to be retrieved automatically by working memory if enough effort is given to orienting attention. Less effort is needed when physical changes occur in the stimulus, whereas more effort is needed when complex and semantic changes occur in unattended stimuli (Cowan, 2010). Both Cowan (1999) and Baddeley (1999) agreed about a passive storage component (activated memory/short-term memory) and an active processing component of working memory, but only Cowan took into account automatic activation during working memory tasks.

2.2.1

Working memory capacity

Cowan (1999, 2010) dedicated an important part of his work to explaining the capacity of working memory. Individual differences in capacity can explain differences in higher-order cognitive domains. "It seems unlikely that, say, seven items could be held in attention at once. Therefore, in addition to attended information, one needs activated sources outside attention and/or supplementary help from the long-term memory" (Cowan, 1999, p.79). Cowan conceded that he does not know if there is a limited capacity of activated memory, but because it is a part of long-term memory, there are likely no limitations. The time limit of activated memory seems to range between 10 and 30 s. Several discriminatory tasks show the decay of activated memory after this period of time.

According to Cowan (2010), the capacity of the focus of attention is the "magical number" 4 ± 1. Capacity is the number of items in the focus of attention at a given time. Different types of stimuli may have different limits (e.g., visuospatial or phonological), but differences are likely attributable to more or less effort that is demanded in attention switching or dual tasks. The time limit of attention is associated with vigilance tasks. Evidence suggest that this limit is around 1 h.

The capacity of working memory leads to the distinctive roles of embedded processes in either working memory or individual performance. According to Cowan (1999), working memory is a global workplace where the information that is needed to perform a task is especially accessible temporarily. Several pieces of memory must be combined and thus are concurrently activated, whereas individual performance can be explained by the mechanisms by which information becomes accessible, which may vary. Thus, performance varies because of activated mechanisms and not the use of working memory. "Thus, there is no single, separate theoretical entity that I would call working memory; that is a practical, task-oriented label" (Cowan, 1999, p.79).

Cowan's (2010) theory suggests that information in long-term memory is activated to allow a person to perform a task. Sometimes, if this information is insufficient, then additional long-term memory is activated. Other previously unused regions of long-term memory can be recombined or co-activated during the same task to complete it. If this happens, then a novel combination of information can be formed within activated memory, leading this new combination to build a new piece of long-term memory. This model of working memory includes attention as a pivotal piece of the puzzle. The focus of attention holds information within consciousness and deals with changes in stimuli. However, activated memory can be outside the attentional range and thus unattended by conscious awareness.

Cowan (2010) finally suggested a neurobiological basis for his embedded-process theory. Cowan's first assumption for a biological foundation of working memory was neuronal activation when the physical characteristics of a stimulus change, thus leading the focus of attention to move from one piece of information to another. Several regions are associated with each feature of working memory. Cowan suggested the following biological underpinnings of the major aspects of working memory: (1) brief sensory system (sensorial cortex; for phonological information, the auditory cortex in the temporal lobe), (2) long-term memory activated portion (association cortex in the parietal lobe), (3) storage and focus of attention (locus coeruleus, hippocampus, and anterior cingulate cortex), (4) central executive (prefrontal cortex),and (5) attentional intervention and entry into the focus of attention (thalamus).

2.3

Prefrontal cortex role and executive function

Executive function or executive control refers to a group of top-down mental processes on which an individual relies when he needs to concentrate and pay attention because doing a task automatically or relying on instinct or intuition would not be advised or sufficient (Diamond, 2013). When someone must deal with and respond to novel information and make the appropriate (not automatic) response, this is considered an executive function task. Using executive control demands effort. It is easier to continue doing what someone has been doing than to change. It is easier to give into temptation than to resist it. It is easier to go on *automatic pilot* than to consider what to do next (Diamond, 2013).

There is general agreement that there are three core executive functions (Miyakeet al., 2000): (1) inhibition (also called inhibitory control) that includes self-control (behavioral inhibition) and interference control (selective attention and cognitive inhibition), (2) working memory, and (3) cognitive flexibility (also called set shifting, mental flexibility, or mental set shifting, closely linked to creativity).Based on these, higher-order executive functions are built, such as reasoning, problem solving, and planning.

Executive control is a set of skills that are essential for mental and physical health, success in school and in life, and cognitive, social, and psychological development. For example, impaired executive functions are found in addictions, attention-deficit/hyperactivity disorder, depression, obsessive compulsive disorder, schizophrenia, and bipolar disorder. The same thing occurs with child development and educational readiness and performance (Diamond, 2013).

Understanding each executive function from a working memory theorist point-of-view is important. First, inhibition involves being able to control one’s attention, behavior, thoughts, and emotions to override a strong internal predisposition or external lure and instead do what is more appropriate or needed in a given situation. Without inhibitory control, the brain would be at the mercy of impulses, old habitual thoughts or conditioned responses, and environmental stimuli that pull us in a given direction.

Inhibitory control allows us to change and choose how to react and behave rather than being unthinking creatures of habit. The classic tasks that are associated with inhibition include the Simon task, Flanker task, Go/No-Go task, stop-signal task, and Stroop task (Diamond, 2013).

Working memory, according to Diamond's (2013) model, refers to the ability to hold information in mind and mentally work with it (i.e., work with information that is no longer perceptually present). According to Diamond, only two types of working memory can be distinguished by encoding processes: verbal and nonverbal (visuospatial). Working memory is critical for making sense of anything that unfolds over time, which requires holding in mind what happened earlier and relating it to what comes later. Thus, to make sense of written or spoken language, one must determine whether it is a sentence, a paragraph, or something longer. Doing mathematics in your head, mentally reordering items, translating instructions into action plans, incorporating new information into thinking (updating), considering alternatives, mentally relating information to derive a general principle, and seeing relationships between items or ideas all require working memory. Reasoning would not be possible without working memory (Diamond, 2013).

Cognitive flexibility requires inhibition and working memory and comes much later in development (Garon, Bryson, & Smith, 2008). To change perspectives, we must inhibit our previous perspective and load a different perspective into working memory, one that can be already established in mind based on long-term memory or recently acquired based on short-term memory. Cognitive flexibility requires inhibitory control and working memory. One aspect of cognitive flexibility is being able to change perspectives spatially (e.g., looking at a dinner table from its longer side and then from its shorter side). Someone can also change perspectives interpersonally (e.g., assuming another person's point-of-view in an argument). Another aspect of cognitive flexibility involves changing how people think about something (i.e., "thinking outside the box"; Diamond, 2013). For example, if one way of solving a problem is not working, then could someone come up with a new idea that taps into the solution that had not been considered before? Cognitive flexibility also involves being sufficiently flexible to adjust to changing demands or priorities, admitting you were wrong, and taking advantage of sudden, unexpected opportunities. For example, when a student is not understanding a concept that the teacher explains, then those teachers often blame the student. But we could think differently and try to figure out a way to teach the content to the student in another fashion so that he can follow and finally grasp the concept (Diamond, 2013).

2.3.1

Working memory and inhibitory control

One of the most important aspects of the theoretical model of executive function is that it separates the control of the focus of attention (considered here as inhibition) from the rest of the working memory system. In Diamond's (2013) words:

> "They generally need one another and cooccur. One prototypical instance of when [executive functions] are needed is the class of situations where you are to act counter to your initial tendency on the basis of information held in mind. [Working memory] and inhibitory control support one another and rarely, if ever, is one needed but not the other" (Diamond, 2013, p. 143).

According to this view, whenever someone is executing a task, he must keep his goal in mind to know what is relevant or appropriate and what to inhibit. By concentrating especially hard on the information that one holds in mind, he increases the likelihood that the information will guide behavior and decrease the likelihood of an inhibitory error (i.e., giving the predisposed response rather than the correct one). This means that constantly holding, manipulating, and updating information in mind supports what someone should do and when he should inhibit a predisposed response to give the correct response. This leads to the conclusion that inhibition relies on working memory to be accurate.

Inhibitory control supports working memory. To relate multiple ideas or stimuli together, someone must be able to resist focusing exclusively on just one thing and recombine ideas and stimuli in new, creative ways. This means a person should be able to resist repeating old thought patterns and keep doing what is right rather than what used to be done. To keep the mind focused on something, one must inhibit internal and external distractions, thus voluntarily controlling the focus of attention. Many of us are familiar with suddenly realizing that we do not know what was in the passage we supposedly just read because our mind was elsewhere (i.e., a flight of thoughts or ideas; Diamond, 2013).

In fact, although inhibitory control and working memory appear to complement each other, some authors (e.g., Diamond, 2013; Wright & Diamond, 2014) believe they are in fact different domain-specific functions. However, other authors do not make a distinction between these two processes, rather considering them as one piece of the other (Baddeley, Allen, & Hitch, 2011; Cowan, 2010; Engle, Kane, & Tuholski, 1999). Diamond (2013) and Wright and Diamond (2014) suggested that three tasks provide evidence of this separation: Hearts and Dots task, spatial Stroop task, and complex span task. The Hearts and Dots task and spatial Stroop task require the person to hold only one rule in mind, meaning that there are low load or no load demands on working memory. Complex span tasks require almost no attentional control because there are no potential distractors that occur during the task (Diamond, 2013). The Hearts and Dots task and spatial Stroop task would be pure inhibition tasks, and the complex span task would be a pure working memory task.

Theorists of working memory models (Baddeley, 2000, 2003; Engle, Kane, & Tuholski, 1999) assert that inhibition is in fact a part of the supervisory attentional system (Norman & Shallice, 1986). According to this model, attentional control or executive control is divided into two subsystems. One system is responsible for processing environmental stimuli that involve perception, automatic attention, memory, and the updating of information. The other system controls and self-regulates actions in a way that keeps the mind constantly focused by inhibiting thoughts and ideas that are not related to the task at hand (Baddeley, 2000; Engle, Kane, & Tuholski, 1999; Norman & Shallice, 1986). According to Baddeley (2000), the supervisory attentional system is a proper model for the central executive and a domain-general set of mental processes that are responsible for maintaining the focus of attention in a task.

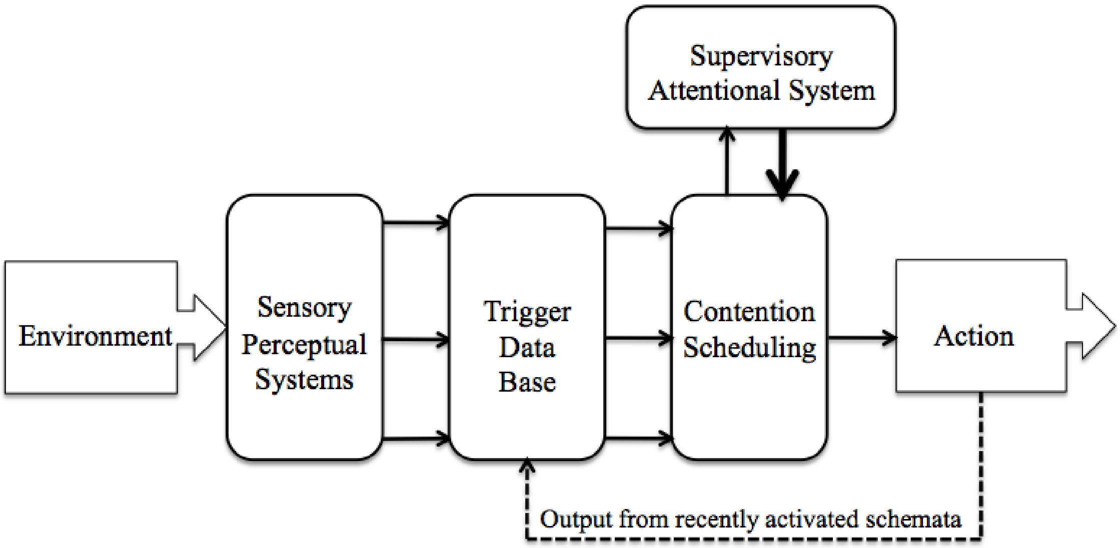

Fig 5. The supervisory attentional system according to and adapted from Norman and Shallice (1986). An environmental stimulus arrives in the mind through the sensory and perceptual systems, trigging long-term memory. To act or behave, consciousness brings from memory a set of thoughts or ideas and holds it in an organized and scheduled part of the consciousness. To behave accordingly, the mind should supervise the thoughts and ideas so they cannot escape from this organization. This is the work of the supervisory attentional system. Finally, the individual acts, and the outcome is judged, and memory is adjusting accordingly.

Baddeley (2000), Cowan (2010), and Engle, Kane, and Tuholski (1999), among others, agree that attentional or executive control is a part of the central executive. Something in dissonance with Diamond (2013) hypothesis of executive functions that suggests attentional control is an integrative part of the attentional system, but it does not influence on inhibitory control or working memory.

2.4

Working memory function and brain activation hypotheses

Based on the three theoretical models presented above (Baddeley's multiple-component model, Cowan's embedded-processes model, and Diamond's executive function model), we sought to test these models using fMRI studies.

To test each model using fMRI data, we must identify mixed and pure measures for each component. Based on several studies (e.g., Allen, Baddeley, & Hitch, 2014; Baddeley, 2003a; Baddeley et al., 1998; Cowan, 2010; Gathercole, Willis, & Baddeley, 1991; Wright & Diamond, 2014), we first separated tasks that are related to each component of the proposed theories. We then depicted regions that are associated with each theory and determined whether the theoretical frameworks consist of components that overlap or are isolated in the brain.

Table 1 shows the most important information for each of the three theoretical models that will be tested. The table presents the following information: theoretical model, authorship, domain (general or specific), system or subsystem/component, pure task, and brain region that is likely activated.

Model	Author	Domain	Subsystem	Task	Brain Regions
Multiple-component working memory	Baddeley, Allen, & Hitch (2011)	Specific	Working memory	Complex span tasks	Overlapping regions
		General	Central Executive	N-Back / Flanker task	Dorsolateral Prefrontal cortex
		Specific	Episodic Buffer	Delayed match-to-sample	Somatosensory Association cortex
		Specific	Visuospatial Sketchpad	Simple spatial span task	Associative visual cortex
		Specific	Phonological Loop	Simple word or letter span tasks	Premotor cortex, Pars opercularis-part of Broca's area, and Supramarginal gyrus-part of Wernicke's area
Embedded-processes model	Cowan (2010)	General	Embedded-processes	Complex span tasks	Overlapping regions
		General	Central Executive	N-Back	Prefrontal cortex
		General	Long-term memory	Recall and Feeling of knowing tasks	Somatosensory Association cortex
		Specific	Short-term memory	Delayed match-to-sample	Hippocampus
		Specific	Sensory memory	Delayed match-to-sample	Sensorial cortex
		Specific	Focus of attention	Orienting of attention / Flanker task	Locus Cerulean, Hippocampus and Anterior Cingulated cortex
Executive functions	Diamond (2013)	General	Executive functions	Maze tasks and tower tasks	Overlapping regions of the Prefrontal cortex
		Specific	Inhibitory Control	Simon task, Stroop task and Flanker task	Orbitofrontal cortex and Anterior Cingulated cortex
		Specific	Working memory	Complex span tasks	Dorsolateral Prefrontal cortex
		Specific	Short-term memory	Simple span tasks	Hippocampus
		Specific	Cognitive flexibility	Fluency tasks and Card sorting tasks	Overlapping regions of the Prefrontal cortex

Table 1. Each theoretical model to explain working memory components, with related pure tasks and brain regions.