Short-term Memory

Authored by
mohammad looti

June 5, 2026

Short-term memory (or "primary" or "active memory") is the capacity for holding a small amount of information in mind in an active, readily available state for a short period of time. The duration of short-term memory (when rehearsal or active maintenance is prevented) is believed to be in the order of seconds. A commonly-cited capacity is 7 ± 2 elements. In contrast, long-term memory indefinitely stores a seemingly unlimited amount of information.

Short-term memory should be distinguished from working memory which refers to structures and processes used for temporarily storing and manipulating information (see more details below).

Existence of a separate store

The idea of a division of memory into short term and long term dates back to the 19th century. A classical model of memory developed in the 1960s, now known to be flawed, assumed that all memories pass from a short-term to a long-term store after a small period of time. This model is referred to as the "modal model" and has been most famously detailed by Shiffrin. The exact mechanisms by which this transfer takes place, whether all or only some memories are retained permanently, and indeed the existence of a genuine distinction between the two stores, remain controversial topics among experts.

One form of evidence, cited in favor of the separate existence of a short-term store comes from anterograde amnesia, the inability to learn new facts and episodes. Patients with this form of amnesia, have intact ability to retain small amounts of information over short time scales (up to 30 seconds) but are dramatically impaired in their ability to form longer-term memories (a famous example is patient HM). This is interpreted as showing that the short-term store is spared from amnesia and other brain diseases

Other evidence comes from experimental studies showing that some manipulations (e.g., a distractor task, such as repeatedly subtracting a single-digit number from a larger number following learning; cf Brown-Peterson procedure, The Brown-Peterson Procedure is used in cognitive experiments, originally as a way to test the limits of short-term memory. Basically, after viewing a list of items to be remembered, a participant will count backwards by 3's from some large number for a short period of time (e.g., 30 seconds). In other words, for 30 seconds, a person would count backwards by 3 from 999. Effectively, this task prevents people from rehearsing the list of items, thereby allowing researchers to test how long short-term memory lasts independently of maintenance rehearsal (the simple repeating of a word over and over again) impair memory for the 3 to 5 most recently learned words of a list (presumably still held in short-term memory), while leaving recall for words from earlier in the list (presumably stored in long-term memory) unaffected; other manipulations (e.g., semantic similarity of the words) affect only memory for earlier list words, but do not affect memory for the last few words in a list. These results show that different factors affect short term recall (disruption of rehearsal) and long-term recall (semantic similarity). Together, these findings show that long-term memory and short-term memory can vary independently of each other.

Not all researchers agree that short-term and long-term memory are separate systems. Some theorists propose that memory is unitary over all time scales, from milliseconds to years. Support for the unitary memory hypothesis comes from the fact that it has been difficult to demarcate a clear boundary between short-term and long-term memory. For instance, Tarnow shows that the recall probability vs. latency curve is a straight line from 6 to 600 seconds (ten minutes), with the probability of failure to recall only saturating after 600 seconds. If there were really two different memory stores operating in this time frame, one could expect a discontinuity in this curve. Other research has shown that the detailed pattern of recall errors looks remarkably similar for recall of a list immediately after learning (presumably from short-term memory) and recall after 24 hours (necessarily from long-term memory).

Further evidence against the existence of a short-term memory store comes from experiments involving continual distractor tasks. In 1974, Robert Bjork and William B. Whitten presented subjects with word pairs to be remembered; however, before and after each word pair, subjects had to do a simple multiplication task for 12 seconds. After the final word-pair, subjects had to do the multiplication distractor task for 20 seconds. In their results, Bjork and Whitten found that the recency effect (the increased probability of recall of the last items studied) and the primacy effect (the increased probability of recall of the first few items) still remained. These results would seem inconsistent with the idea of short-term memory as the distractor items would have taken the place of some of the word-pairs in the buffer, thereby weakening the associated strength of the items in long-term memory. Bjork and Whitten hypothesized that these results could be attributed to the memory processes at work for long-term memory retrieval versus short-term memory retrieval.

Ovid J.L. Tzeng (1973) also found an instance where the recency effect in free recall did not seem to result from the function of a short-term memory store. Subjects were presented with four study-test periods of 10 word lists, with a continual distractor task (20-second period of counting-backward). At the end of each list, participants had to free recall as many words from the list as possible. After free-recall of the fourth list, participants were asked to free recall items from all four lists. Both the initial free recall and the final free recall showed a recency effect. These results went against the predictions of a short-term memory model, where no recency effect would be expected in either initial or final free recall.

Koppenaal and Glanzer (1990) attempted to explain these phenomena as a result of the subjects' adaptation to the distractor task, which therefore allowed them to preserve at least some of the functions of the short-term memory store. As evidence, they provided the results of their experiment, in which the long-term recency effect disappeared when the distractor after the last item differed from the distractors that preceded and followed all the other items (e.g. arithmetic distractor task and word reading distractor task). Thapar and Greene challenged this theory. In one of their experiments, participants were given a different distractor task after every item to be studied. According to Koppenaal's and Glanzer's theory, there should be no recency effect as subjects would not have had time to adapt to the distractor; yet such a recency effect remained in place in the experiment.

One proposed explanation of the existence of the recency effect in a continual distractor condition, and the disappearance of it in an end-only distractor task is the influence of contextual and distinctive processes. According to this model, recency is a result of the final items' processing context being similar to the processing context of the other items and the distinctive position of the final items versus items in the middle of the list. In the end distractor task, the processing context of the final items is no longer similar to the processing context of the other list items. At the same time, retrieval cues for these items are no longer as effective as without the distractor. Therefore, the recency effect recedes or vanishes. However, when distractor tasks are placed before and after each item, the recency effect returns, because all the list items once again have similar processing context.

Biological basis

It is proposed by Tarnow that short term memory involves the firing of neurons which depletes the Readily Releasable Pool (RRP) of neurotransmitter vesicles at presynaptic terminals. The pattern of depleted presynaptic terminals represents the long term memory trace and the depletion itself is the short term memory. After the firing has slowed down, endocytosis causes short term memory to decay (Endocytosis is the process by which cells absorb molecules (such as proteins) by engulfing them. It is used by all cells of the body because most substances important to them are large polar molecules that cannot pass through the hydrophobic plasma or cell membrane). If the endocytosis is allowed to finish (the memory is not activated again), the pattern of exhausted postsynaptic terminals becomes invisible and the short term memory disappears. The long term memory remains as the metastable pattern of the neuronal excitations.

Relationship with working memory

The relationship between short-term memory and working memory is described differently by various theories, but it is generally acknowledged that the two concepts are distinct. Working memory is a theoretical framework that refers to structures and processes used for temporarily storing and manipulating information. As such, working memory might also be referred to as working attention. Short-term memory generally refers, in a theory-neutral manner, to the short-term storage of information, and it does not entail the manipulation or organization of material held in memory. Thus while there are short-term memory components to working memory models, the concept of short-term memory is distinct from these more hypothetical concepts. Within Baddeley's influential 1986 model of working memory there are two short-term storage mechanisms: the phonological loop and the visuospatial sketchpad. Most of the research referred to here involves the phonological loop, because most of the work done on short-term memory has used verbal material. In recent years, however, there has been a surge in research on visual short term memory, and also increasing work on spatial short term memory.

Duration of short-term memory

The limited duration of short-term memory immediately suggests that its contents spontaneously decay over time. The decay assumption is part of many theories of short-term memory, most notably Baddeley's model of working memory (The original model of Baddeley & Hitch was composed of three main components; the central executive which acts as supervisory system and controls the flow of information from and to its slave systems: the phonological loop and the visuo-spatial sketchpad. The slave systems are short-term storage systems dedicated to a content domain (verbal and visuo-spatial, respectively). In 2000 Baddeley added a third slave system to his model, the episodic buffer). The decay assumption is usually paired with the idea of rapid covert rehearsal: In order to overcome the limitation of short-term memory, and retain information for longer, information must be periodically repeated, or rehearsed -- either by articulating it out loud, or by mentally simulating such articulation. In this way, the information will re-enter the short-term store and be retained for a further period.

Several researchers, however, dispute that spontaneous decay plays any significant role in forgetting over the short term, and the evidence is far from conclusive.

Authors doubting that decay causes forgetting from short-term memory often offer as an alternative some form of interference: When several elements (such as digits, words, or pictures) are held in short term memory simultaneously, their representations compete with each other for recall, or degrade each other. Thereby, new content gradually pushes out older content, unless the older content is actively protected against interference by rehearsal or by directing attention to it.

Capacity of short-term memory

Whatever the cause or causes of forgetting over the short term may be, there is consensus that it severely limits the amount of new information that we can retain over brief periods of time. This limit is referred to as the finite capacity of short-term memory. The capacity of short-term memory is often called memory span, in reference to a common procedure of measuring it. In a memory span test, the experimenter presents lists of items (e.g. digits or words) of increasing length. An individual's span is determined as the longest list length that he or she can recall correctly in the given order on at least half of all trials.

In an early and highly influential article, The Magical Number Seven, Plus or Minus Two, the psychologist George Miller suggested that human short-term memory has a forward memory span of approximately seven items plus or minus two and that that was well known at the time (it seems to go back to the 19th century researcher Wundt). More recent research has shown that this "magical number seven" is roughly accurate for college students recalling lists of digits, but memory span varies widely with populations tested and with material used. For example, the ability to recall words in order depends on a number of characteristics of these words: fewer words can be recalled when the words have longer spoken duration; this is known as the word-length effect, or when their speech sounds are similar to each other; this is called the phonological similarity effect. More words can be recalled when the words are highly familiar or occur frequently in the language. Recall performance is also better when all of the words in a list are taken from a single semantic category (such as sports) than when the words are taken from different categories. According to the available evidence, the best overall estimate of short-term memory is about four pieces or "chunks" of information. In free recall it has been shown, to the contrary, that there is no such "quantized" limit, rather it is a function of memory decaying with time.

Chunking

Chunking is the process by which we can expand our ability to remember things in the short term. Chunking is also a process by which a person organizes material into meaningful groups. Although the average person may only retain about four different units in short-term memory, chunking can greatly increase a person's recall capacity. For example, in recalling a phone number, the person could chunk the digits into three groups: first, the area code (such as 215), then a three-digit chunk (123) and lastly a four-digit chunk (4567). This method of remembering phone numbers is far more effective than attempting to remember a string of 10 digits.

Practice and the usage of existing information in long-term memory can lead to additional improvements in one's ability to use chunking. In one testing session, an American cross-country runner was able to recall a string of 79 digits after hearing them only once by chunking them into different running times (e.g. the first four numbers were 1518, a three-mile time.)

Factors affecting short term memory

It is very difficult to demonstrate the exact capacity of STM because it will vary depending on the nature of the material to be recalled. There is currently no way of defining the basic unit of information to be stored in the STM store. It is also possible that STM is not the store described by Atkinson and Shiffrin. In that case, the task of defining the task of STM becomes even more difficult.

However, capacity of STM can be affected by the following: Influence of long-term memory, Reading aloud, Pronunciation time and Individual differences.