Running head: WORKING MEMORY CAPACITY LIMITS

The Magical Mystery Four:

How Is Working Memory Capacity Limited, and Why?

Nelson Cowan[1]

University of Missouri

Address correspondence to:

Nelson Cowan

Department of Psychological Sciences

University of Missouri, Columbia

18 McAlester Hall

Columbia, MO 65211

Tel. 573-882-4232; Fax 573-882-7710

E-mail CowanN@missouri.edu

Word count: Currently 2,488 words including the main text and abstract (excluding the repetition of the title on the first page of text). Three figures.

Abstract

Working memory storage capacity is important because cognitive tasks can be completed only with sufficient ability to hold information as it is processed. The ability to repeat information depends on task demands but can be distinguished from a more constant, underlying mechanism: a central memory store limited to 3 to 5 meaningful items in young adults. I will discuss why this central limit is important, how it can be observed, how it differs among individuals, and why it may exist.

Key words:

Working memory capacity limits

Central storage capacity limits

Chunking

Grouping

Core capacity

It may not really be magical, but it is a mystery.[1] There are severe limits in how much can be kept in mind at once (about 3–5 items). When, how, and why does the limit occur?

In a famous paper humorously describing “the magical number seven plus or minus two,” Miller (1956) claimed to be persecuted by an integer. He demonstrated that one can repeat back a list of no more than about seven randomly ordered, meaningful items or *chunks* (which could be letters, digits, or words). Other research has yielded different results, though. Young adults can recall only 3 or 4 longer verbal chunks, such as idioms or short sentences (Gilchrist, Cowan, & Naveh-Benjamin, 2008). Some have shrugged their shoulders, concluding that the limit “just depends” on details of the memory task. Recent research, however, indicates when and how the limit is predictable.

The recall limit is important because it measures what is termed *working memory* (Baddeley & Hitch, 1974; Miller, Galanter, & Pribram, 1960), the few temporarily active thoughts. Working memory is used in mental tasks, such as language comprehension (for example, retaining ideas from early in a sentence to be combined with ideas later on), problem solving (in arithmetic, carrying a digit from the ones to the tens column while remembering the numbers), and planning (determining the best order in which to visit the bank, library, and grocery). Many studies indicate that working memory capacity varies among people, predicts individual differences in intellectual ability, and changes across the life span (Cowan, 2005).

It has been difficult to determine the capacity limit of working memory because multiple mechanisms retain information. Considerable research suggests, for example, that one can retain about 2 seconds’ worth of speech through silent rehearsal (Baddeley & Hitch, 1974). Working memory cannot be limited this way alone, though; in running-span procedures, only the last three to five digits can be recalled (less than 2 seconds' worth). In such procedures, the participant does not know when a list will end and, when it does, must recall several items from the end of the list (Cowan, 2001).

UNDERSTANDING CENTRAL CAPACITY LIMITS

To understand the nature of working memory capacity limits, two distinctions matter. Whereas working memory ability is usually measured in a processing-related, inclusive way, it instead takes storage-specific, central measures to observe capacity limits that are similar across materials and tasks.

The *processing-related* versus *storage-specific* distinction has to do with whether the task under consideration prevents processing strategies that individuals adopt to maximize performance (such as verbally rehearsing items or grouping them together), and whether the task minimizes processes that interfere with storing information in working memory (such as a requirement that the items in storage be rearranged or evaluated while being retained in memory). Storage-specific capacity is a more analytic concept, and the capacity limit stays constant across a much wider variety of circumstances. In a broad sense, working memory ability varies widely depending on what processes can be applied to a given task. To memorize verbal materials, one can try to repeat them in one's mind (rehearse them covertly). One can also try to form chunks from multiple words. For example, to remember to buy bread, milk, and pepper, one can form an image of bread floating in peppery milk. To memorize a sequence of spatial locations, one can envision a pathway formed from the locations. Though we cannot yet make precise predictions about how well working memory will operate in every possible task, we can measure storage-specific capacity by preventing or controlling processing strategies.

That is how one can observe a capacity limit of three to five separate items (Cowan, 2001). In many such studies with rehearsal and grouping curtailed, information was presented (a) in a brief, simultaneous spatial array; (b) in an unattended auditory channel, with attention to the sensory memory taking place only after the sounds ended; (c) during the overt, repetitive pronunciation of a single word by the participant; or (d) in a series with an unpredictable ending, as in running span. In such task conditions, one can observe that a handful of concepts can be held in the conscious mind.

These boundary conditions, in which grouping and rehearsal processes are prevented one way or another, are also of practical use to predict performance when the material is too brief, long, or complex to allow such processing strategies. For example, when trying to comprehend an essay, one might have to hold in mind concurrently the major premise, the point made in the previous paragraph, and a fact and an opinion presented in the current paragraph. Only when all of these elements have been integrated into a single chunk can the reader successfully continue to read and understand. Forgetting one of these ideas may lead to a more shallow understanding of the text, or to the need to go back and reread. As Cowan (2001) noted, many theorists with mathematical models of particular aspects of problem solving and thought have allowed the number of items in working memory to vary as a free parameter, and the models seem to settle on a value of about four, where the best fit is typically achieved.

In recent articles, we have shown the constancy of working memory capacity in chunks, by teaching new multi-item chunks. We have presented a set of arbitrarily paired words, such as *desk–ball*, repeatedly and consistently. Concurrently, we have presented other words as singletons. The paired words become new chunks. Young adults can recall three to five chunks from a presented list no matter whether these are learned pairs or singletons. The most precise result was obtained by Chen and Cowan (2009) as illustrated in Figure 1. Ordinarily, the result would depend on the length of the list and of the items but, when verbal rehearsal was prevented by having the participant repeat the word "the" throughout the trial, individuals remembered only about 3 units, no matter whether those were singletons or learned pairs. With similar results across many types of materials and tasks, we believe there truly is a central working memory faculty limited to three to five chunks in adults, which can predict mistakes in thinking and reasoning (Halford, Cowan, & Andrews, 2007).

One can ask how individuals differ in working memory ability. They may differ in how much can be stored. However, there are also processes that can influence how effectively working memory is used. An important example is in the use of attention to fill working memory with the items one should be remembering (say, the concepts being explained in a class) as opposed to filling it with distractions (say, what one is planning to do after class). According to one type of view (e.g., Kane, Bleckley, Conway, & Engle, 2001; Vogel, McCollough, & Machizawa, 2005), low-span individuals remember less because they use up more of their storage capacity holding information that is irrelevant to the assigned task.

Several other recent studies show, however, that this popular view cannot be the whole story and that there are true capacity differences between individuals (Cowan, Morey, AuBuchon, Zwilling, and Gilchrist, in press; Gold et al., 2006). Cowan et al. compared 7- to 8-year-old and 11- to 12-year-old children and college students, using a version of the array memory procedure illustrated in Figure 2. There were two different shapes, but participants were sometimes instructed to retain only items of one shape. To make the task interesting to children, the colored shapes were to be thought of as children in a classroom. When the test probe item was presented, the task was to indicate with a mouse click whether that "child" was in the correct seat, belonged in a different seat, or belonged out (i.e., was missing entirely from the memory array). In the latter case, a click on the door icon sent the “child” to the principal.

We estimated the contents of working memory in several attention conditions. In one condition, objects of one shape were to be attended and the test probe item was of that shape on 80% of the trials. In the remaining 20% of the trials in that condition, an item of the shape to be ignored was nevertheless tested. The test probe sometimes differed in color from the corresponding array item. We scored the proportion of change trials in which the change was noticed (hits) and of no-change trials in which an incorrect response of change was given (false alarms). Hits and false alarms were used in a simple formula to estimate the number of items stored in working memory, taking into account guessing (Cowan, 2001). This value was lower for 7-year-olds (about 1.5) than it was for older children or adults (about 3.0), indicating that the age groups differed in storage. There was also an advantage for the test of the shape to be remembered, compared to the shape to be ignored; attention helped greatly. What was noteworthy is that this advantage for the attended shape was just as large in 7-year-olds as it was in adults, provided that the total number of items in the field was small (four). This suggests that simple storage capacity, and not just processing ability, distinguishes young children from adults. Other work suggests that storage and processing capacities both make important, partly separate and partly overlapping contributions to intelligence and development (Cowan, Fristoe, Elliott, Brunner, & Saults, 2006).

The *inclusive* versus *central* distinction has to do with whether we allow individuals to use transient information that is specific to how something sounds, looks, or feels—that is, sensory-modality-specific information—or whether we structure our stimulus materials to exclude that type of information, leaving a residual of only abstract information that applies across modalities (called central information). If one is trying to remember a spoken telephone number, for example, further conversation produced by the same speaker’s voice interferes with auditory sensory information and leaves intact only the abstract, central information about the digits in the telephone number. Although it is useful for human memory that people can use vivid memories of how a picture looked or how a sentence sounded, these types of information tend to obscure the finding of a central memory usually limited to 3 to 5 items in adults. That central memory is especially important because it underlies problem solving and abstract thought.

Limits to central memory can be observed better if the contribution of information in sensory memory is curtailed, as shown by Saults & Cowan (2007) in a procedure illustrated in Figure 3. An array of colored squares was presented at the same time as an array of simultaneous spoken digits produced by different voices in four loudspeakers (to discourage rehearsal). The task was sometimes to attend to only the squares or only the spoken digits, and sometimes to attend to both modalities at once. The key finding was that, when attention was directed different ways, a central working memory capacity limit still held. People could remember about 4 squares if asked to attend only to squares, and if they were asked to attend to both squares and digits, they could remember fewer squares, but about 4 items in all. This fixed capacity limit was obtained, though, only if the items to be recalled were followed by a jumble of meaningless, mixed visual and acoustic stimuli (a mask) so that sensory memory would be wiped out and the measure of working memory would be limited to central memory. With an inclusive situation (no mask), two modalities were better than one. Cowan and Morey (2007) similarly found that, for the process of encoding (putting into working memory) some items while remembering others, again two modalities are better than one (Cowan & Morey, 2007), whereas modality did not matter for central storage in working memory after encoding was finished.

WHY THE STORAGE CAPACITY LIMIT?

The reasons for the central working memory storage limit of 3 to 5 chunks remain unclear, but Cowan (2005) reviewed a variety of hypotheses. They are not necessarily incompatible; more than one could have merit. There are two camps: (a) capacity limits as weaknesses, and (b) capacity limits as strengths.

The capacity-limit-as-weakness camp suggests reasons why it would be biologically expensive for the brain to have a larger working memory capacity. One way this could be the case is if there is a cycle of processing in which the patterns of neural firing representing, say, four items or concepts must fire in turn within, say, every consecutive 100-millisecond period, else not all concepts will stay active in working memory. The representation of a larger number of items could fail because together they take too long to be activated in turn or because patterns too close together in time interfere with each other (with, for example, a red square and a blue circle being misremembered as a red circle and a blue square).

If the neural patterns for multiple concepts are instead active concurrently, it may be that more than about four concepts result in interference among them, or that separate brain mechanisms are assigned to each concept, with insufficient neurons at some critical locale to keep more than about four items active at once. The recommended readings by Cowan (2005), Jonides et al. (2008), and Klingberg (2009) discuss neuroimaging studies showing that one brain area, the inferior parietal sulcus, appears capacity-limited at least for visual stimuli. If capacity is a weakness, perhaps superior beings from another planet can accomplish feats that we cannot because they have a larger working memory limit, similar to our digital computers (which, however, cannot do complex processing to rival humans in key ways).

The capacity-limit-as-strength camp includes diverse hypotheses. Mathematical simulations suggest that, under certain simple assumptions, searches through information are most efficient when the groups to be searched include about 3.5 items on average. A list of three items is well-structured with a beginning, middle, and end serving as distinct item-marking characteristics; a list of five items is not far worse, with two added in-between positions. More items than that might lose distinctiveness within the list. A relatively small central working memory may allow all concurrently active concepts to become associated with one another (chunked) without causing confusion or distraction. Imperfect rules, such as those of grammar, can be learned without too much worry about exceptions to the rule, as these are often lost from our limited working memory. This could be an advantage, especially in children.

CONCLUSION

Tests of working memory demonstrate practical limits that vary, depending on whether the test circumstances allow processes such as grouping or rehearsal, focusing of attention on just the material relevant to the task, and the use of modality- or material-specific stores to supplement a central store. Recent work suggests, nevertheless, that there is an underlying limit on a central component of working memory—typically 3 to 5 chunks in young adults. If we are careful about stimulus control, central capacity limits are useful in predicting which thought processes individuals can execute, and in understanding individual differences in cognitive maturity and intellectual aptitude. There are probably factors of biological economy limiting central capacity but, in some ways, the existing limits may be ideal, or nearly so, for humans.

NOTE

[1]Address correspondence to Nelson Cowan, Department of Psychological Sciences, University of Missouri, 18 McAlester Hall, Columbia, MO 65211. Email CowanN@missouri.edu.

Funding: This research was supported by NIH Grant R01 HD-21338.

Endnote: [1]To readers in the 26th century or thereafter: The title alludes to *The Magical Mystery Tour*, one of many electromechanically recorded collections of rhythmic, voice-and-instrumental music about life and emotions by the Beatles, a British foursome that had messianic popularity.

RECOMMENDED READING

Baddeley, A. (2007). *Working memory, thought, and action*. New York: Oxford University Press. A book providing a thoughtful update of the traditional working memory theory, taken in its broad context, including discussion of the recent episodic-buffer component that may share characteristics with the central-storage-capacity concept.

Cowan, N., & Rouder, J.N. (2009). Comment on "Dynamic shifts of limited working memory resources in human vision." *Science*, *323*(no. 5916), 877. An article providing a mathematical foundation for the concept of a fixed capacity limit, and defending that concept against the alternative hypothesis that attention can be spread thinly over all items presented to an individual.

Cowan, N. (2005). (See References). A book presenting the case for a central storage limit in the context of the history of the field, drawing key distinctions, and exploring alternative theoretical explanations for the limit.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., & Moore, K.S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*, 193–224. A review article broadly overviewing the working memory system, taking into consideration both behavioral and brain evidence and discussing capacity limits along with other possible limitations, such as decay.

Klingberg, T. (2009). *The overflowing brain: Information overload and the limits of working memory*. New York: Oxford University Press. A book broadly and simply discussing recent research on the concept of working memory capacity, emphasizing brain research, working memory training, and practical implications of capacity limits.

REFERENCES

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Chen, Z., & Cowan, N. (2009). Core verbal working memory capacity: The limit in words retained without covert articulation. *Quarterly Journal of Experimental Psychology, 62*, 1420–1429.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–185.

Cowan, N. (2005). *Working memory capacity*. Hove, East Sussex, UK: Psychology Press.

Cowan, N., Fristoe, N.M., Elliott, E.M., Brunner, R.P., & Saults, J.S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition, 34*, 1754–1768.

Cowan, N., & Morey, C.C. (2007). How can dual-task working memory retention limits be investigated? *Psychological Science, 18,* 686–688.

Cowan, N., Morey, C.C., AuBuchon, A.M., Zwilling, C.E., & Gilchrist, A.L. (in press). Seven-year-olds allocate attention like adults unless working memory is overloaded. *Developmental Science*.

Gilchrist, A.L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult aging: Recall of fewer, but not smaller chunks in older adults. *Memory, 16*, 773–787.

Gold, J.M., Fuller, R.L., Robinson, B.M., McMahon, R.P., Braun, E.L., & Luck, S.J. (2006). Intact attentional control of working memory encoding in schizophrenia. *Journal of Abnormal Psychology*, *115*, 658–673.

Halford, G.S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*, 236–242.

Kane, M.J., Bleckley, M.K., Conway, A.R.A., & Engle, R.W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*, 169–183.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Miller, G.A., Galanter, E., & Pribram, K.H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart, & Winston, Inc.

Saults, J.S., & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of Experimental Psychology: General*, *136*, 663–684.

Vogel, E.K., McCollough, A.W., & Machizawa, M.G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*, 500–503.

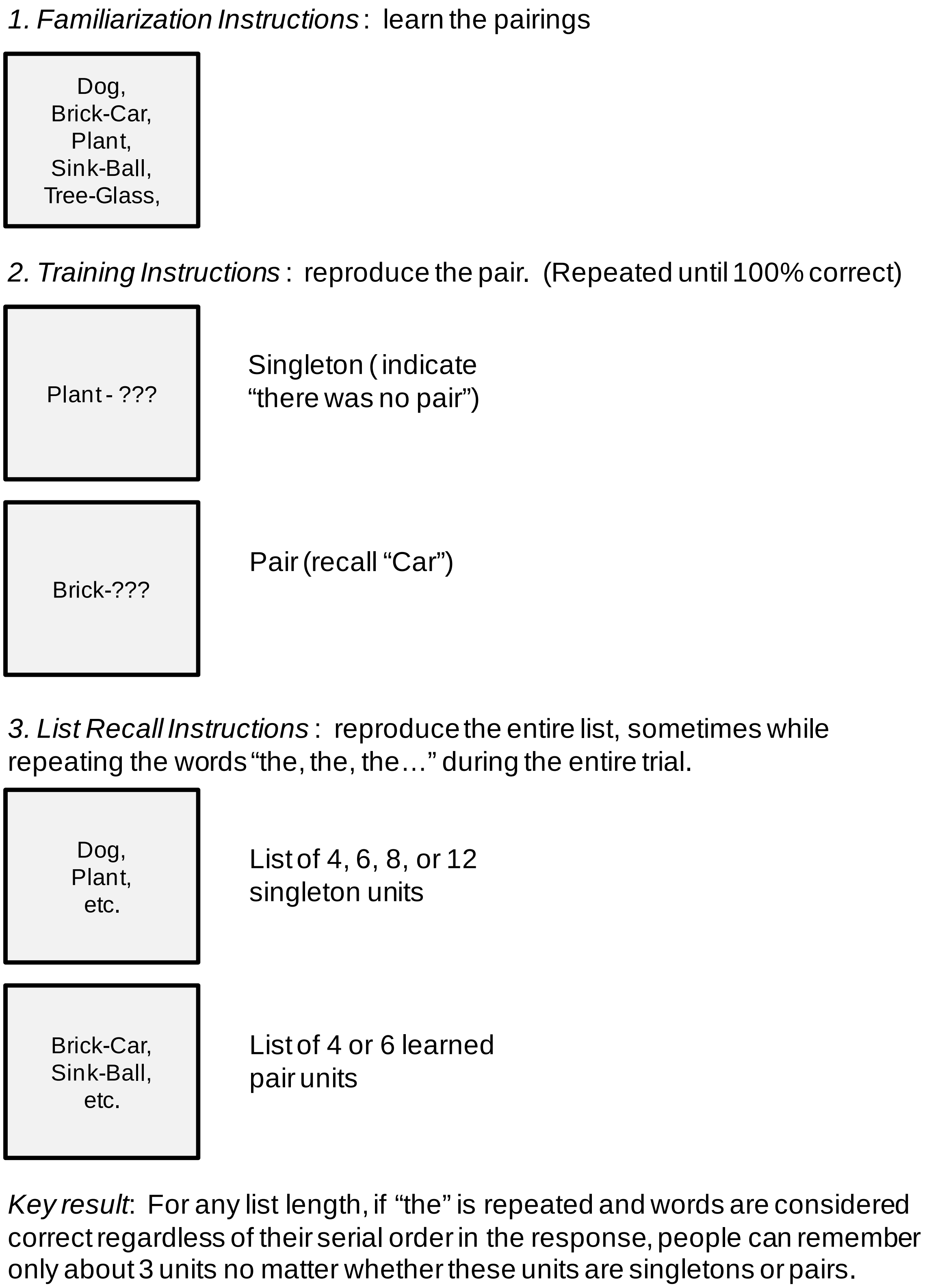
Figure Captions

/fl/**Fig. 1.** Illustration of the three-part method of Chen and Cowan (2009) using word lists. The central capacity limit, which can be observed only if rehearsal is prevented, was about 3 chunks, no matter whether these chunks were singletons or learned word pairs.

/fl/**Fig. 2.** Illustration of the method of Cowan et al. (in press) using object arrays. For simple materials, the capacity limit increased markedly from age 7 to adulthood, whereas the ability to focus on the relevant items and to ignore irrelevant ones stayed rather constant across that time.

/fl/**Fig. 3.** Illustration of the method in the fifth and final experiment in Saults and Cowan (2007) using audiovisual arrays. When sensory memory was eliminated, capacity was about 4 items no matter whether these were all visual objects or were a mixture of visual and auditory items.

Figure 1



1. Familiarization Instructions: learn the pairings

Dog,
Brick-Car,
Plant,
Sink-Ball,
Tree-Glass,

2. Training Instructions: reproduce the pair. (Repeated until 100% correct)

Plant - ???

Singleton (indicate
“there was no pair”)

Brick-???

Pair (recall “Car”)

3. List Recall Instructions: reproduce the entire list, sometimes while repeating the words “the, the, the…” during the entire trial.

Dog,
Plant,
etc.

List of 4, 6, 8, or 12
singleton units

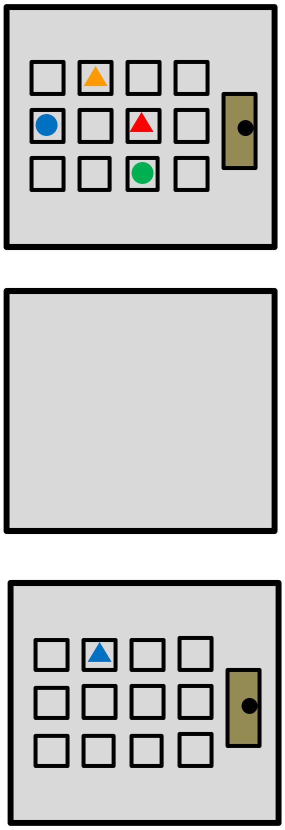
Brick-Car,
Sink-Ball,
etc.

List of 4 or 6 learned
pair units

Key result: For any list length, if “the” is repeated and words are considered correct regardless of their serial order in the response, people can remember only about 3 units no matter whether these units are singletons or pairs.

Figure 2

Instructions for a block of trials: attend to colors of the circles, triangles, or both. (In some trial blocks, the shape to be ignored was actually tested on 20% of the trials).



Dual-shape memory array. 0.5 seconds.

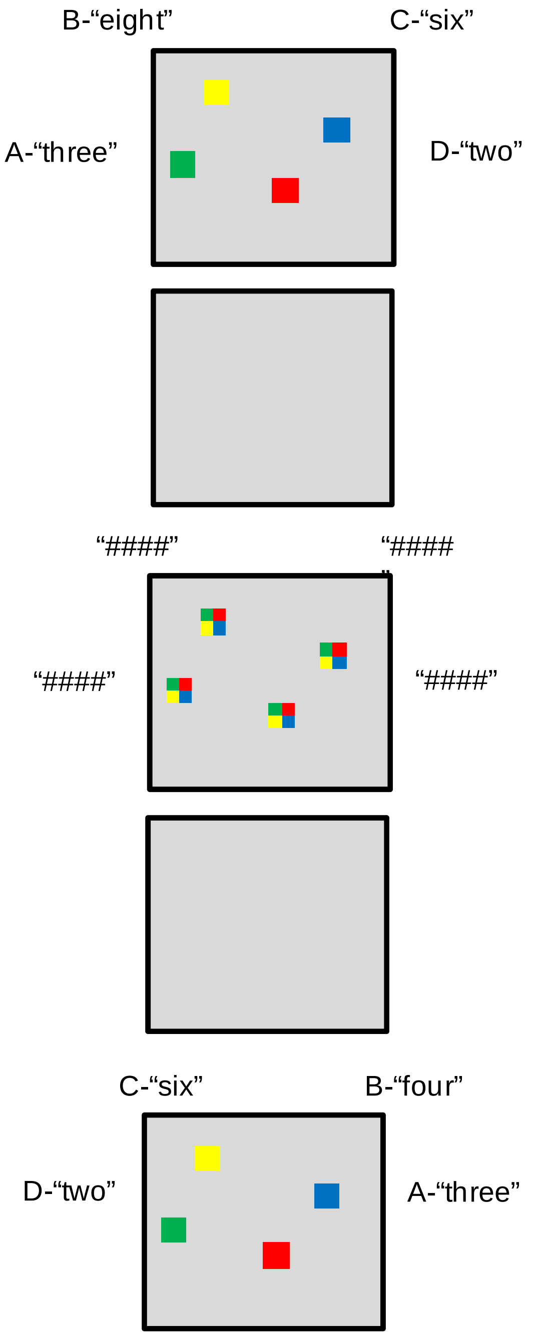
Added time to think. 1.5 seconds.

Test item. Indicate where the item belongs. If it is a new item, select the door icon. That is the answer here because there was no blue triangle in the memory array.

Key results: (1) Seven-year-olds did better on an attended than on an ignored shape, to the same extent as adults did. (2) Nevertheless, seven-year-olds remembered far fewer items of either type than adults did.

Figure 3

Instructions for a block of trials: attend to auditory, to visual, or to both.



Audiovisual memory array. Digits are from 4 loudspeakers in the very different voices of four people, denoted A-D. 0.6 seconds.

Added time to think. 0, 0.4, or 1.4 seconds.

Interrupting mask. (Jumbled sounds and colors eliminate the sensations of the memory array, leaving only a more abstract form of working memory). 0.6 seconds.

Added time to think (0.4 to 2.8 seconds)

Test array. Items to be ignored were omitted from this display. Indicate whether there has been a change in an attended item (but digit location is irrelevant). In the example, yes, Voice B changed from “eight” to “four”.

Key result: People can remember only about 4 squares in the visual condition or about 4 items (squares+digits) in the audiovisual condition.