1 About memory

To understand how to practice or revise effectively, you need to understand some basic principles about how memory works. This chapter covers:

- the 8 basic principles of memory
- how neurons work
- what working memory is and why it's so important
- what consolidation is and why it matters

'Practice' (a term I use to cover both revision of information and the practice of skills) is a deceptively simple concept. Everyone thinks they know what it means, and what they think it is, basically, is repetition. When we have to remember an unknown phone number long enough

to dial it, we repeat it to ourselves. As children told to learn a poem, we repeated it until we had it memorized. Learning to drive, we repeated the necessary actions over and over again. Repetition is at the heart of learning.

But simple repetition is the least effective learning strategy there is.

How do we reconcile these statements? Well, repetition is crucial to cement a memory, but the untutored way of doing it wastes a great deal of time, and still results in learning that is less durable than more efficient strategies.

In this book, I'm going to tell you the 10 principles of effective repetition, why they work, and how they work. We'll look at examples from science, mathematics, history, foreign language learning, and skill learning. At the end of it, you'll know how to apply these principles in your study and your daily life.

Let's start with a very brief look at how memory works.

THE 8 BASIC PRINCIPLES OF MEMORY

The most fundamental thing you need to understand about memory is that it is not a recording. When we put information into our memory, we don't somehow copy the real-world event, as a video camera might, but rather we select and edit the information. For this reason, putting information into memory is called **encoding**. This is why I habitually talk about memory codes rather than memories. It's a reminder that nothing in your memory is a 'pure' rendition, a faithful copy. We create our memory codes, and when we try and retrieve a memory, it is this coded information that we are looking for. (Like a computer, what

our brain processes is information — when I use the word 'information', I don't just mean 'facts', but images and skills and events and everything else we file in long-term memory.)

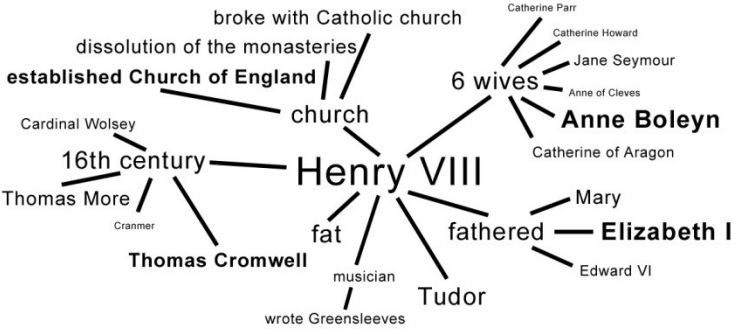
Why does it matter that the information is coded?

Because what you think you are looking for may not be precisely what is there. How easy it is to remember something (retrieve a memory code) depends on the extent to which the code matches what you think you're looking for.

For example, say you are trying to remember someone's name. You might think it begins with T, or that it's unusual, or very common, or sounds something like -immy, or that it's old-fashioned, or ... Whatever your idea is, the point is that there *is* an idea, a starting point, a clue (we call it a **retrieval cue**). How likely you are to retrieve the memory code depends on how good a clue it is.

This is because memory codes are linked together in a network. Remembering is about following a trail through the network, following the links. No surprise then that your starting point (the retrieval cue) is crucial.

For example, consider this simplified memory code for Henry VIII:



The size of the words reflect how strong those parts of the code are — Anne Boleyn, for example, is for most of us the most memorable of Henry's wives; Elizabeth the most memorable of his children.

Accordingly, it would be a lot easier to retrieve "Henry VIII" if the retrieval cue was "father of Elizabeth I" than if it was "father of Edward VI", or if the cue was "established the Church of England" rather than "Cranmer's king", or, worst of all, "16th century musician". (Do note that information in a memory code is not necessarily true! For example, Henry VIII did not actually write the song *Greensleeves*, but it is a common belief that he did. 'Information' is a blanket word to cover a type of content; the statement "The grass is green" and the statement "The grass is red" contain the same amount of information, although only one of the statements is true.)

The trail through memory resembles a trail through a jungle. Much-travelled paths will be easier and quicker to follow. Paths that have been used recently will be easier to find than old disused trails.

There are eight fundamental principles encompassed in these simple ideas:

1. **code principle**: memories are selected and edited codes.

2. **network principle**: memory consists of links between associated codes.

3. **domino principle**: the activation of one code triggers connected codes.

4. **recency effect**: a recently retrieved code will be

more easily found.
If you were watching the TV program *The Tudors* last night, it would be much easier to call up Henry VIII's name up again than it would be if you hadn't thought of him since school.

5. **priming effect**: a code will be more easily found if linked codes have just been retrieved.
Having been thinking of Henry VIII, you will find it easier to retrieve "Walter Raleigh" (linked to Elizabeth I), compared to a situation where you were asked, out of the blue, who that guy was who put his cloak across the puddle for Queen Elizabeth to walk over.

6. **frequency** (or **repetition**) **effect**: the more often a code has been retrieved, the easier it becomes to find.

7. **matching effect**: a code will be more easily found the more closely the retrieval cue matches the code.
This can be seen in jokes: if you were asked, "What did the tree do when the bank closed?", you'd probably realize instantly that the answer had something to do with "branch", because "branch" is likely to be a strong part of both your "tree" code and your "bank" code. On the other hand, if you were asked, "What tree is made of stone?", the answer (lime tree) is not nearly as easily retrieved, because "lime" is probably not a strong part of either your "tree" code or your "stone" code.

8. **context effect**: a code will be more easily found if the encoding and retrieval contexts match.
If you learned about Henry VIII from watching *The Tudors* on TV, you will find it easier to remember facts about Henry VIII when you're sitting watching TV. We use this principle whenever we try and

remember an event by imagining ourselves in the place where the event happened.

These principles all affect how practice works and what makes it effective, but three are especially important. The recency and priming effects remind us that it's much easier to follow a memory trail that has been activated recently, but that's not a strength that lasts. Making a memory trail permanently stronger requires repetition (the frequency effect). This is about neurobiology: every time neurons fire in a particular sequence (which is what happens when you 'activate' a memory code), it makes it a little easier for them to fire in that way again.

The frequency effect is at the heart of why practice is so important. The recency and priming effects are at the heart of why most people don't practice effectively.

HOW NEURONS WORK

Let's take a very brief look at this business of neurons firing. Neurons are specialized brain cells. We might think of them as nodes in the network. Neurons are connected to each other through long filaments, one long one (the **axon**) and many very short ones (**dendrites**). It is the long axon that carries the outgoing signal from the neuron. The dendrites receive the incoming signals, and they do this through specialized receptors called **synapses**. For here's the thing: neurons aren't physically connected. Messages are carried through the network by electrical impulses along the filaments, which induce chemical responses at the synapses. Specialized chemicals called **neurotransmitters** travel the very short gap between the synapses on one neuron to those on a nearby one.

In other words, information is carried within the neuron in the form of electrical impulses (as it is in your radio and television), is then transformed into a chemical format so that it can cross the gap between neurons, and then translated back into electrical impulses in the receiving neuron.

Being carried as an electrical signal has an important implication: how fast we think (and how well, as we'll see in the next section) depends on how fast the signals are flowing. The speed of the electrical signal depends on the wiring. As with the wiring in your home, the 'wires' (axons) are sheathed in insulation. The better insulated, the less 'loss' in the signal, the faster the signal can travel. In the brain, this insulation is called **myelin**. Because myelin is white(ish), and the cell bodies are gray, we commonly refer to 'gray matter' and 'white matter'.

Myelin tends to degrade over time, and this degradation is one of the factors implicated in cognitive decline in old age. Myelin degradation can also occur in certain medical conditions (multiple sclerosis being the prime example).

But how quickly the signals move is only part of the story — the other part is how far the signals have to travel. Axons can be very long, but information moves more quickly when the connection between two neurons is very short. Consider how many neurons you need to activate to have a coherent thought, and you'll realize that you'll do your best thinking when the neurons you need are all clustered tightly together.

Here's the last crucial concept: a neuron doesn't care what information it carries; a neuron, like your brain, is flexible. However, if you keep sending the same (or similar) information through, a small network of closely arranged

neurons will develop to carry that specific information. With practice (the frequency effect), the connections between the neurons will grow ‘stronger’ — more used to carrying that information across particular synapses, more easily activated when triggered.

WORKING MEMORY — A CONSTRAINING FACTOR

It seems incredible that we can store all the memories we accumulate in such a system, but we have some 200 billion neurons in our brain, and each neuron has about 1000 synapses on average. These are unimaginable numbers. But although our memory store is vast, as in a real jungle we can't see very much of it at a time. In fact, it's quite remarkable how little we can 'see' at any point, and this limitation is one of the critical constraints on our learning and our understanding.

We call the tiny part of memory that we are aware of, **working memory**. When you put information into your memory, the encoding takes place in working memory. When you drag it out of your memory, you pull it into working memory. When you read, you are using working memory to hold each word long enough to understand the complete sentence. When you think, it is working memory that holds the thoughts you are thinking.

As we all know to our cost, working memory is very small. Try and hold an unfamiliar phone number in your mind long enough to dial it and you quickly realize this. Probably the most widely known ‘fact’ about working memory is that it can only hold around seven **chunks** of information (between 5 and 9, depending on the individual). But we know now that working memory is even smaller than that.

The ‘magic number 7’ (as it has been called) applies to how much you can hold if you actively maintain it — that is, repeat it to yourself. In the absence of this deliberate circulation, it is now thought that working memory can only hold around four chunks (between three and five), of which only one can be attended to at any one time (that is, only one is ‘in focus’).

Although it sounds like a small difference, the difference between having a working memory capacity of three chunks or one that can hold five chunks has significant effects on your cognitive abilities. Your working memory capacity is closely related to what is now called **fluid intelligence**, meaning the part of an IQ test that has nothing to do with knowledge but depends almost entirely on your ability to reason and think quickly.

While working memory capacity may seem to be a ‘fixed’ attribute, something you are born with, it does increase during childhood and adolescence, and tends to decrease in old age. There have been a number of attempts to increase people's working memory capacity through training, some of which have had a certain amount of success, but most of this success has been with people who have attention difficulties. It is much less clear that training can increase the working memory capacity of an individual without cognitive disabilities.

At a practical level, however, differences in working memory capacity have a lot to do with how well we form our memory codes — with our skill in leaving out irrelevant material, and our skill at binding together the important stuff into a tightly-bound network. This is implicit in the word ‘chunks’.

Although working memory can hold only a very small

number of chunks, ‘chunks’ is the escape clause, as it were — for what constitutes a chunk is a very flexible matter. For example, 1 2 3 4 5 6 7 are seven different chunks, if you remember each digit separately (as you would if you were not familiar with the digits, as a young child isn't). But for those of us who are well-versed in our numbers, 1 through to 7 could be a single chunk. Similarly, these nine words:

1. brown
2. the
3. jumps
4. dog
5. over
6. quick
7. lazy
8. the
9. fox

could be nine chunks, or, in a different order (“the quick brown fox jumps over the lazy dog”), one chunk (for those who know it well as an example used in typing practice). At a much higher level of expertise, a chess master may have whole complex sequences of chess moves as single chunks.

Think back to what I said about how clusters of neurons become more strongly and closely connected with practice, and you'll see why practice is the key to functionally increasing your working memory capacity. The key is your chunks. A chunk is a very tight cluster. Such clusters enable you to increase how much information you can 'hold' in working memory.

I said that working memory contains a certain number of chunks, but to a large extent this way of thinking about it is a matter of convenience. It's more precise to say that the amount of information you can hold depends on how fast you are moving it. Because here's the thing about working memory — nothing stays in it for more than a couple of seconds, if you're not consciously keeping it active. This is why you have to keep repeating that phone number: you have to bring each digit back into 'focus' before its time is up and it fades back into the long-term store.

Let's go back to my earlier statement that the 'magic number 7' has now been reduced to 4 in the absence of deliberate repetition. We can reconcile these two numbers through another concept: that working memory (the 'inner circle') is surrounded by an outer area, in which, say, 3 items that have recently been in working memory can hover for a while, ready to be pulled back in easily.

Let's see that at work in a simple equation. Say you were given this problem to solve:

$(38 \times 4)/3$

How you solve this will depend on your mathematical expertise, but a common way would be to break it down to:

$30 \times 4 = 120$

$8 \times 4 = 32$

$120 + 32 = 152$

$150/3 = 50$

$2/3 = {}^{2}/_{3}$

$50 + {}^{2}/_{3} = 50\ {}^{2}/_{3}$

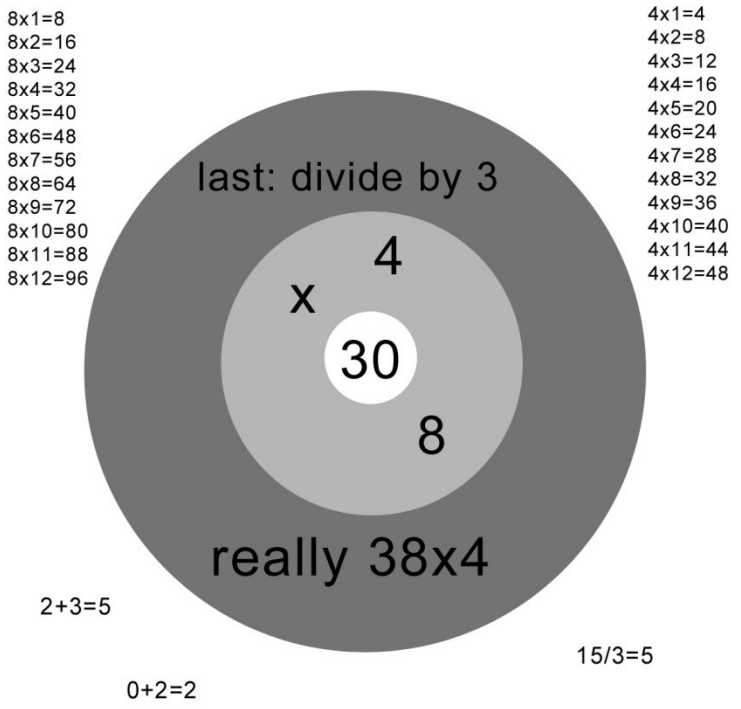
Now think of the working memory flow needed to achieve this. First, you separate 38 into 30 and 8, moving your focus from 38 to 30, between 30 and 4 as you produce a new number (120) which briefly becomes the focus as you update working memory (30 and 4 can be discarded, replaced by this new number). Now you pull the waiting 8 into focus, perform the updating operation on it (multiply by 4), and replace it with the new number, 32. Now you must pull 120 back into focus as you add it to 32 and update working memory again, replacing 32 and 120 with the new number 152. Now you need the "divide by 3" (I hope it's still waiting! It all depends how fast you've been.) And so on.

Let's think about what you have to hold in working memory while you perform this fairly simple calculation:

> 30 x 4 = 120 (while performing this operation you need to hold in mind that 30 is just part of 38 and that you'll need to multiply the 8 by 4 and then add the two sums together to get a new sum that will then need to be divided by 3)
>
> 8 x 4 = 32 (performing this operation while holding in mind: the previous sum (120); the need to add the two sums together; the later need to divide by 3)
>
> 120 + 32 = 152 (holding in mind the division by 3)
>
> 150/3 = 50 (holding in mind the 2 left over)
>
> 2/3 = $^{2}/_{3}$ (holding in mind the 50)
>
> 50 + $^{2}/_{3}$ = 50 $^{2}/_{3}$

Here's a visual representation of the first step (the numbers floating in the space beyond the rings represent

information in your long-term memory store):



You can see working memory is already filling up, but if you're not skilled at math, each of these operations may involve a little more work, especially if you don't know your tables well, so that "32" doesn't come instantly to mind when you see "8 x 4". In such a case, there'll be even less space available for those other parts of the equation that you're holding in mind.

Moreover, if you suffer from 'math anxiety', you might also have other thoughts cluttering up the space — thoughts like "oh I'm hopeless at math", memories of failure, and the like.

On the other hand, you might have approached the equation this way:

> 38 doubled is 76 (at this point, you're holding in mind the intention to double it again, and to divide the answer by 3)
>
> 76 doubled is 152 (at this point, you only need to hold in mind the intention to divide the answer by 3)
>
> 152/3 is 50 with 2 left over
>
> $50\ ^{2}/_{3}$

Because working memory is fundamentally time-based, a lot comes down to how quickly you can perform basic operations — which is simply another way of saying how accessible the information is. Because I had a teacher at school who used to put a number on the board and tell us to keep doubling (or halving) until he returned to the room, I have no trouble instantly transforming 38 to 76, 76 to 152. It would be a different story if I had to laboriously say to myself: 70 + 70 = 140, 6 + 6 = 12, 140 + 12 = 152.

This is why you can functionally (i.e., for practical purposes) increase your working memory capacity in specific areas by making your long-term knowledge more readily accessible — it all comes down your level of expertise, or to put it another way: practice.

Understanding that memory codes flow between long-term memory, the focus, the inner ring of working memory, and the outer ring of working memory, is critical for understanding the fundamental principles of effective practice. The flow is governed by time and attention. Once an item is out of focus, it only has a couple of seconds before

it will retire out of the inner ring into the outer ring, unless you bring it back into focus. Once it's in the outer ring, it can only stay there a short time before it will fade back into the vast sea of long-term memory — unless, of course, you bring it back into focus.

It is by keeping the information moving, therefore, that you keep it in working memory. Fast speakers, and fast thinkers, have an advantage.

Understanding the limits of your working memory capacity helps you work out the most effective approach to your own practice, in particular, how much information you should try and handle at one time. It also helps you recognize when a particular learning task might be more demanding of your working memory.

The demands any cognitive task put on working memory we call **cognitive load**. Being able to assess the cognitive load of a task gives you the opportunity to reduce the load where necessary. Often, the reason a student has trouble understanding or performing a task is because the cognitive load is too much for them. Note that the cognitive load of a task is not a fixed amount, but varies for the individual, because, of course, it depends on the chunks (that is, on your long-term memory codes).

To a large extent, then, cognitive load is reflected in task difficulty.

Here are some factors to consider when trying to assess cognitive load:

- **how much information there is**
- **how complex the information is**

How difficult it is to craft it into a tightly-bound cluster — that is, how hard it is for you to connect the information together and make sense of it.

- h**ow often you have to shift your focus**
  For example, if you are trying to do more than one task at a time (writing an essay and checking your Facebook page), then you are appreciably adding to your cognitive load; if you are trying to translate a text in a foreign language and have to keep diverting to the dictionary and a grammar book, then this will add considerably to the cognitive load (as compared to being able to translate without needing to check vocabulary and grammar points).

- **whether information already in focus has to be altered and how much time and effort is needed to change it**
  For example, in the math example above, you were performing calculations that 'updated' the number in focus.

There are basically two approaches to reducing cognitive load:

- break the task into smaller components

- practice the task or parts of the task until they are so easily retrieved from long-term memory that they are essentially automatic.

Thus, to reduce the cognitive load of dual essay-writing and Facebook-checking, you simply stop doing one activity and focus on the other. To reduce the cognitive load of trying to translate while constantly checking words and

grammar, you need to practice your vocabulary and grammar until they are readily accessible.

Practice, then, is at the heart of reducing cognitive load and functionally improving your working memory capacity.