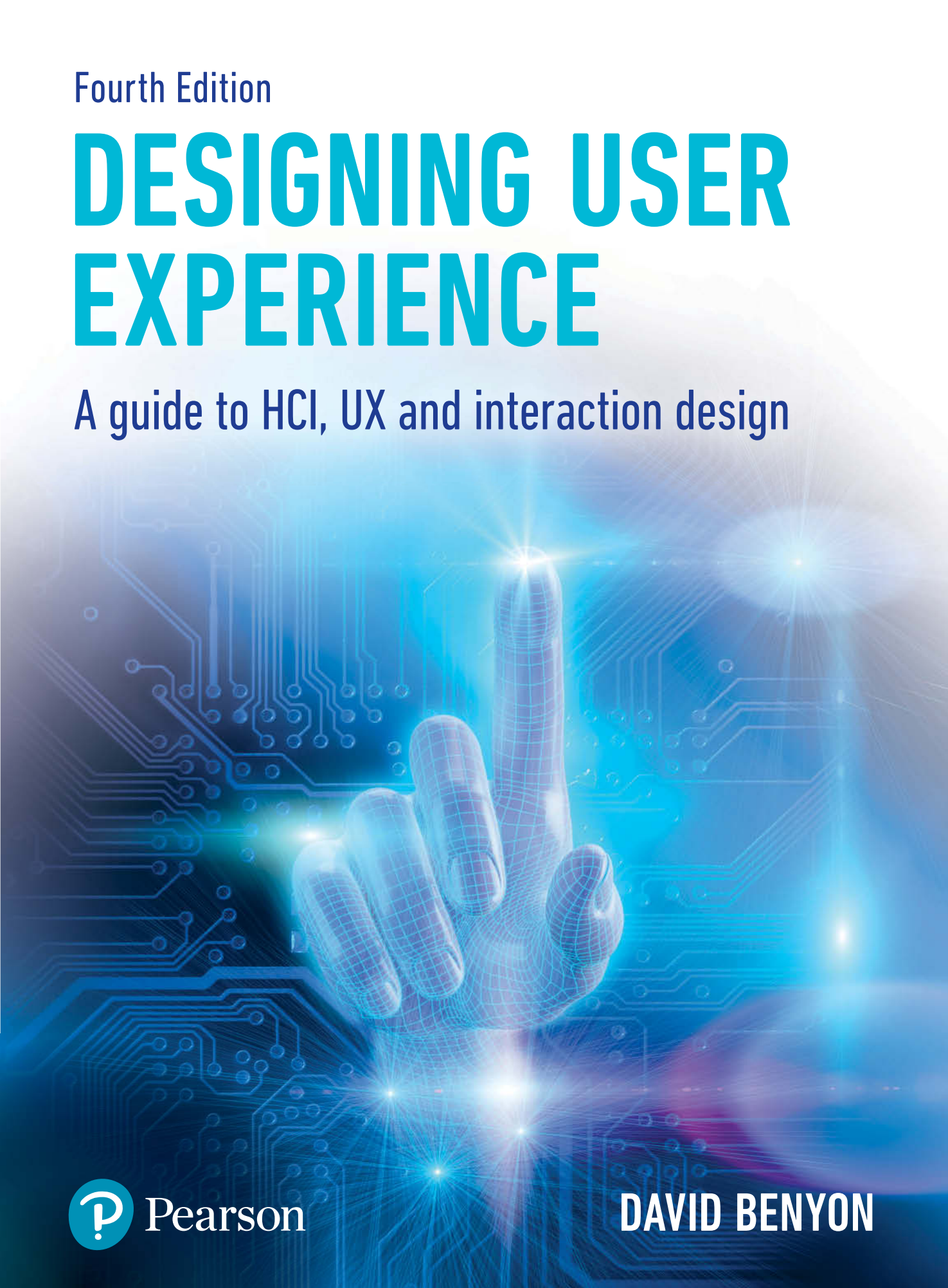
Fourth Edition

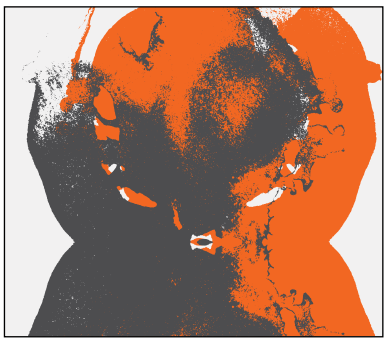
DESIGNING USER EXPERIENCE

A guide to HCI, UX and interaction design



Pearson

DAVID BENYON



Chapter 21
Memory and attention

Contents

Aims

Memory and attention are two of the key abilities that people possess. They work together to enable us to act in the world. For UX designers there are some key features of memory and attention that provide important background to their craft. Some useful design guidelines that arise from the study of memory and attention are presented in Chapter 12 and influence the design guidelines in Chapter 5. In this chapter we focus on the theoretical background.

After studying this chapter you should be able to:

- Describe the importance of memory and attention and their major components and processes
- Understand attention and awareness; situation awareness, attracting and holding attention
- Understand the characteristics of human error and mental workload and how they are measured.

21.1 Introduction

It is said that a goldfish has a memory that lasts only three seconds. Imagine this were true of you: everything would be new and fresh every three seconds. Of course, it would be impossible to live or function as a human being. This has been succinctly expressed by Blakemore (1988):

> without the capacity to remember and to learn, it is difficult to imagine what life would be like, whether it could be called living at all. Without memory, we would be servants of the moment, with nothing but our innate reflexes to help us deal with the world. There could be no language, no art, no science, no culture.

Memory is one of the main components of a psychological view of humans that aims to explain how we think and act. It is shown in Figure 21.1 along with awareness, motivation, affect and attention (and other unnamed functions). These subjects are covered in the next few chapters. In introducing memory we begin with a brief discussion of what memory is *not* and in doing so we hope to challenge a number of misconceptions.

There are several key issues about memory. First, it is not just a single, simple information store – it has a complex structure, one that is still argued over. There are differences between short-term (or working) and long-term memory. Short-term memory (STM) is very limited but is useful for holding such things as telephone numbers while we are calling them. In contrast, long-term memory (LTM) stores, fairly reliably, things such as our names and other biographical information, the word describing someone

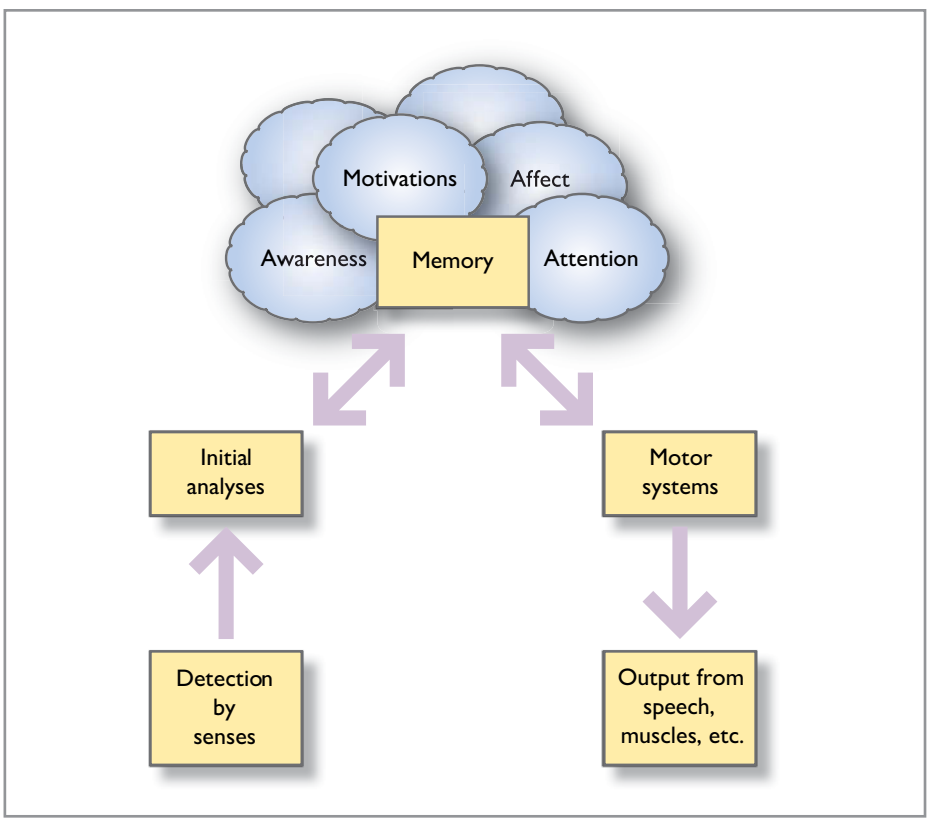


Figure 21.1 A more detailed information-processing paradigm

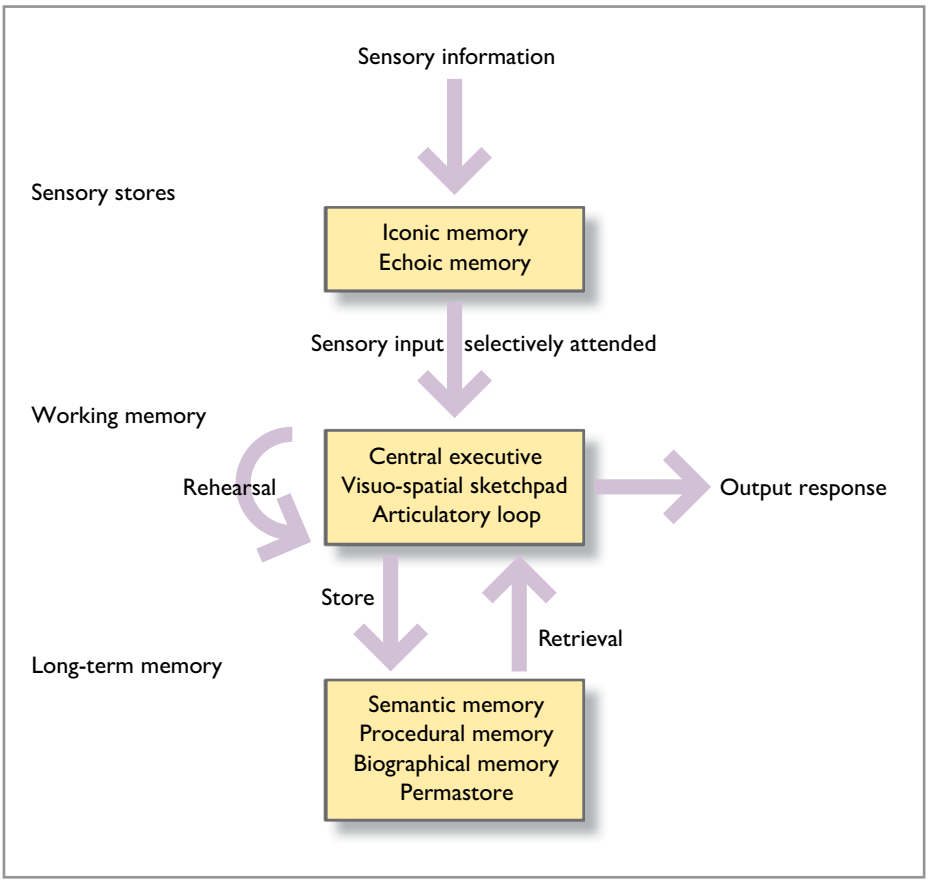


Figure 21.2 A schematic model of multi-store memory
(Source: After Atkinson and Shiffrin, 1968)

who has lost their memory, and how to operate a cash dispenser. This common-sense division reflects the most widely accepted structure of memory, the so-called multi-store model (Atkinson and Shiffrin, 1968), illustrated in Figure 21.2.

Second, memory is not a passive repository: it comprises a number of active processes. When we remember something we do not simply file it away to be retrieved whenever we wish. For example, we will see that memory is enhanced by deeper or richer processing of the material to be remembered. Third, memory is also affected by the very nature of the material to be remembered. Words, names, commands or images for that matter which are not particularly distinctive will tend to interfere with their subsequent recognition and recall. Game shows (and multiple-choice examination questions) rely on this lack of distinctiveness. A contestant may be asked:

> For £10,000 can you tell me: Bridgetown is the capital of which of the following?
>
> (a) Antigua
> (b) Barbados
> (c) Cuba
> (d) Dominica

As the islands are all located in the Caribbean they are not (for most contestants) particularly distinctive. However, this kind of problem can be overcome by means of **elaboration** (e.g. Anderson and Reder, 1979). Elaboration allows us to emphasize similarities and differences among the items.

Fourth, memory can also be seen as a **constructive process**. Bransford *et al*. (1972) were able to show that we construct and integrate information from, for example, individual sentences. In an experiment they presented a group of people with a series of thematically related sentences and then presented them with a second set of sentences, asking: 'Have you seen this sentence before?' They found that most people estimated that they had seen approximately 80 per cent of these sentences before. In fact, all of the sentences were new. Bransford *et al*. concluded that people are happy to say that they recognized sentences they have not seen providing they are consistent with the theme of the other sentences.

Finally, many researchers would now argue that memory cannot be meaningfully studied in isolation, as it necessarily underpins all other aspects of cognition (thinking). For example, object recognition relies on memory; the production and understanding of language relies on some form of internal lexicon (or dictionary); finding our way about town relies on an internal representation of the environment, sometimes described as a cognitive map (Tversky, 2003); the acquisition of skills often begins with internalizing and remembering instructions.

Memory is related to attention and these two are related to making mistakes, having accidents or doing things unintentionally. Memory, attention and error are also related to emotion. In this chapter we discuss the first three of these, devoting the next chapter to looking at emotion, or 'affect'.

21.2 Memory

Memory is usually divided into a set of **memory processes** and a number of different types of **memory store**. Table 21.1 is a summary of the main memory stores and their sub-components and associated processes. Figure 21.1 is an illustration of this multi-store model of memory (note the role of attention).

Memory stores: working memory

As we have already noted, working memory, first identified and named by Baddeley and Hitch (1974), is made up from three linked components: a **central executive**, a **visuo-spatial sketchpad** and an **articulatory loop** (also called the phonological loop). The central executive is involved in decision making, planning and related activities. It is also closely linked to managing our ability to perform more than one thing at a time (see the section below which discusses the role of attention). The articulatory or phonological loop can be thought of as behaving like a loop of audio tape. When we are trying to call an unfamiliar telephone number or repeating a phrase in a foreign language, we tend to repeat the string of numbers (or words) either out loud or silently to ourselves. This process is called **rehearsal**. When we are doing this we are making use of the articulatory loop, which can also account for our experience of the **inner voice**. The analogy of the audio tape is useful as it allows us to see that the articulatory loop is limited in both capacity and duration.

The visuo-spatial sketchpad (also called the scratchpad) is the visual and spatial information equivalent of the articulatory loop and has been linked to our **mind's eye**. We use our mind's eye to visualize a route through a town or building or for the mental rotation of figures (visualize a coin and then rotate it to see what is on the other side). The visuo-spatial sketchpad is also limited in capacity and duration unless refreshed by means of rehearsal. Finally, the capacity of working memory itself is approximately three or four items (e.g. MacGregor, 1987; LeCompte, 1999) where an item may be a

Table 21.1 A summary of the structure of memory

Main components	Key processes associated with this particular store
Sensory stores The **iconic** store (visual) and the **echoic** store (auditory) are temporary stores where information is held before it enters working memory.	The contents of these stores are transferred to working memory within a fraction of a second.
Working memory (WM) Working memory is made up of three key elements: the **central executive**, the **articulatory loop** and the **visuo-spatial sketchpad**. The central executive is involved in decision making, the articulatory loop holds auditory information and the visuo-spatial sketchpad, as the name suggests, holds visual information.	**Rehearsal** is the process of refreshing the contents of WM, such as repeating aloud a phone number. The contents of WM are said to **decay** (are lost/forgotten) if they are not rehearsed. Another way of forgetting from WM is **displacement** which is the process by which the current contents of WM are pushed out by new material.
Long-term memory Long-term memory comprises the following: • **Semantic** memory. This holds information related to meaning. • **Procedural** memory. This stores our knowledge of how to do things such as typing or driving. • **Episodic** and/or **autobiographical** memory. This may be one or two different forms of memory that are related to memories personal to an individual such as memories of birthdays, graduation or getting married. • **Permastore**. This has been suggested by Bahrick (1984) as the name for the part of LTM which lasts for our lifetime. It stores the things you never forget.	**Encoding** is the process by which information is stored in memory. **Retrieval** is the means by which memories are recovered from long-term storage. **Forgetting** is the name of a number of different possible processes by which we fail to recover information.

word or a phrase or an image. It should be noted that older textbooks and papers suggest that the limit of short-term memory is 7 ± 2 items, sometimes called the *magical number* 7: this is now known to be incorrect.

BOX 21.1

Distinguishing between short-term and working memory

In their multi-store model of memory, Atkinson and Shiffrin (1968) distinguish between short- and long-term memory (reflecting William James's primary and secondary memory division seventy years earlier). While the term short-term memory is still widely used, we have chosen to employ the term working memory (WM) instead. STM is usually characterized by a limited, temporary store for information before it is transferred to long-term memory, while WM is much more flexible and detailed in structure and function. Our use of WM instead of STM also better reflects our everyday experience.

Memory stores: long-term memory

Long-term memory has an effectively unlimited capacity and memories stored there may last as long as an individual's lifetime. The coding (the internal representation) of the information it holds is primarily semantic in nature, that is, it is stored in terms of its

meaning, for example knowledge of facts and the meaning of words (contrast this with the binary encoding of information in a computer). However, research has indicated that other forms of encoding are present too – for example, memories of music or the bark of a dog are encoded as auditory information; similarly, haptic (touch) encoding allows us to remember the feeling of silk and the sting of a cut. Finally, olfactory (smell) and gustatory (taste) encoding allow us to recognize and distinguish between the smell and taste of fresh and rotten food.

In addition to semantic memory, long-term memory includes other kinds of memories such as **episodic** or **autobiographical** memory (memory of our personal history, for example our first kiss, graduation day, the death of a parent) and **procedural memory** (e.g. the knowledge of how to ride a bike, type, play the euphonium). This neat three-way division of long-term memory into component parts – semantic, episodic and procedural – has been questioned by Cohen and Squire (1980), who argue that the real distinction is between 'knowing that' (declarative memory) and 'knowing how' (procedural memory), but in practice there is little between these two accounts.

Challenge 21.1

Contrast listing the components of a bicycle (e.g. frame, wheels, etc.) with knowing how to ride a bicycle (e.g. sitting on the saddle and pedalling) and with your memory of the first time you rode a bicycle (e.g. how old were you? What sort of day was it? Who else was there?). Which is hardest to describe?

How do we remember?

In everyday English, to remember means both to retrieve information ('I think her birthday is the 18th of June') and to store information in memory ('I'll remember that'). To remove this ambiguity we will use the terms **store** and **encode** to mean place in memory, and **retrieve** and **recall** to mean bring back from memory.

If what we want to store is not too complex (that is, it does not exceed the capacity of working memory), we will typically rehearse it, that is, repeat the string of words either aloud or using our inner voice. This is useful for remembering unfamiliar names or strings of numbers or words such as a foreign phrase, for example '*Dos cervezas, por favor*'. This technique exploits the articulatory loop of working memory. Similar strategies are used to remember, for a short time, the shape of an object or a set of directions. The capacity of working memory can effectively be enhanced by chunking the material to be remembered first. Chunking is the process by which we can organize material into meaningful groups (chunks). For example, an apparently random string of numbers such as 00441314551234 may defeat most people unless it is chunked. This particular number may be seen to be a telephone number made up from the code for international calls (0044), the area code for Edinburgh (131) and the prefix for Edinburgh Napier University (455), leaving only 1234 to remember. Thus the string of numbers has been reduced to four chunks.

So how do we remember things for longer periods? One answer is **elaboration**, which has been developed as an alternative view of memory in itself. The **levels of processing** (LoP) model proposed by Craik and Lockhart (1972) argues that rather than focusing on the structural, multi-store model of memory we should emphasize the memory processes involved. The LoP model recognizes that any given stimulus (piece of information) can be processed in a number of ways (or levels), ranging from the trivial

or shallow all the way through to a deep, semantic analysis. Superficial processing may involve the analysis of the stimulus's surface features such as its colour or shape; a deeper level of analysis may follow which may test for such things as whether the stimulus (e.g. cow) rhymes with the word 'hat'. The final and deepest level of analysis is the semantic, which considers the stimulus's meaning – does the word refer to a mammal?

Finally, we are able to retrieve stored information by way of **recall** and/or **recognition**. Recall is the process whereby individuals actively search their memories to retrieve a particular piece of information. Recognition involves searching our memory and then deciding whether the piece of information matches what we have in our memory stores.

How and why do we forget?

There are numerous theories of forgetting. However, before we discuss their strengths and weaknesses, we begin with another key distinction, namely the difference between **accessibility** and **availability**. Accessibility refers to whether or not we are able to retrieve information that has been stored in memory, while the availability of a memory depends on whether or not it was stored in memory. The metaphor of a library is often used to illustrate this difference. Imagine you are trying to find a specific book in a library. There are three possible outcomes:

1 You find the book (the memory is retrieved).
2 The book is not in the library (the memory is not available).
3 The book is in the library but has been misfiled (not accessible).

There is, of course, a fourth possibility, namely that someone else has borrowed the book, which is where the metaphor breaks down!

As we described earlier, information is transferred from working memory to long-term memory to be stored permanently, which means that availability is the main issue for working memory while accessibility is the main (potential) problem for long-term memory.

Challenge 21.2

Demonstrating recency and the serial order effect, the serial position curve is an elegant demonstration of the presence of (a) a short/long-term divide in memory and (b) the primacy and recency effects in forgetting. This is easily demonstrated. First, create a list of, say, 20–30 words. Present them in turn (read them or present them on a screen – try using PowerPoint) to a friend, noting the order in which the words were presented. At the end of the list, ask them to recall as many of the words as they can. Again note the order of the words. Repeat this process with another 6–10 people. Plot how many words presented first (in position 1) were recalled, then how many in positions 2, 3, 4, etc., up to the end of the list.

Forgetting from working memory

The first and perhaps oldest theory is **decay theory**, which argues that memory simply fades with time, a point which is particularly relevant to working memory which maintains memories for only thirty seconds or so without rehearsal. Another account

is **displacement theory**, which has also been developed to account for forgetting from working memory. As we have already seen, working memory is limited in capacity, so it follows that if we were to try to add another item or two to this memory, a corresponding number of items must be squeezed out.

Forgetting from long-term memory

We turn now to more widely respected theories of forgetting from long-term memory. Again psychology cannot supply us with one, simple, widely agreed view of how we forget from LTM. Instead there are a number of competing theories with varying amounts of supporting evidence. Early theories (Hebb, 1949) suggested that we forget from *disuse*. For example, we become less proficient in a foreign language learned at school if we never use it. In the 1950s it was suggested that forgetting from LTM may simply be a matter of decay. Perhaps memory engrams (= memory traces) simply fade with time, but except in cases of explicit neurological damage such as Alzheimer's disease, no evidence has been found to support this.

A more widely regarded account of forgetting is **interference theory**, which suggests that forgetting is more strongly influenced by what we have done before or after learning than the passage of time itself. Interference takes two forms: **retroactive interference** (RI) and **proactive interference** (PI). Retroactive interference, as the name suggests, works backwards, that is, newer learning interferes with earlier learning. Having been used to driving a manual-shift car, spending time on holiday driving an automatic may interfere with the way one drives after returning home. In contrast to RI, proactive interference may be seen in action in, for example, moving from word processor v1 to v2. Version 2 may have added new features and reorganized the presentation of menus. Having learned version 1 interferes with learning version 2. Thus earlier learning interferes with new learning. However, despite these and numerous other examples of PI and RI, there is surprisingly little outside the laboratory to support this theory.

Retrieval failure theory proposes that memories cannot be retrieved because we have not employed the correct retrieval cue. Recalling the earlier library metaphor, it is as if we have 'filed' the memory in the wrong place. The model is similar to the tip-of-the-tongue phenomenon (Box 21.2). All in all, many of these theories probably account for some forgetting from LTM.

BOX 21.2

The tip-of-the-tongue phenomenon

Researchers Brown and McNeill (1966) created a list of dictionary definitions of unfamiliar words and asked a group of people to provide words that matched them. Not surprisingly, not everyone was able to provide the missing word. However, of those people who could not, many were able to supply the word's first letter, or the number of syllables or even words that sounded like the missing word itself. Examples of the definitions are:

- Favouritism, especially governmental patronage extended to relatives (nepotism).
- The common cavity into which the various ducts of the body open in certain fish, birds and mammals (cloaca).

21.3 Attention

Attention is a pivotal human ability and is central to operating a machine, using a computer, driving to work or catching a train. Failures in attention are a frequently cited reason for accidents: car accidents have been attributed to the driver using their mobile phone while driving; aircraft have experienced 'controlled flight into terrain' (to use the official jargon) when the pilots have paid too much attention to the 'wrong' cockpit warning; and control room operators can be overwhelmed by the range and complexity of instruments to which they must attend. Clearly we need to be able to understand the mechanism of attention, its capabilities and limitations, and how to design to make the most of these abilities while minimizing its limitations.

Attention is an aspect of cognition that is particularly important in the design and operation of safety-critical interactive systems (ranging from the all too frequently quoted control room operator through to inspection tasks on mundane production lines). While there is no single agreed definition of attention, Solso (1995) defines it as 'the concentration of mental effort on sensory or mental events', which is typical of many definitions. The problem with definitions in many ways reflects how attention has been studied and what mental faculties researchers have included under the umbrella term of attention. However, the study of attention has been split between two basic forms: selective attention and divided attention. Selective (or focused) attention generally refers to whether or not we become aware of sensory information. Indeed, Cherry (1953) coined the term the cocktail party effect to illustrate this (Box 21.3).

BOX 21.3

The cocktail party effect

Cherry (1953), presumably while at a cocktail party, had noticed that we are able to focus our attention on the person we are talking to while filtering out everyone else's conversation. This principle is at the heart of the search for extra-terrestrial intelligence (SETI), which is selectively listening for alien radio signals against the background of natural radio signals.

Studies of selective attention have employed a *dichotic* listening approach. Typically, participants in such experiments are requested to shadow (repeat aloud) one of the two voices they will hear through a set of headphones. One voice will be played through the right headphone while another is played through the left – hence *dichotic*. In contrast to selective attention, divided attention recognizes that attention can be thought of in terms of mental resources (e.g. Kahneman, 1973; Pashler, 1998) that can in some sense be divided between tasks being performed simultaneously (commonly referred to as *multi-tasking*). For example, when watching television while holding a conversation, attention is being split between two tasks. Unless an individual is very well practised, the performance of two simultaneously executed tasks would be expected to be poorer than attending to just one at a time. Studies of *divided* attention might employ the same physical arrangements as above but ask the participant to attend (listen) to both voices and, say, press a button when they hear a keyword spoken in either channel.

BOX 21.4

The Stroop effect

Stroop (1935) showed that if a colour word such as 'green' is written in a conflicting colour such as red, people find it remarkably difficult to name the colour the word is written in. The reason is that reading is an automatic process which conflicts with the task of naming the colour of the 'ink' a word is written in. The Stroop effect has also been shown to apply to suitably organized numbers and words.

Try saying aloud the colour of the text – not the word itself:

Column 1	*Column 2*
RED	RED
GREEN	GREEN
BLUE	BLUE
RED	RED
GREEN	GREEN
RED	RED

You should find that saying the colour of each word in column 1 is slower and more prone to error owing to the meaning of the word itself. The word 'red' interferes with the colour (green) it is printed in and vice versa.

How attention works

To date, there have been a number of accounts or models of attention. The earliest date from the 1950s and are characterized by likening attention to a bottleneck. Later theories have concentrated on an allocation model that treats attention as a resource that can be spread (or allocated) across a number of different tasks. Other views of attention have concentrated on the automatic/controlled processing divide and on sequential/parallel processing. As in most aspects of psychology, there is no single account of attention; instead there is a mosaic of complementary views.

'Bottleneck' theories of attention

We begin with Donald Broadbent's single-channel theory of attention (Broadbent, 1958). He proposed that information arriving at the senses is stored in short-term memory before being filtered or selected as being of interest (or being discarded), which in practice means that we attend to one particular channel and ignore others. This information (channel) is then processed by a limited-capacity processor. On being processed, instructions may be sent to motor effectors (the muscles) to generate a response. The presence of short-term memory, acting as a temporary buffer, means that information that is not selected is not immediately discarded either. Figure 21.3 is an illustration of Broadbent's model. Broadbent realized that we might be able to attend to this information stored in the short-term memory but switching between two different channels of information would be inefficient. (A number of researchers have observed that Broadbent's thinking reflects the technology of his day, as in many ways this single-channel model of attention is similar to the conventional model of a computer's central processing unit (CPU), which also has a single channel and is a serial processing device – the von Neumann architecture.) This original single-channel model (sometimes referred to

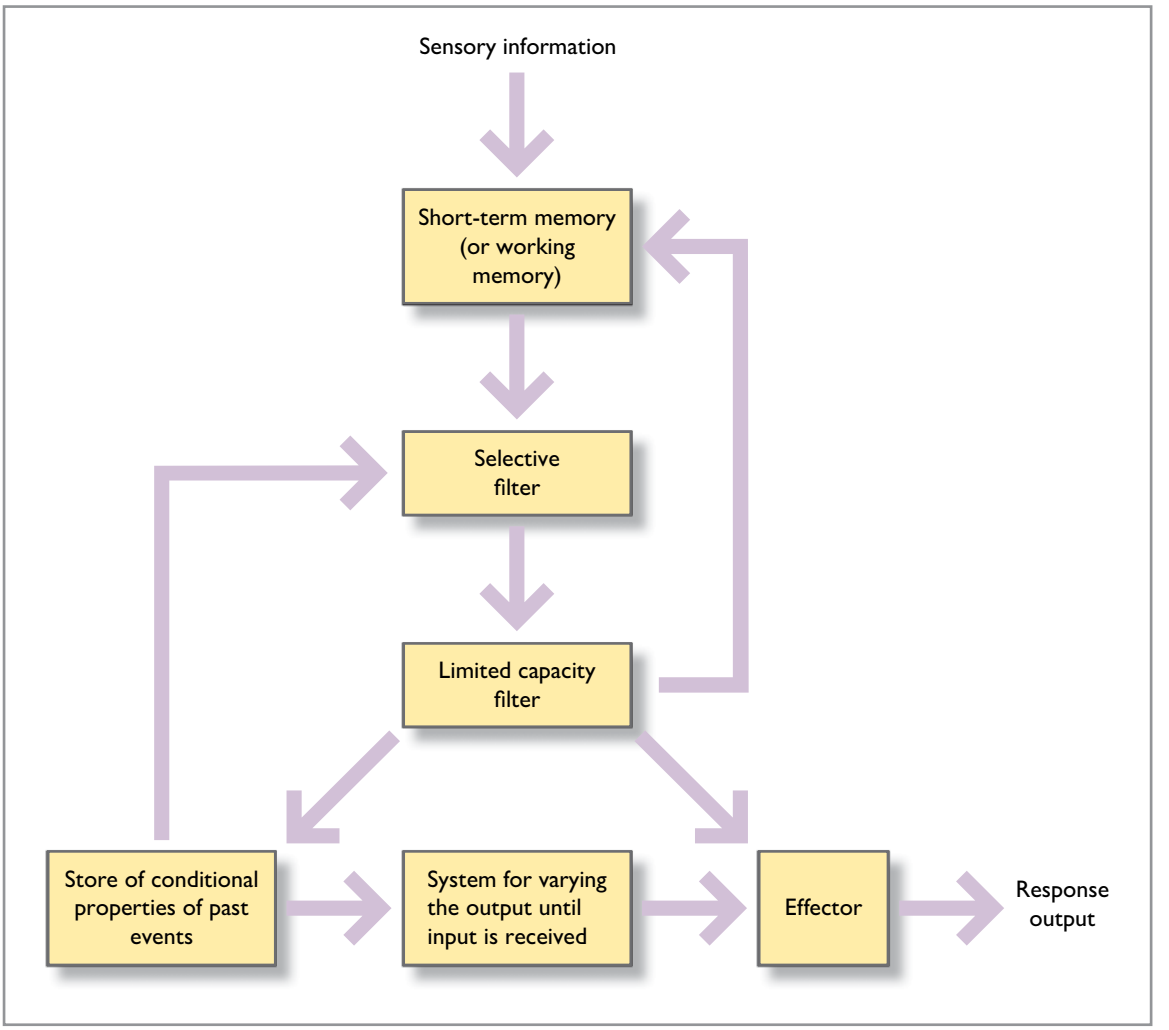


Figure 21.3 Broadbent's single-channel model of attention

as a bottleneck account of attention) was refined and developed by Broadbent's co-workers and others (Triesman, 1960; Deutsch and Deutsch, 1963; Norman, 1968) but remained broadly similar.

Triesman (1960) argued for the **attenuation** of the unattended channel, which is like turning down the volume of a signal, rather than an on–off switch. In Triesman's model, competing information is analyzed for its physical properties, and for sound, syllable pattern, grammatical structure and meaning, before being attended. The later Deutsch and Deutsch (1963) and Deutsch–Norman (Norman, 1968) models rejected Broadbent's early selection model, instead arguing for a later-selection filter/pertinence account. Selection (or filtering) occurs only after all of the sensory inputs have been analyzed. The major criticism of this family of single-channel models is their lack of flexibility, particularly in the face of a competing allocation model discussed below. Whether any single, general-purpose, limited-capacity processor can ever account for the complexity of selective attention has also been questioned. The reality of everyday divided attention presents even greater problems for such accounts. As we have just discussed, models of selective attention assume the existence of a limited-capacity filter capable of dealing with only one information channel at a time. However, this is at odds with both everyday experience and experimental evidence.

Attention as capacity allocation

Next, we briefly discuss an example of a group of models of attention which treats attention as a limited resource that is allocated to different processes. The best known is Kahneman's **capacity allocation** model (Kahneman, 1973). Kahneman argued that we have a limited amount of processing power at our disposal and whether or not we are able to carry out a task depends on how much of this capacity is applied to the task. Of course, some tasks require relatively little processing power and others may require more – perhaps more than we have available. This intuitively appealing account allows us to explain how we can divide our attention across a number of tasks depending upon how demanding they are and how experienced we are in executing them. However, there are a number of other variables that affect the ways in which we allocate this attentional capacity, including our state of arousal and what Kahneman describes as enduring dispositions, momentary intentions and the evaluation of the attentional demands. Enduring dispositions are described as the rules for allocating capacity that are not under voluntary control (e.g. hearing your name spoken), while momentary intentions are voluntary shifts in attention, such as responding to a particular signal. There is a further variable – how aroused we are. Arousal in this context may be thought of as how awake we are. Figure 21.4 illustrates the capacity allocation model in which we can see the limit capacity; the central processor has been replaced by an allocation policy component that governs which of the competing demands should receive attention. While Kahneman portrays attention as being more flexible and dynamic than the single-channel models, he is unable to describe how attention is channelled or focused. Similarly, he is unable to define the limits of what is meant by 'capacity'.

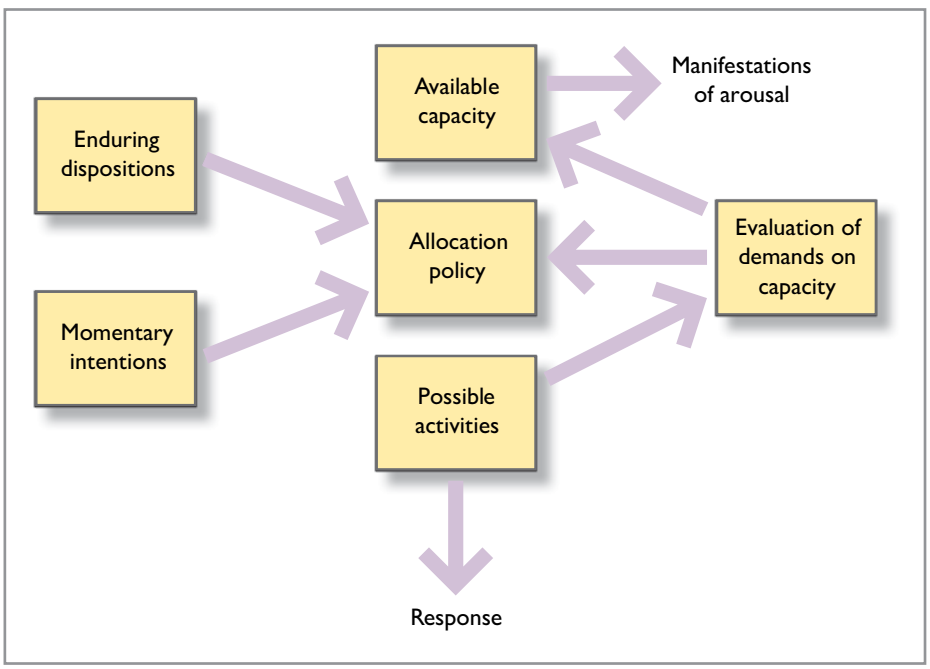


Figure 21.4 Kahneman's capacity allocation model

Automatic and controlled processing

In contrast to the foregoing models of attention, Schneider and Shiffrin (1977) observed that we are capable of both automatic and controlled information processing. We generally use automatic processing with tasks we find easy (and this, of course, is

dependent upon our expertise in this task) but use controlled processing on unfamiliar and difficult tasks.

Schneider and Shiffrin distinguish between controlled and automatic processing in terms of attention as follows. Controlled processing makes heavy demands on attention and is slow, limited in capacity and involves consciously directing attention towards a task. In contrast, automatic processing makes little or no demand on attention, is fast, unaffected by capacity limitations, unavoidable and difficult to modify, and is not subject to conscious awareness.

Schneider and Shiffrin found that if people are given practice at a task, they can perform it quickly and accurately, but their performance is resistant to change. An example of apparent automaticity in real life occurs when we learn to drive a car. At first, focused attention is required for each component of driving and any distraction can disrupt performance. Once we have learned to drive, and as we become more experienced, our ability to attend simultaneously to other things increases.

Moving from this very brief treatment of models of attention, we now consider how a wide range of internal and external factors can affect our ability to attend.

Factors affecting attention

Of the factors that affect our ability to pay attention to a task, stress is the most important. Stress is the effect of external and psychological stimuli and directly affects our level of arousal. Arousal is different from attention in that it refers to a general increase or decrease in perceptual and motor activity. For example, sexual arousal is typified by heightened levels of hormonal secretions, dilation of the pupils, increased blood flow and a whole range of mating behaviours.

Stressors (stimuli which cause stress) include such things as noise, light, vibration (e.g. flying through turbulence) and more psychological factors such as anxiety, fatigue, anger, threat, lack of sleep and fear (think about the days before an examination, for example). As long ago as 1908, Yerkes and Dodson found a relationship between performance of tasks and level of arousal. Figure 21.5 is an illustration of this relationship – the so-called Yerkes–Dodson law. There are two things to note about this relationship. First, for both simple and complex tasks there is an optimal level of

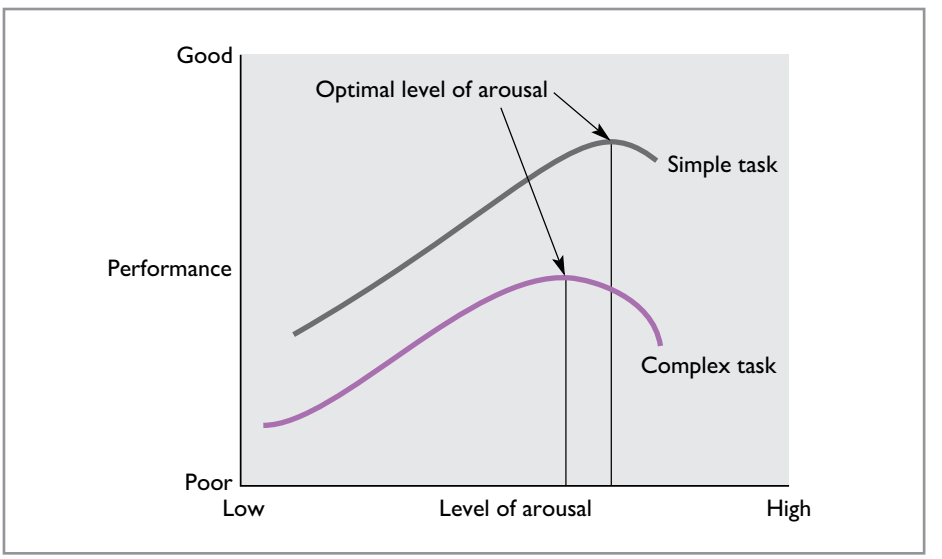


Figure 21.5 The Yerkes–Dodson law

arousal. As our level of arousal increases, our ability to execute a task increases until we reach a point when we are too aroused and our performance falls off sharply. Second, simple tasks are more resistant to increased levels of arousal than are complex tasks. The other aspect of this is the skill of the individual involved. A simple task to a highly skilled individual is likely to be seen as complex by a less skilled or able individual.

➔ Arousal is also important to the study of emotion, described in Chapter 22

Vigilance

Vigilance is a term applied to the execution of a task wherein an individual is required to monitor an instrument or situation for a signal. Perhaps the classic example of a vigilance task is being on watch on board a ship. During the Second World War mariners were required to be vigilant in scanning the horizon for enemy ships, submarines, aircraft or icefloes. Wartime aside, vigilance is still an important element of many jobs – consider the role of the operator of a luggage X-ray machine at an airport, or a safety inspector checking for cracks or loose fittings on a railway track.

BOX 21.5

Attention drivers!

Wikman *et al*. (1998) reported differences in the performance of inexperienced (novice) and experienced drivers when given a secondary task to perform while driving. The drivers were asked to do such things as changing a CD, operating the car radio or using a mobile phone. Unsurprisingly, the novice drivers were distracted more (allocated their attention less effectively) than the experienced drivers. Experienced drivers took their eyes off the road for less than three seconds, while novice drivers were found to weave across the road.

FURTHER THOUGHTS

In-car systems

The use of spoken messages in-car, particularly for satellite navigation (satnav) systems, is becoming commonplace. The challenge for the designers of these systems is:

(a) To attract the attention of the driver without distracting them
(b) To avoid habituation – that is, the driver learning to ignore the nagging voice.

The choice of voice is also critical. Honda has decided upon 'Midori', the name given to the voice of an unnamed bilingual Japanese actress whose voice is 'smooth as liqueur'. In contrast, Italian Range Rovers are equipped with a voice which is argumentative in tone, and Jaguar (the English motor manufacturer) retains British colloquialisms to reinforce its brand image. These brand images aside, manufacturers have found that drivers tend to listen to female voices more than male voices.

Other issues in in-car HCI concern the design of devices such as phones and satellite navigation systems that require complex operation and hence result in divided attention (Green, 2012).

Mental workload

Mental workload addresses issues such as how busy the user or operator is and the difficulty of the tasks assigned to them – will they be able to deal with an additional workload? A classic example of this occurred in the 1970s when a decision was made

to remove the third crew member from a flight team on board a medium to large passenger jet. The Federal Aviation Administration now requires measures of the mental workload on the crew prior to the certification of a new aircraft or new control system.

Turning now to design issues in respect of mental workload, the first observation is that a discussion of mental workload does not necessarily equate workload with overload. Indeed, the reverse is often true – just consider the potential consequences of operator/user boredom and fatigue (Wickens and Hollands, 2000, p. 470). There are a number of ways in which workload can be estimated, one of which is the NASA TLX scale (Table 21.2). This scale is a subjective rating procedure that provides an overall workload score based on a weighted average of ratings on six sub-scales.

Table 21.2 Measuring workload

Title	Endpoints	Description
Mental demand	Low/end	How much mental and perceptual activity was required (e.g. thinking, deciding, etc.)? Was the task easy or demanding, simple or complex?
Physical demand	Low/high	How much physical effort was required (e.g. pushing, pulling, etc.)? Was the task easy or demanding, slack or strenuous, restful or laborious?
Temporal demand	Low/high	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Perfect/failure	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/high	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration level	Low/high	How insecure, discouraged, irritated, stressed and annoyed as opposed to secure, gratified, content, relaxed and complacent did you feel during your task?

Source: Wickens, C.D. and Hollands, J.G. (2000) *Engineering Psychology and Human Performance* (3rd edn), © 2000. Printed and electronically reproduced by permission of Pearson Education, Inc., Upper Saddle River, New Jersey.

Visual search

Visual search, researched extensively by psychologists and ergonomists, refers to our ability to locate particular items in a visual scene. Participants in a visual search study, for example, may be required to locate a single letter in a block of miscellaneous characters. Try to find the letter 'F' in the matrix in Figure 21.6. This is a good example of how perception and attention overlap and an understanding of the issues involved in visual search can help in avoiding interactive systems such as that shown in Figure 21.7.

Research has revealed that there is no consistent visual search pattern which can be predicted in advance. Visual search cannot be presumed to be left to right, or clockwise rather than anti-clockwise, except to say that searching tends to be directed towards where the target is expected to be. However, visual attention will be drawn towards features which are large and bright and changing (e.g. flashing, which may be used for warnings). These visual features can be used to direct attention, particularly if they have a sudden onset (i.e. a light being switched on, or a car horn sounding). Megaw and Richardson (1979) found that physical organization can also have an effect on search

E E E E E E E E
E E E E E E E E
E E E E E E E E
E E E E E E E E
E E E E E E E E
E E E E E E E E
E E E E E E F E
E E E E E E E E

Figure 21.6 A matrix of letters

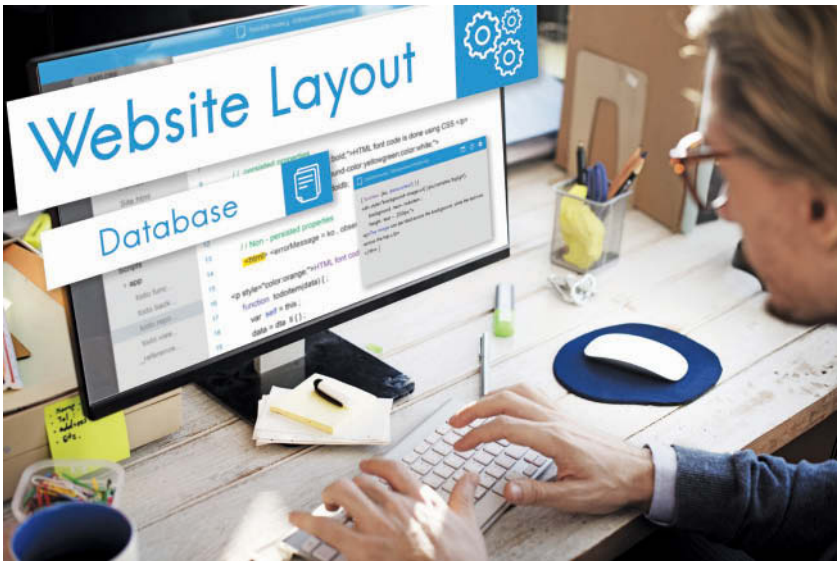


Figure 21.7 A practical example of the challenge of visual search

patterns. Displays or dials organized in rows tended to be scanned from left to right (just as in reading Western languages, but raising the question of cultural bias – would the same be true for those cultures in which people read from right to left or from top to bottom?). Parasuraman (1986) reported evidence of an **edge effect** wherein during supervisory tasks (that is, the routine scanning of dials and displays) operators tended to concentrate on the centre of the display panel and tended to ignore the periphery. As Wickens and Hollands (2000) note, research into visual scanning behaviour has yielded two broad conclusions. First, visual scanning reveals much about the internal expectancies that drive selective attention. Second, these insights are probably most useful in the area of diagnostics. Clearly those instruments most frequently watched are likely to be the most important to an operator's task. This should guide design decisions to place the instruments in prominent locations or to locate them adjacent to one another.

BOX 21.6

Just how long is it reasonable to wait?

It is generally accepted that delays of less than 0.1 second are taken to be effectively instantaneous, but delays of a second or two may be perceived by the user of an interactive system as being an interruption in the free flow of their interaction. Delays of more than 10 seconds present problems for people. Minimizing delay is important in the design of websites for which numerous, often contradictory, guidelines have been published. Here are two perfectly reasonable suggestions:

- The top of your page should be meaningful and fast.
- Simplify complex tables as they display more slowly.

Signal detection theory

It is late at night. You are asleep alone in your apartment. You are woken by a noise. What do you do? For many people the first thing is to wait and see (as it were) whether they hear the noise again. Here we are in the domain of signal detection theory (SDT) – was there really a signal (e.g. the sound of breaking glass by the local axe murderer) and if so, are we to act on it, or was it just the wind or a cat in the dustbin? SDT is applicable in any situation in which there are two different, non-overlapping states (i.e. signal and noise) that cannot be easily discriminated – for example, did a signal appear on the radar screen, did it move, has it changed size or shape? In such situations we are concerned with signals which must be detected, and in the process one of two responses may be produced – for example, 'I detected the presence of a signal, so I shall press the stop button', or 'I failed to see anything, so I shall continue to watch'. This may vary in importance from the trivial, such as recognizing that a job has been printed (the printer icon has disappeared from the application's status bar), through to the safety-critical, for instance a train driver spotting (or not) a stop light.

The following compelling examples of the importance of SDT have been identified by Wickens and Hollands (2000): the detection of a concealed weapon by an airport security guard; the identification of a malignant tumour on an X-ray plate by a radiologist; and a system malfunction detected by a nuclear plant supervisor. Their list goes on to include identifying critical incidents in the context of air traffic control, proof-reading, detecting lies from a polygraph (lie detector) and spotting hairline cracks in aircraft wings, among other things. SDT recognizes that an individual faced with such a situation can respond in one of four ways: in the presence of a signal, the operator may detect it (hit) or fail to detect it (miss); in the absence of a signal, the operator may correctly reject it (correct rejection) or incorrectly identify it (false alarm). This is illustrated in Table 21.3.

The probability of each response is typically calculated for a given situation and these figures are often quoted for both people and machines. So, a navigational aid on board an aircraft (for example, ground collision radar) might be quoted as producing false

Table 21.3 SDT decision table

Response	State	
	Signal	*Noise*
Yes	Hit	False alarm
No	Miss	Correct rejection

alarms (also called false positives) at a rate of less than 0.001 – one in a thousand. Similar figures are quoted as targets for medical screening operators (for instance, no more than 1 in 10,000 real instances of, say, breast cancer should be missed while 1 in 1000 false alarms are acceptable).

BOX 21.7

Transcript from Apollo XIII: barber-poles and the Moon

The Apollo flights to the Moon in the late 1960s and early 1970s are excellent examples of both user-centred design and brilliant and innovative ergonomic design. One of the innovations can be found in the design of the Apollo spacecraft, which used barber-poles to provide status information to the astronauts. A barber-pole is a striped bar signalling that a particular circuit or function is active (for example, the communication system – the talkback system), or, as the transcript below illustrates, measures of liquid helium and the state of the electrical systems. In the transcript we see that Jim Lovell reports to Mission Control that main bus *'B is barber poled and D is barber poled, helium 2, D is barber pole':*

55:55:35 – Lovell: *'Houston, we've had a problem. We've had a main B bus undervolt.'*
55:55:20 – Swigert: *'Okay, Houston, we've had a problem here.'*
…
55:57:40 – DC main bus B drops below 26.25 volts and continues to fall rapidly.
55:57:44 – Lovell: *'Okay. And we're looking at our service module RCS helium 1. We have – B is barber poled and D is barber poled, helium 2, D is barber pole, and secondary propellants, I have A and C barber pole.'* AC bus fails within 2 seconds.

Interestingly, the use of a barber-pole (Figure 21.8) can be found in modern operating systems. For example, the macOS system uses barber-poles.

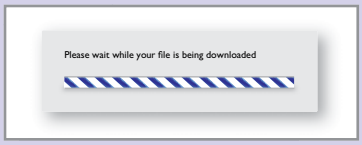


Figure 21.8 Barber-pole

21.4 Human error

Human error is studied in a wide variety of ways. Some researchers conduct laboratory investigations while others investigate the causes of major accidents after the event. A typical example of a laboratory study is that by Hull *et al.* (1988), who asked twenty-four ordinary men and women to wire an electric plug. They found that only five succeeded in doing so safely, despite the fact that twenty-three of the twenty-four had wired a plug in the previous twelve months. In analyzing the results of this study it was found that a number of factors contributed to these failures, including:

← Chapter 2 discussed mental models

- Failure to read the instructions
- Inability to formulate an appropriate mental model
- Failure of the plug designers to provide clear physical constraints on erroneous actions. This last point was regarded as the most significant.

Unhappily, error is an inescapable fact of life. Analysis of the causes of major accidents has found that human error is primarily responsible in 60–90 per cent of all major accidents (Rouse and Rouse, 1983; Reason, 1997). This figure is consistent with the

findings of commercial organizations – for example, Boeing, the aircraft manufacturer, estimates that 70 per cent of all 'commercial airplane hull-loss accidents' are attributable to human error.

Understanding action slips

Research conducted by Reason (1992) has given insight into everyday errors. In one study he asked thirty-six people to keep a diary of action slips (i.e. actions that have deviated from what they intended) for a period of four weeks. Analysis of the reported 433 slips revealed that storage failures (for example, repeating an action which has already been completed) were the most frequently reported. Figure 21.9 summarizes the key findings of this study and Table 21.4 describes each type of action slip (the miscellaneous errors are too diverse to discuss).

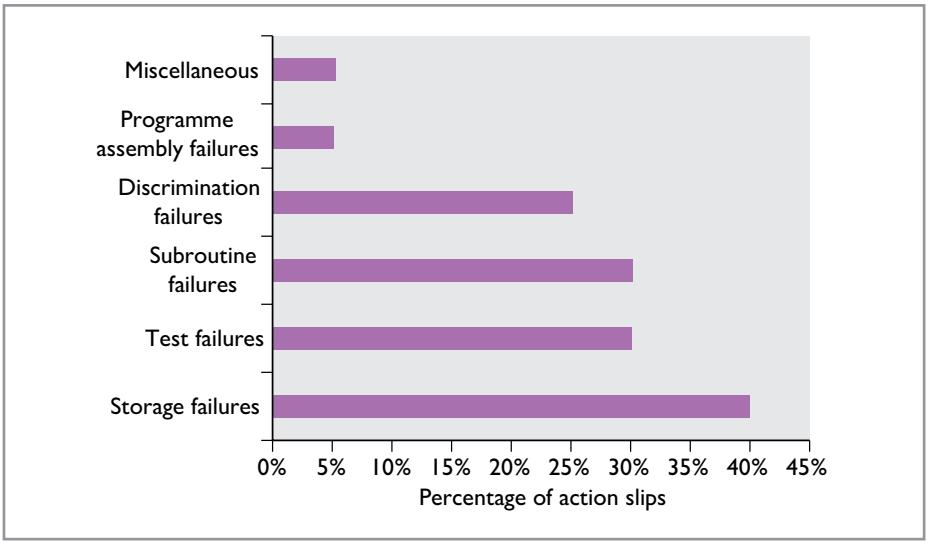


Figure 21.9 Five categories of action slips
(Source: After Reason, 1992, Fig. 15.24)

Table 21.4 Action slips

Type of action slip	Description
Storage failures	These were the most common and involved errors such as repeating an action which has already been completed, e.g. sending the same email twice.
Test failures	These refer to forgetting what the goal of the action was, owing to failing to monitor the execution of a series of actions, e.g. starting to compose an email and then forgetting to whom you are sending it.
Subroutine failures	These errors were due to omitting a step in the sequence of executing an action, e.g. sending an email and forgetting to include the attachment.
Discrimination failures	Failure to discriminate between two similar objects used in the execution of an action resulted in this category of error, e.g. intending to send an email and starting Word instead by mistake.
Programme assembly failures	This was the smallest category, accounting for only 5 per cent of the total. They involved incorrectly combining actions, e.g. saving the email and deleting the attachment instead of saving the attachment and deleting the email.

Each of these slips (and there are other classifications of errors, e.g. Smith *et al.*, 2012) presents challenges for the interactive systems designer. Some can be reduced or managed, others cannot.

Reducing action slips

Designers should design to minimize the chance of slips. For example, 'wizards' prompt people for, and help them recall, the steps which need to be undertaken to complete a task, such as installing a printer. In the sequence in Figure 21.10 the system prompts to obtain information to allow the operating system to install a printer. The advantage of this approach is that only relatively small amounts of information are required at any one time. It also has the advantage of an error correction system (i.e. use of the *Back* and *Next* steps).

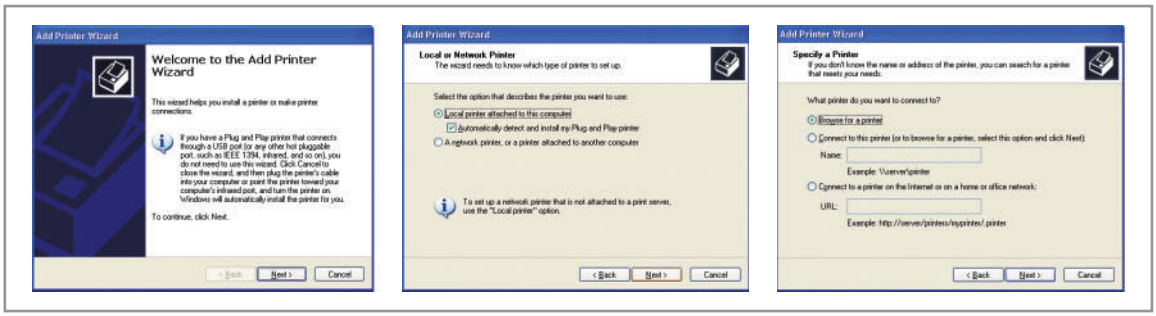


Figure 21.10 Using a Microsoft wizard to prompt a user to supply information one step at a time

One of the most demanding tasks in the work of an academic is marking coursework and examination scripts and tabulating the results without making mistakes. Figure 21.11 is a snapshot of a spreadsheet designed by Professor Jon Kerridge of the School of Computing, Edinburgh Napier University, to help reduce errors in this process. It is an example of good practice in this kind of manual tabulation of data as it employs a number of semi-automated checks (with corresponding error messages): in the column labelled *Checked* is a *note* indicating that an error message will appear if 'either the mark

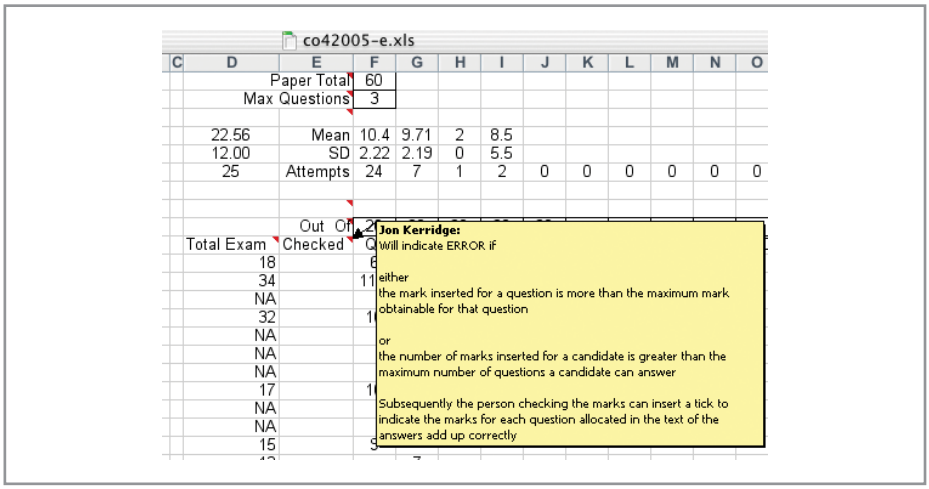


Figure 21.11 Avoiding mistakes with automated error checking

(Source: Courtesy of Jon Kerridge)

inserted for a question is more than the maximum mark obtainable for that question or . . .'. The author of the system has annotated the spreadsheet using *comments* and has used a series of *if statements* to check the inputted data. So, for example, marks should be entered for three questions only and an error is signalled if this number is exceeded.

Challenge 21.3

What is wrong with the error message in Figure 21.12? How would you reword it?

Figure 21.12 An unexpected error message

Summary and key points

We have seen that memory is divided into a number of stores, each of different size, make-up and purpose. Information arriving at the senses is held briefly in the sensory stores before moving on to working memory. Working memory (the modern equivalent of short-term memory) holds three or four items for up to thirty seconds unless rehearsed. Information may subsequently be stored in the long-term memory store with additional processing. The contents of long-term memory last a long time (minutes, hours, days, even years) and are held in several different types of memory, including memory for skills (procedural memory), semantic memory which holds the meaning of words, facts and knowledge generally, autobiographical memory which holds our personal experiences, and finally perma-store which holds information which literally may last a lifetime.

- In terms of design these limitations and capabilities translate into two important principles: the need to chunk material to reduce the load on working memory and the importance of designing for recognition rather than recall.
- Attention can be thought of in terms of being divided or selective. Divided attention refers to our ability to carry out more than one task at a time, though our ability to carry out multiple tasks also depends upon our skill (expertise) and the difficulty of the task. In contrast, selective attention is more concerned with focusing on particular tasks or things in the environment.
- It should come as no surprise that we make errors while using interactive devices. These errors have been classified and described by a range of researchers, storage failures being the most common. While all errors cannot be prevented, measures can be taken to minimize them, using devices such as wizards and automated error checking.

Exercises

1 As wizards can be used to prevent action slips, does it make good sense to use them for all dialogues with the system or application? When would you not use an error-preventing dialogue style such as wizards?

2 Compare and contrast how you would design a web browser for recall as compared with one for recognition. What are the key differences?

3 (Advanced) You are responsible for designing the control panel for a nuclear reactor. Operators have to monitor numerous alerts, alarms and readings which (thankfully) indicate normal operating conditions almost all the time. If and when an abnormal state is indicated, the operator must take remedial action immediately. Discuss how you would design the control panel to take into account the qualities of human attention.

4 (Advanced) How far (if at all) does psychological research into the mechanisms of human memory support the effective design of interactive systems? Give concrete examples.

Further reading

Reason, J. (1990) ***Human Error***. Cambridge University Press, Cambridge. *Perhaps a little dated now but a highly readable introduction to the study of error.*

Wickens, C.D. and Hollands, J.G. (2000) ***Engineering Psychology and Human Performance*** **(3rd edn)**. Prentice-Hall, Upper Saddle River, NJ. *One of the definitive texts on engineering psychology.*

Getting ahead

Baddeley, A. (1997) ***Human Memory: Theory and Practice.*** Psychology Press, Hove, Sussex. *An excellent introduction to human memory.*

Ericsson, K.A. and Smith, J. (eds) (1991) ***Towards a General Theory of Expertise.*** Cambridge University Press, Cambridge. *This is an interesting collection of chapters written by experts in expertise.*

Web links

The accompanying website has links to relevant websites. Go to
www.pearsoned.co.uk/benyon

Comments on challenges

Challenge 21.1

The most difficult of these to describe is usually the procedural knowledge of how to ride a bicycle. Most people find the other two aspects reasonably easy. Procedural knowledge is notoriously difficult to articulate – hence the advisability of having users show you how they perform particular tasks rather than try to tell you.

Challenge 21.2

The plot should resemble Figure 21.13. Words presented first, second, third . . . are recalled well, as are the last four or five words. The twin peaks represent recall from long-term (primacy) and working memory (recency), respectively. This is a well-known effect and explains why, when asking directions or instructions, we tend to remember the beginning and end but are vague about what was said in the middle.

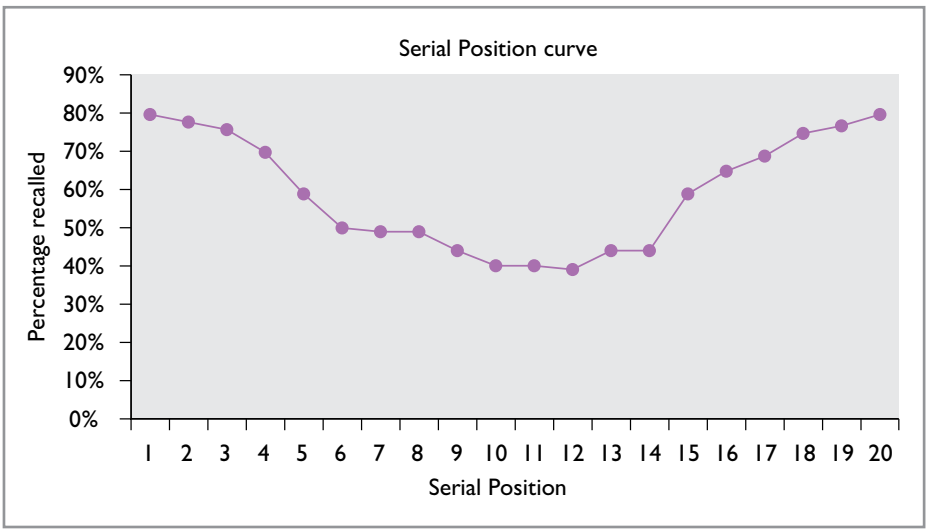


Figure 21.13

Challenge 21.3

This error message violates a number of good practice guidelines. 'Critical' sounds scary. 'Please contact support' is not helpful – what is the user supposed to do next? How does the user avoid the error in the future? See the guidelines above. Perhaps a better form of wording might be 'System problem encountered. Please restart application'.