2 Attention

Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.[1]

We start with attention and with William James. William James was one of the first modern experimental psychologists (although he himself did few experiments). He believed in studying the mind by using every tool he could find: logic, introspection, and experimentation. His goal was to describe the functions of psychological processes, and he succeeded admirably. His massive textbook *The Principles of Psychology*, first published in 1890, makes good reading today.

As our first quotation from James states, the effects of attention are known by everyone. We cannot fully appreciate all that takes place at any one time. When we concentrate fully on a book, noises in the environment fade from consciousness; when our thoughts wander in a lecture, we find ourselves unable to recall the speaker's message, although we were aware that he was speaking. We can generate examples endlessly. Let us try rather to determine more exactly the nature of attention and the quantitative bounds on its limitations. Then, we may be able to construct the type of logical process that is involved in limiting and controlling attention.

[1] From William James, *The Principles of Psychology*, Vol. 1. New York: Henry Holt and Co., 1890. Pages 403–404. (Republished by Dover, 1950.)

The Phenomena

William James deserves our study because he has carefully set forth his opinions about attention, letting us gain a deeper understanding of the many aspects of the phenomenon before we attempt to explain them. First, we follow James as he examines the problem of simultaneous attention. Granted that we have trouble attending to several simultaneous events, exactly how many items can we attend to at the same time? This problem is basic. We can immediately consider two different types of processes which might be responsible for the limitations of attention: one, a serial process which can do but one thing at a time, and the other, a parallel process which can do a number of things simultaneously, but with some upper limit to the total number of operations it can do at any one time.

A serial device requires some method of switching among the tasks that it is trying to do. If the switching can be performed with sufficient rapidity, there may be little loss in the information obtained from any given task. Parallel processes do not need to be switched from one task to another, but in turn, imply a good deal of complexity and redundancy in the mechanism that analyzes incoming information. The distinction between serial and parallel processes is of much current interest. Unfortunately, although James raises the issue, he gives us little help in resolving it, for he concludes that both possibilities might exist.[2]

*Attention**

WILLIAM JAMES

If, then, by the original question, how many ideas or things can we attend to at once, be meant how many entirely disconnected systems or processes of conception can go on simultaneously, the answer is, *not easily more than one, unless the processes are very habitual; but then two, or even three,* without very much oscillation of the attention. Where, however, the processes are less automatic, as in the story of Julius Caesar dictating four letters whilst he writes a fifth, there must be a rapid oscillation of the mind from one to

* *William James.* The Principles of Psychology. *Vol. 1. New York: Henry Holt and Co., 1890. Page 409. (Republished by Dover, 1950.)*

[2] In the readings that follow throughout the book, the original writings have been edited for continuity. Figures and footnotes have been renumbered to correspond to the numbers used here, and occasional sentences referring to sections of the author's paper that are not included in this book have been deleted. The symbol . . . marks a short deletion of material from a quotation, and the symbol ——— marks a lengthy deletion. My comments are printed in this style of type and lengthy quotations appear in this style of type.

the next, and no consequent gain of time. Within any one of the systems the parts may be numberless, but we attend to them collectively when we conceive the whole which they form.

The point seems to be that the number of things we can do depends on the difficulty of each task. A well-learned task, such as walking, takes little effort and does not impede us in our performance of another. A more difficult task such as walking along a high, narrow ledge requires more concentration and may completely impede our efforts to hold a conversation.

Given that the number of things to which we can attend at once is very limited, what is our perception of events to which we are not attending? The act of switching our attention to an event may both blur our perception of that event and cause confusion in our judgments of its temporal properties. These are the critical observations for the experimental investigation of attention, and they suggest a method of study. If attention is the result of a serial device, should there not be difficulty in determining the details and time sequence of events that occur during the absence of our attention? Again, let us return to the descriptions of William James before we go to more modern versions.[3]

When the things to be attended to are small sensations, and when the effort is to be exact in noting them, it is found that attention to one interferes a good deal with the perception of the other. A good deal of fine work has been done in this field, of which I must give some account.

It has long been noticed, when expectant attention is concentrated upon one of two sensations, that the other one is apt to be displaced from consciousness for a moment and to appear subsequently; although in reality the two may have been contemporaneous events. Thus, to use the stock example of the books, the surgeon would sometimes see the blood flow from the arm of the patient whom he was bleeding, *before* he saw the instrument penetrate the skin. Similarly the smith may see the sparks fly *before* he sees the hammer smite the iron, etc. There is thus a certain difficulty in perceiving the exact *data* of two impressions when they do not interest our attention equally, and when they are of a disparate sort.

Professor Exner . . . makes some noteworthy remarks about the way in which the attention must be *set* to catch the interval and the right order of the sensations, when the time is exceeding small. The point was to tell whether two signals were simultaneous or successive; and, if successive, which one of them came first.

The first way of attending which he found himself to fall into, was when the signals did not differ greatly—when, e.g., they were similar sounds heard each by a different ear. Here he lay in wait for the *first* signal, whichever it might be, and identified it the next moment in memory. The second, which could then always be known by default, was often not clearly distinguished in itself. When the time was too short, the first could not be isolated from the second at all.

The second way was to accommodate the attention for a certain *sort* of signal, and the next moment to become aware in memory of whether it came before or after its mate.

> *This way brings great uncertainty with it. The impression not prepared for comes to us in the memory more weak than the other, obscure as it were, badly fixed in time. We tend to take the subjectively stronger stimulus, that which we were intent upon, for the first, just as we are apt to take an objectively stronger stimulus to be the first. Still, it may happen otherwise. In the experiments from touch to sight it often seemed to me as if the impression for which the attention was* not *prepared were there already when the other came.*

Exner found himself employing this method oftenest when the impressions differed strongly.

— — —

THE EFFECTS OF ATTENTION

Its remote effects are too incalculable to be recorded. The practical and theoretical life of whole species, as well as of individual beings, results from the selection which the habitual direction of their attention involves.

— — —

Suffice it meanwhile that each of us literally *chooses,* by his ways of attending to things, what sort of a universe he shall appear to himself to inhabit.

The immediate effects of attention are to make us:

(a) perceive
(b) conceive
(c) distinguish
(d) remember

better than otherwise we could—both more successive things and each thing more clearly. It also

(e) shortens 'reaction-time.'

a and *b*. Most people would say that a sensation attended to becomes stronger than it otherwise would be. This point is, however, not quite plain, and has occasioned some discussion. From the strength or intensity of a sensation must be distinguished its clearness; and to increase *this* is, for some

[3] William James. *The Principles of Psychology, Op. cit.* Pages 409, 410, and 424–427.

psychologists, the utmost that attention can do. When the facts are surveyed, however, it must be admitted that to some extent the relative intensity of two sensations may be changed when one of them is attended to and the other not. Every artist knows how he can make a scene before his eyes appear warmer or colder in color, according to the way he sets his attention. If for warm, he soon begins to see the red color start out of everything; if for cold, the blue. Similarly in listening for certain notes in a chord, or overtones in a musical sound, the one we attend to sounds probably a little more loud as well as more emphatic than it did before. When we mentally break a series of monotonous strokes into a rhythm, by accentuating every second or third one, etc., the stroke on which the stress of attention is laid seems to become stronger as well as more emphatic. The increased visibility of optical after-images and of double images, which close attention brings about, can hardly be interpreted otherwise than as a real strengthening of the retinal sensations themselves. And this view is rendered particularly probable by the fact that an imagined visual object may, if attention be concentrated upon it long enough, acquire before the mind's eye almost the brilliancy of reality, and (in the case of certain exceptionally gifted observers) leave a negative after-image of itself when it passes away. Confident expectation of a certain intensity or quality of impression will often make us sensibly see or hear it in an object which really falls far short of it. In face of such facts it is rash to say that attention cannot make a sense-impression more intense.

But, on the other hand, the intensification which may be brought about seems never to lead the judgment astray. As we rightly perceive and name the same color under various lights, the same sound at various distances; so we seem to make an analogous sort of allowance for the varying amounts of attention with which objects are viewed; and whatever changes of feeling the attention may bring we charge, as it were, to the attention's account, and still perceive and conceive the object as the same.

> *A gray paper appears to us no lighter, the pendulum-beat of a clock no louder, no matter how much we increase the strain of our attention upon them. No one, by doing this, can make the gray paper look white, or the stroke of the pendulum sound like the blow of a strong hammer—everyone, on the contrary, feels the increase as that of his own conscious activity turned upon the thing.*

Were it otherwise, we should not be able to note *intensities* by attending to them. Weak impressions would, as Stumpf says, become stronger by the very fact of being observed.

> *I should not be able to observe faint sounds at all, but only such as appeared to me of maximal strength, or at least of a strength that increased with the amount of my observation. In reality, however, I can, with steadily increasing attention, follow a diminuendo perfectly well.*

The subject is one which would well repay exact experiment, if methods could be devised. Meanwhile there is no question whatever that attention augments the *clearness* of all that we perceive or conceive by its aid. But what is meant by clearness here?

c. Clearness, so far as attention produces it, *means distinction from other things* and *internal analysis* or *subdivision.* These are essentially products of intellectual *discrimination,* involving comparison, memory, and perception of various relations. The attention *per se* does not distinguish and analyze and relate. The most we can say is that it is a condition of our doing so. And as these processes are to be described later, the clearness they produce had better not be farther discussed here. The important point to notice here is that it is not attention's *immediate* fruit.

d. Whatever future conclusion we may reach as to this, we cannot deny that *an object once attended to will remain in the memory,* whilst one inattentively allowed to pass will leave no traces behind.

William James leaves us with a very complete description of the phenomenon of attention. He describes its variety, its nature, and its effects. Not much can be added to the overall picture, but all of the details must be filled in. We have just read that attention can alter the temporal order of our perceptions; why? We have read that attention affects retention, clarity, and reaction time; why? Even with a good description of the phenomenon we still know little of the mechanism.

Attention received much study following William James. Eighteen years later, in 1908, Edward B. Titchener wrote from his experimental laboratory at Cornell that one of the few things psychology could credit itself with achieving was the discovery of attention. Titchener credited the German introspectionist Wundt with the doctrine of attention, dating its inception as 1860, but, commenting further on the critical importance of attention, Titchener pointed out that "the discovery of attention did not result in any immediate triumph of the experimental method. It was something like the discovery of a hornet's nest: the first touch brought out a whole swarm of insistent problems." (Titchener, 1908, Chapter 5.)

Titchener tried hard to specify the attributes of attention by such "laws" as the law of prior entry in which he stated that, "the stimulus for which we are predisposed requires less time than a like stimulus, for which we are unprepared, to produce its full conscious effect. Or, in popular terms, the object of attention comes to consciousness more

quickly than the objects that we are not attending to." But Titchener was forced to conclude that, "although the discovery of a reliable measure of attention would appear to be one of the most important problems that await solution by the experimental psychology of the future" (Titchener quoting Külpe), "the discovery has not yet been made." All these statements apply today, some 70 years later.

The study of attention declined from the early years of the century until the 1950s. Then, in England, a group of researchers started a whole new series of studies, this time with a specific theoretical model of the attention process in mind. Several dramatic changes in scientific technique had occurred in those interim years. Mostly as a result of tremendous impetus to scientific work produced by the second world war, communication engineers had developed powerful electronic systems and analytical techniques, including the digital computer and related topics in automata and network theory.

Selective Attention and the Cocktail Party Problem

One of the first studies to come out of the new era of experimentation exemplifies many of the characteristics of the research. It was conducted by an Englishman, E. Colin Cherry, in an American laboratory at the Massachusetts Institute of Technology. The study was one of experimental psychology, but it was performed in the MIT Research Laboratory of Electronics and was published in a physics journal, the *Journal of the Acoustical Society of America*. Such interdisciplinary research is characteristic of modern psychology.

Cherry addressed himself to the problem of selective attention, or as he put it, "the cocktail party problem." The cocktail party serves as a fine example of selective attention. We stand in a crowded room with sounds and conversations all about us. Often the conversation to which we are trying to listen is not the one in which we are supposedly taking part. There are many different aspects of the cocktail party to interest psychologists. (We ignore the idea that it is a comfortable way in which to conduct research.) First, what is our selective ability? How are we able to select the one voice that interests us out of the many that surround us? Second, how much do we retain of the conversations to which we do not pay attention?

The first problem, selective attention, is not trivial. It implies a very complex analysis of the sounds that arrive at our ears—an analysis so complex that it cannot yet be performed by electronic devices. The second problem, the measure of our knowledge of rejected sources of speech, tells us how well the attention mechanism selects and rejects channels of information. These two problems, we will see, characterize the most recent research, for they are at the core of the phenomenon: we select what is relevant; we reject the rest.

The first complete theory of attention came in 1958 by Donald Broadbent, from the psychological laboratories in Cambridge, England. Broadbent developed a series of experiments, the most famous involving simultaneous memorization of two simultaneously presented sequences of digits.[6]

When Broadbent presented his subjects with three pairs of digits dichotically, so that one set of three digits read serially was heard at one ear at the same time that a second set of digits was heard at the other ear, he found surprising results. First, his subjects could barely recall 4 or 5 digits, whereas in more normal situations people have little trouble in remembering a string of 7 to 10 digits. Second, subjects preferred to organize their output by ears, rather than by the apparently more natural order, the order in which they heard the digits. That is, if the right ear has presented to it the digits 1, 7, 6 at the rate of one digit every half-second and the left ear the digits 8, 5, 2, the actual order of presentation of the digits is by the three pairs 1–8, 7–5, and 6–2. The preferred order of recall, however, is to give one ear's sequence first and then the other's: 1, 7, 6 and then 8, 5, 2. Usually, subjects get all the digits correct from the first ear, but make errors in the other sequence.

What can we make of these results? Broadbent concluded that they illustrated the properties of selective attention; selection was made on the basis of the physical channels by which the digits were presented. After considering a large set of experimental findings of various sorts, including the dichotically presented digits and Cherry's results, Broadbent put together a theoretical structure which he felt represented the underlying processes.

[6] A glossary of terminology might be appropriate here. In experiments involving auditory information presented to the two ears, it is necessary to distinguish among the various ways in which that information might be presented.

Consider two sources of sounds, *A* and *B*, which we wish to present simultaneously to a listener. If *A* and *B* are both presented to one ear only (through an earphone) we say that the presentation is *monaural*. If *A* and *B* are mixed together and then presented to both ears, so that both ears hear exactly the same material, we say the presentation is *binaural*. If the two channels are fed into separate ears, so that the left ear hears only *A* and the right ear only *B*, we say the presentation is *dichotic*. Finally, if we feed *A* through one loudspeaker and *B* through another placed nearby in such a way as to recreate the sound patterns resulting when two persons might simultaneously take the part of *A* and *B*, we say the presentation is *stereophonic*. Similar distinctions can be made for visual material: *monoptic, dioptic, stereoscopic*.

The Filter Model

Broadbent was attempting to piece together a model of human capability that would account for a wide variety of data, not just those from experiments in attention. Basically, Broadbent suggested that the limit to our ability to perceive competing messages is perceptual; we are able to analyze and identify only a limited amount of the information that arrives at our sensory inputs. He proposed that the brain contains a "selective filter" that can be "tuned" to accept the desired message and reject all others. The filter thus manages to block undesired inputs, reducing the processing load on the perceptual system. In the following excerpt from his book, Broadbent summarizes the model and tries to show how it is compatible with the evidence from a rather wide variety of psychological tasks. His model is important, for it shaped the direction for further research in attention. It is appropriate that we review it now, before we examine other, later experiments.

*Perception and Communication**

DONALD E. BROADBENT

SUMMARY OF PRINCIPLES

(a) A nervous system acts to some extent as a single communication channel, so that it is meaningful to regard it as having a limited capacity.

(b) A selective operation is performed upon the input to this channel, the operation taking the form of selecting information from all sensory events having some feature in common. Physical features identified as able to act as a basis for this selection include the intensity, pitch, and spatial localization of sounds.

(c) The selection is not completely random, and the probability of a particular class of events being selected is increased by certain properties of the events and by certain states of the organism.

(d) Properties of the events which increase the probability of the information, conveyed by them, passing the limited capacity channel include the following: physical intensity, time since the last information from that class of events entered the limited capacity channel, high frequency of sounds as opposed to low.

— — —

(h) Incoming information may be held in a temporary store at a stage previous to the limited capacity channels: it will then pass through the channel when the class of events to which it belongs is next selected. The maximum time of storage possible in this way is of the order of seconds.

(i) To evade the limitations of (h) it is possible for information to return to temporary store after passage through the limited capacity channel: this provides storage of unlimited time at the cost of reducing the capacity of the channel still further and possibly to zero. (Long-term storage does not affect the capacity of the channel, but rather is the means for adjusting the internal coding to the probabilities of external events; so that the limit on the channel is an informational one and not simply one of a number of simultaneous stimuli.)

(j) A shift of the selective process from one class of events to another takes a time which is not negligible compared with the minimum time spent on any one class.

— — —

An information-flow diagram incorporating the more probable principles is shown in Fig. 2.1.

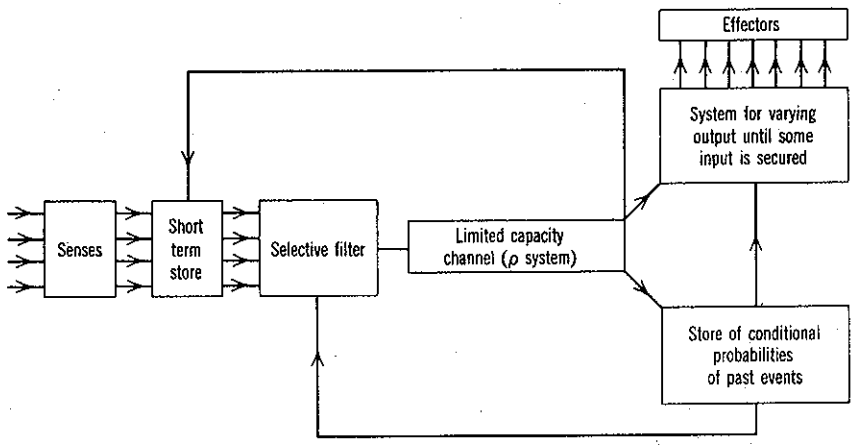

Fig. 2.1. A tentative information-flow diagram for the organism, as conceived at the present time.

* *Donald E. Broadbent.* Perception and Communication. *London: Pergamon Press, 1958. Pages 297–300. Copyright © 1958 by D. E. Broadbent. With permission of author and publisher.*

Major Implications of These Principles

Now that these principles are stated thus boldly, it may be urged that they are not particularly surprising. Do we not know that attention is limited, that noises distract us, that we consciously rehearse any matter which must be remembered for a short period, and so on? What gain is there from putting these everyday experiences into this stilted language? They have already been formulated by the classical introspective psychologists: why is time and effort wasted on rephrasing them?

These are two answers to this objection. First, it is indeed true that the principles lead to no prediction which is contrary to everyday observation. It would be a poor set of scientific principles which did do so: though it is not uncommon for psychologists to feel that they ought to contradict common beliefs about behaviour.

But secondly, as a matter of history it is not true that these principles are obvious nor that they were adequately formulated by classical psychologists. It is quite possible to say that mentalistic statements are consistent with them: to say that our limited capacity single channel is to be equated with the unitary attention of the introspectionists. Indeed, the writer believes that the one is simply a more exact version of the other. But a view of attention as unitary might also be taken to mean that a man cannot perform two tasks at once: which experimental evidence shows to be untrue. A view of noise as distracting might be taken to mean that card-sorting or mental arithmetic will be worse performed in noise, which they are not.

Broadbent provides us with a complete structure for selection and attention which agrees with intuition and with the facts known at that time (1958). Basically the hypothesis can be summarized by saying that a limited-capacity processor deals with but one channel of information at a time. The human selects among the various sources of information impinging on his sense organs on the basis of the physical characteristics of the information. When necessary, he switches attention among the various input sources. A short-term memory system prevents loss of information about the immediate past history of the unselected channels.

The Filter Model Fails and Is Modified

Broadbent's theory indicated a specific theoretical structure that, in turn, had strong implications on how people behave. Some of these implications were tested and proved wrong. One method of probing the theoretical structure is by means of the question raised earlier with Cherry's paper: how are we able to switch our attention among inputs if we are unaware of the content of unselected inputs?

A crucial experiment on this point was done by two undergraduates at Oxford University, Gray and Wedderburn (1960). They rejected the idea that attention was based on the physical characteristics of sensory channels and suggested that psychological attributes played an important role in selection. Their experiments were simple in concept, but the results were devastating to Broadbent's theory. Suppose we listened to one word so divided that different syllables of the word are presented alternately to different ears. At the same time, another word is decomposed in a similar fashion and presented to the complementary ear. Would not the attention switch from ear to ear and thus re-create each word correctly, rather than stick to one physical channel and get a nonsensical mixture of syllables? If this is so—and Gray and Wedderburn showed that it is—the attention mechanism must be able to extract the meaning of information from the two ears in order to know which to choose. But Broadbent's system (Fig. 2.1) requires attention to be switched at an early stage in the processing of sensory information, much before any of the meaning has been extracted.

The distinction between selection on the basis of physical or sensory analyses and selection on the basis of a more thorough analysis of meaning is very important. Broadbent's theoretical scheme makes good sense; we would not like to discard the overall structure. Yet it is clear from even the simple points raised by Gray and Wedderburn that something is wrong. A much more complete discussion of this problem was put forth by Anne Treisman in her doctoral dissertation at Oxford University in 1961. Treisman examined the role of verbal and linguistic features on her subjects' ability to select one message from among several. In particular, she hoped to test Broadbent's suggestion that "classes" of words may behave in the same way as sensory channels do by presenting messages in different languages. Certainly a difference in language is an extreme case of a verbal distinction between two messages without any general differences in physical characteristics. In addition, she examined the role of familiarity, redundancy, meaningfulness, and similarity.

Treisman evaluated Broadbent's theory of a selective filter by continuing and elaborating on Cherry's experimental methods. The technique of this research is to keep the subject occupied performing a shadowing task while various types of competing messages and signals are presented to him. It has been found that if the subject manages to keep up the shadowing task, gross physical changes of the nonshadowed message are noticed (change from a man's voice to a woman's), simple changes are not noticed (changes in the language of the nonshadowed message), and "important" words on the nonshadowed ear are often noticed (the subject's name or material that would be relevant within the context of the shadowed material).

The problem discussed for the rest of the chapter deals with the

way in which our ability to perceive material is limited. The selective filter proposed by Broadbent works to minimize the amount of processing that must be performed by more complex, higher-level processes. Broadbent evidently had in mind an ascending chain of complexity with some central mechanism performing the final analysis on incoming information. In this scheme one wants to eliminate irrelevant messages from the central mechanism. The problem, as we shall see, is that the properties required of the filter became so complex that the filter seemed to be almost as complicated as the final mechanism it was attempting to serve.

In her experiments, Treisman studied selective attention to one of two competing messages, presented binaurally. The messages were created by varying a number of different attributes. The irrelevant material was sometimes read with the same voice as the relevant material (female) and sometimes in a different voice (male). The nature of the irrelevant material also varied, sometimes being a technical discussion, other times passages from novels, and sometimes passages from the same novel as the relevant channel. Finally, the language of the irrelevant passage varied from English to Latin, French, German, and Czech (with a deliberate English accent), to English played backward over the tape recorder, and even to a French translation of the English shadowed message. The subjects' job was to shadow the relevant channel and ignore entirely the irrelevant material. As we can guess from Cherry's study, this is an easy task when the voices and materials used for the two messages are different, but a difficult task when the same voices and similar materials are used.

The aim of these studies, remember, was to determine at what level the selection of relevant from irrelevant material was made. Broadbent postulated selection almost entirely based on sensory features. Treisman hoped that by using a variety of irrelevant material she could distinguish between the relative effects of cues based entirely on sensory features and cues that required the determination of familiarity and meaning. Her experiments showed that sensory cues alone were not sufficient. She summarized her results and put forth a hypothesis to account for the types of errors made by her subjects (Ss).

*Verbal Cues in Selective Attention**

ANNE M. TREISMAN

(1) It was shown that a difference in voice (male vs. female) and a difference in language have quite different effects on tasks requiring selective

* *Anne M. Treisman. Verbal cues, language and meaning in selective attention.* American Journal of Psychology. *1964, 77, 215–216.*

response to one of two messages. The dlifference in voice allows the irrelevant message to be rejected much more efficiently, and this probably takes place at an earlier stage in the perceptual analysis of inputs. (2) When two messages share the same general physical characteristics, a difference in language allows some selection between them; however this seems more similar to selection between two English messages on the basis of subject matter than to the efficient performance obtained with different voices. Thus, complete rejection of one language as such appears to be impossible. (3) Phonetic cues make an unknown foreign language less distracting than a message which is phonetically similar to English (the Czech nonsense), and allow the *Ss* to name the irrelevant foreign language. Reversed speech however, causes a relatively high degree of interference. (4) The *Ss*' knowledge of the language affects the amount of interference it produces. In most cases, however, little of the content of the rejected message can be reported. Many *Ss* failed to notice that the Czech was not normal English and not all the *Ss*, even among those who knew the language fluently, realize that a rejected French message was a translation of the selected English one. While similarity of the languages at the phonetic level makes selection more difficult (as with the Czech), similarity of meaning (with the French translation) produced no general increase in interference. (5) Finally, when both messages are in the same language and same voice, selection is based chiefly on transitional probabilities between words and its efficiency varies with the degree of contextual constraint within both the selected and irrelevant messages. This is the only condition in which a considerable number of overt intrusions from the irrelevant message is made.

These findings are relevant to the following problems: (1) Are both messages fully analyzed before one is selected to determine the responses? (2) If not, at what stage in the analysis is one of the two discarded? (3) What determines selection and switching between messages? How does this depend on the familiarity and phonetic structure of the languages and on the transitional probabilities between words?

(1) Analysis of irrelevant message. For the following reasons, it seems unlikely that both messages are always fully analyzed and that selection takes place only for the overt responses: (a) less than half the *Ss* recognized either the French translation or the Czech nonsense, although this would have added no more load to response or memory than noticing the male voice. Those who did identify these messages may have done so by switching to the irrelevant message.

— — —

(2) Stage at which selection is made. The question then arises: at what stage is one of the two messages rejected from further analysis? It does not seem possible to reject or filter out irrelevant messages which differ only in verbal characteristics in the same efficient way as those which differ in general physical features.

— — —

If this early selection were possible between messages in the same voice but different languages, known and unknown languages should cause equal interference and both should be easier to reject than a message in the same language as the selected message. Neither of these predictions was confirmed by the results.

The results thus suggest that features of incoming messages are analyzed successively by the nervous system, starting with general physical features and proceeding to the identification of words and meaning, and that selection between messages in the same voice, intensity, and localization takes place during, rather than before or after, the analysis which results in the identification of their verbal content. It seems to be at this stage that the information-handling capacity becomes limited and can handle only one input at a time, either keeping to one message where possible, or switching between the two. Broadbent's suggestion that one may think of classes of words as constituting separate "input channels" which can be rejected, as such, is not supported by these results.

(3) Factors determining selection and switching between messages. What, then, determines selection and switching when both messages arrive on one input channel? The irrelevant messages seem to fall into two main classes: (*a*) those which the *Ss* potentially could identify; and (*b*) those whose verbal content they could never identify at all, because the language was unknown or the tape played backwards.

(*a*) When the irrelevant message was in a known foreign language or in English, the interference often took the form of making the *Ss* shift their attention to the wrong message and lose the correct message altogether for a time. The results with the statistical approximations suggest that the *Ss* repeat the correct message until its transitional probabilities fall to a low value. Differences in the competing messages do not affect the point at which they switch, but do affect their subsequent performance. Having switched, the *Ss* have two decisions to make: whether to repeat aloud what they hear, or to switch back. The instructions were to repeat as much as possible, rather than remaining silent when in doubt. This would encourage the *Ss* to make overt responses until they switched back, except when these were obvious errors, such as words in a different language.

We can now try to interpret these experiments and hypothesize what must be happening. To summarize Treisman's results once more, she found a graded effect on the ability of her subjects to reject an irrelevant message. When there was a distinct physical difference between relevant and irrelevant channels, subjects had no difficulty in shadowing one without being bothered by the other. When the messages had similar physical characteristics but belonged to different languages, they were much less successful. The better the subjects knew the irrelevant languages, the more it interfered. The most difficult task was to maintain shadowing one message when both were read in the same language and spoken with the same voice.

To explain these results, Treisman postulates an analytical mechanism that performs a series of tests on incoming messages. The first tests distinguish among the inputs on the basis of sensory or physical cues; later tests distinguish among syllabic patterns, specific sounds, individual words and, finally, grammatical structure and meaning. The sequence of tests can be thought of as a tree, with incoming sensory information starting at the bottom and working its way up to a unique end point, with tests at each spot where there is a choice of branches that might be taken. Moreover, Treisman suggests that the tests be flexible, so that if a particular word is expected, all the tests relevant to selecting that word might be prebiased or presensitized toward it. Thus, analysis is much simplified for items that are expected to occur.

If channels have physical distinctions, then at a very early stage of testing it will be possible to separate one from the other. This is done by attenuating the irrelevant channel so that it no longer interferes with the later testing procedure. Thus, words that appear on the irrelevant channel will be severely attenuated because they fail the physical test. If, however, a word on the irrelevant channel fits within the context of the material which has just been analyzed, it might very well be detected because the sensitization of each test toward the expected event would tend to cancel the effects of the attenuation of the irrelevant channel. Note, by the way, that by this model we ought to make mistakes, often claiming to have heard an item that was not actually presented. These false recognitions are a result of the lowered decision criteria for the tests relevant to expected events. Thus, although this presensitization makes detection of the correct event more likely, it also increases the likelihood that similar sounding items will pass the test incorrectly.

Treisman moves us one step further in the specification of the level at which attention becomes selective. She suggests that all incoming signals are analyzed to some extent by a sequence of operations. Signals are separated from one another by their physical features when that is possible and by their grammatical features when that becomes necessary. Grammatical information is used to bias or sensitize the criterion for identifying certain signals. Thus, in the middle of a sentence we might expect a certain grammatical class of words to occur, so we presensitize our analytical mechanisms for the possibility. This explains why we are sometimes able to pick out material presented in competing messages when that material appears to be relevant to the context to which we are primarily attending. The details of this procedure are left unexplained. Obviously, it implies that all signals, whether thought to be relevant or irrelevant, must receive a good deal of analysis, if only so that they may be discarded with some certainty. The infrequent relevant

signal from interfering channels would never get through the attention mechanism had it not received some identification before the final selection process took place.

Early versus Late Selection

Our neat, pretty picture of an attention mechanism has disappeared. William James described how one stream of thought was separated from all possible ones. It appeared that this was an automatic process, requiring but little mental effort and resulting in a major simplification of duties for higher level processes. But now the story is not so simple. Evidently, we choose among incoming channels of information on the basis of rather complex analyses of the incoming signals. We had assumed that the purpose of selective attention was to allow a central mechanism to concentrate its efforts on analyzing and responding to one problem at a time. But now it appears that selection among alternative channels itself requires complex processing. What have we gained by the concept of a selective mechanism?

The main argument against Treisman's explanation concerns the complexity of the operations she proposes. Presumably, the selective nature of attention serves the purpose of reducing the amount of analysis that some central device must perform on incoming information by feeding it only one signal at a time. This concept is fine as long as the selection can be performed by looking for simple, physical differences among the signals. It is very easy to conceive of a system that separates a man's voice from a woman's or a voice on the left from one on the right. As soon as we are forced to use the meaning of signals to aid us in our selection, the problem becomes very complex. The meaning of the peculiar sound waveform that comprises a word cannot be determined without extensive analysis of the signal, an analysis that must use information stored in memory. At this point the whole purpose of a selective mechanism seems to disappear, for if we need to extract the meaning of all incoming signals to determine what to attend to, how does the selectivity help us?

An alternative theory of selective attention requires us to move the selection mechanism back a bit. That is, suppose we admit that every incoming signal does indeed find its match in memory and receive a simple analysis for its meaning. Let the selective attention mechanism take over from there. We still save some work because there is a lot more to understanding the meaning of the sequence of signals arriving on an information channel than simply looking up each one in memory. By this procedure, however, we have to assume that the way by which a sensory signal gets to memory is done automatically and by means of the sensory features of the signal alone. This has important implications

for a theory of memory, as well as for the theory of attention and selection.

The general framework for this type of theory of attention was first stated in 1963 by the psychologists J. Anthony Deutsch and Diana Deutsch, then at Oxford, England (Deutsch and Deutsch, 1963). The Deutsches' theory was elaborated in a way not unrelated to the suggestion of Treisman by Norman (1968). Consider the scheme outlined in Fig. 2.2.

All signals arriving at sensory receptors pass through a stage of analysis performed by the early physiological processes. The parameters extracted from these processes are used to determine where the representation of the sensory signal is stored. Thus, as shown in Fig. 2.2, all sensory signals excite their stored representation in memory. Now, at the same time, we assume that an analysis of previous signals is going on. This establishes a class of events deemed to be *pertinent* to the

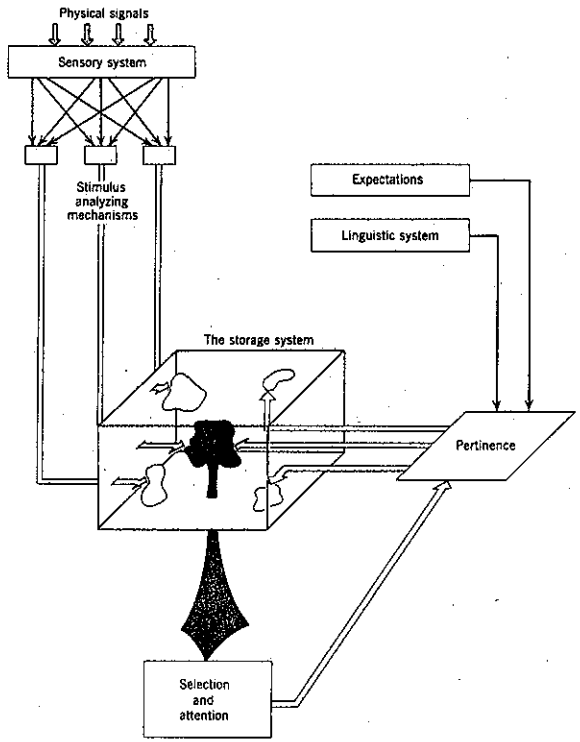

Fig. 2.2. The selection process. Both the physical inputs and the pertinence of information determine what will be selected for further processing. Physical inputs pass through the sensory system and stimulus-analyzing mechanisms before exciting their representation in the storage system. Simultaneously, the analysis of previously encountered material, coupled with the history of expectations and the rules of perception, determine the class of events assumed to be most pertinent at the moment. The material that receives the greatest combined excitation is selected for further attention.

ongoing analysis. The set of pertinent items also excites their representation in memory. The item most highly excited by the combination of sensory and pertinence inputs is selected for further analysis (the shaded item in Fig. 2.2.).

CHOOSING BETWEEN APPROACHES

Since 1968, much of the research on attention can be characterized as studies of two models. One model tends to favor early selection of incoming information, with the site of the attentional restriction early in the stages of processing. The other model tends to favor late selection of incoming information, with the site of attentional restriction late in the stages of processing. In the years since 1968, there has been increasing evidence favorable to both notions. Unfortunately, there has also been evidence that does not support either notion.

Suppose subjects are asked to shadow material presented to one ear, but also to push a button every time they hear a critical word (perhaps the word "tap") in either ear. Proponents of late selection would say that responses to the word "tap" should be equally good regardless of to which ear it is presented, for it has equal "pertinence" in all cases. Proponents of early selection say that if subjects are forced to attend to one sensory channel by the attentional demands of the shadowing task, then they will respond well to the word "tap" when it is in the same ear as the material to be shadowed and poorly otherwise. Indeed, when Treisman and Geffen (1967) did this study, the results supported the Treisman model of early selection and could not be easily explained by the model of late selection.

Suppose that while subjects listen to and shadow sentences presented to one of their ears, words semantically related to the shadowed sentences are presented to their other ears. In the early selection models, there should be no effect of the semantic relationship between unattended words and the attended message. But the late selection model predicts that the two sources of meaning will interact with one another. Notice that here we are talking about an interaction of meaning, not simple signals. It takes a considerable amount of processing to extract the meaning of a word from its sound.

Several studies have shown that the meaning of "unattended" words does affect performance on the "attended" task. Lewis (1970) showed that shadowing of the message on the attended ear was delayed when the word presented to the unattended ear was a synonym of the word in the attended one. McKay (1973) had subjects shadow sentences such as:

(a) "They threw stones toward the bank yesterday."

The word presented in the unattended ear was either "river" or "money."

Then, when presented with a recognition test for the meaning of the sentence to which they had attended, subjects were asked to choose between:

(b) "They threw stones toward the side of the river yesterday."
(c) "They threw stones toward the savings and loan association yesterday."

Subjects who had the word "river" presented to the unattended ear tended to select (b) as their choice for the meaning, whereas subjects who had the word "money" in the unattended ear tended to select (c). Moreover, the subjects did not remember what word had been presented to the unattended ear, although it obviously must have been processed to some considerable amount in order to influence the results.

Von Wright, Anderson, and Stenman (1975) tried a different approach. They presented an electric shock to their subjects (at a strength judged to be "very unpleasant") when one of a set of critical words was spoken to them. In one such training session, subjects were allowed to escape the shock if they depressed a switch within a half second of the start of the word. In such a procedure, subjects become conditioned to the word-shock combination, and presentation of a critical word causes a slight decrease in the skin resistance (probably caused by a slight amount of sweat secretion). Thus, one can detect whether the subject has responded to the word by measuring the skin resistance: This is called the GSR measure (galvanic skin response). After the conditioning had been successfully established, subjects shadowed a message presented to one ear while a list of words was presented to the other ear. In the unattended list, there were some neutral words, some of the critical words that had been conditioned to the shock, and both synonyms and homonyms of the critical words. (The study was done in Finland, and all the words were two-syllable Swedish words, accented on the first syllable.) Despite the fact that *no* shocks were ever given in the experiment itself, there was an appreciable GSR to the critical words presented in the unattended ear, and a lesser, though still substantial, response to the homonyms and synonyms. Again, this pattern of results supports the suggestion that information in the unattended channel is processed reasonably deeply. Thus, these experiments support the model that proposes late selection and are not easily explained by models proposing early selection. (This experiment is similar to one performed earlier by Corteen and Wood, 1972.)

Now where are we? First, we have at least two theories, each better than the other for some class of conditions, but each with its own difficulties. Moreover, there are some results not easily accountable by either. What do we do now? Whenever a situation like this occurs, it is usually wise to back up a bit and review what we are trying to do. These theories all have served a valuable purpose in guiding us to a good understanding of some phenomena, but if we go back and look at what is really meant by attention, we see that we have only scratched the surface. Another look at the phenomena might help us take a different view, one that might be more successful. Indeed, this is a useful tactic in scientific research in general. After awhile, theories and experiments tend to get lost, buried in the fine details of experimental techniques and results. The overall broad picture can be forgotten. We must remember that we are attempting to understand the wide range of human cognitive phenomena, and a new look at those phenomena might be rewarding, especially because of the tantalizing nature of our theories—so near, but yet not quite there.

4
Attention, Effort, and Resources

"Everyone knows what attention is." So said William James and so began Chapter 2. But does everyone know? Moray (1970) proposed six different meanings. Posner and Boies (1971) suggested three different components. Everyone may know what it is, but maybe it is more than one thing.

What is attention? Consider again the subjective feeling. When you get deeply engrossed in a task, so deeply engrossed that the world closes in and only the central task that is being performed is in focal awareness, then the rest of the world might as well not exist; we characterize that feeling as that of being in a highly attentive state. It is like a trance. When the state ends, there is often a feeling of exhaustion, as if a good deal of mental effort had been expended. In fact, attentional states are a key part of meditative techniques and many religious ceremonies.

Emotion and Arousal

Consider the plight of a diver, weighted down with equipment and faced with danger. The diver equipped with a self-contained underwater breathing apparatus (SCUBA), diving in cold water, may be heavily burdened by all the paraphernalia of that sport. He or she is probably wearing a bulky sponge-rubber wet suit, including a hood over the head, an air tank (weighing around 35 pounds) strapped to the back, perhaps 15 pounds of lead weights on a belt around the waist, fins, gloves, mask, snorkle, knife, depth gauge, compass, watch, decompression meter, air regulator and mouthpiece, air pressure meter, and some type of buoyancy compensator (inflatable life vest). It should come as no surprise that a person encumbered this way is not very agile. Suppose that

after 60 minutes of diving in very cold water, the diver is chilled, perhaps to the point of slight trembling. Now suppose the air tank and hose get tangled in the kelp (large seaweed). Such occurrences, though reasonably uncommon, should be of little concern to the well-trained diver. It is possible to remove the tank, and there are several well-known ways to get back to the surface without an air supply. Indeed, there is too much air, because the air in the lungs will expand as the diver rises to the surface.) But there is danger. So we have a tired, cold, and apprehensive diver about to perform a straightforward but dangerous task.

The psychological result of these conditions is to cause a state of high arousal. The diver will be anxious, under high stress. A number of different activities must be attended to, seemingly all at the same time. Failure to do them properly before exhausting the air supply could lead to injury or death. The result of such conditions seems to be the focusing of attention on a more and more narrow set of tasks. The result can often be tragic:

> *The question of panic occurs throughout the accounts of the diving accidents, panic that seems to override certain aspects of training even in divers who have had formal instruction. For example, it has been reported that in all of the deaths attributed to diving in California the diver was found still wearing his weight belt despite the attempts in diving courses to make jettisoning of the weight belt automatic in emergencies. The death of a young woman in Tucson a couple of years ago is illustrative. This woman, enrolled in a diving course but lacking experience, was reported to have surfaced in panic and drowned while diving for golf balls in a twelve-foot trap in dark water. When her body was recovered, she was wearing her weight belt, and in addition, was still clutching a heavy bag of golf balls.*
>
> **(Bachrach, 1970, Page 122; also see Egstrom and Bachrach, 1971)**

In general, it has been found that when humans become aroused, their performance on a task changes. If we draw a graph, plotting performance on the vertical axis and amount of arousal on the horizontal axis, the shape of the curve looks like an inverted U. At first, increases in arousal lead to improved performance, but then, as arousal builds up to its highest level, performance deteriorates. This relation has been named the Yerkes-Dodson law after its discoverers (Yerkes and Dodson, 1908). An explanation that seems both intuitively satisfying and supported by observations goes like this. As arousal level increases, attention becomes more narrowly focused. Attention to peripheral tasks decreases while attention to the central task increases; this improves performance. With ever-increasing levels of arousal, attention becomes more and more narrowly focused. This excessive narrowing eventually proves detrimental. Attention tends to focus entirely on one detail of the task, a detail that may be irrelevant. (This explanation is essentially a modern restatement of the position suggested by Easterbrook, 1959).

The role of arousal on attention, and thereby on performance, plays an important part of our everyday concept of attention. Most of us have probably noticed the relative narrowing or focusing of attention when apprehensive, or when a dangerous situation appears near. Instructors of sports such as underwater diving, flying, or parachute jumping try to combat the effects of anxiety on performance by *overtraining*. The goal is to make the appropriate responses to different situations become so automatic that they require no conscious attention. Automatic, nonconscious actions seem to be less susceptible to disruption by level of arousal.

AUTOMATICITY

Suppose we are active at several things at once: reading this book, eating, scratching, and tapping with one foot. It is easy to imagine reading this book while walking, book in one hand, apple in the other. Do we attend to all the things we do? Certainly, there would have to be more attention paid to the reading, less to the scratching, tapping, or walking. But is not some attention required to walk, to avoid obstacles, to keep in balance? The answer depends on one's view of attention. Some would equate attention with conscious awareness, and in this case it might only be the reading that is receiving attention. Others might equate attention with the general distribution of mental effort or mental resources to the various activities being performed. In this case, all the tasks receive some amount of attention, but the reading receives the most. Both those views, therefore, lead to similar interpretations.

A general rule appears to be that when a skill is highly learned—perhaps because it has been practiced for years and years—then it becomes automated, requiring little conscious awareness, little allocation of mental effort. Thus, how many tasks it is possible to do at once depends more on the level of training than on the task itself. A novice driver cannot both drive and talk, whereas a skilled driver can easily do both. Similarly, almost everyone can read while walking about. But if the passage being read requires deep thought, quite often the walking will stop. Highly skilled tasks seem to become automated, and thereby not as susceptible to disruption by withdrawing attention. The relationship between automatic processing and attentional control is so strong that one researcher, David LaBerge, has stated that

> *. . . to describe theoretically how automaticity develops is tantamount to describing the gradual elimination of attention in the processing of information. For example, imagine learning the name of a completely unfamiliar letter. This is much like learning the name that goes with the face of a person recently met. When presented again with the visual stimulus one recalls a time-and-place episode which subsequently produces the appropriate response. With further practice, the name emerges almost at the same time as the episode. This "short circuiting" is represented by the formation of a direct line between the visual and name codes. The process still requires attention, so this line is dashed, and the episodic code is used now more as a check on accuracy than as the mediator of the association. As more and more practice accumulates, the direct link becomes automatic (Mandler, 1954), represented by the solid line joining a letter with its name. At this point the presentation of the stimulus evokes the name without any contribution by the Attention Centre. Indeed, in such cases, we often observe that we cannot prevent the name from "popping into our head."*
>
> **(LaBerge, 1975, Page 58)**

LaBerge has shown how it is possible to disrupt tasks not yet fully automated by diversion of attention away from them. Remember the diver who, in panic, drowned in relatively shallow water, still clutching a heavy bag of golf balls and wearing her weight belt? Diving instructors worry about this problem. How do you train someone so they will perform the proper actions, even when in panic? The solution is to overtrain anyone who performs dangerous tasks. Make all actions become automated. Practice the set of possible responses to any situation over and over again. In this way, a minimum of attention is required, and in time of danger, the appropriate sequences get performed automatically. The trouble is that such training is hard to give and hard to take. The person being overtrained in some activity gets bored and wonders what all the fuss is about. Practicing the release of one's weight belt over and over again while diving in a swimming pool seems a pointless exercise to the student. But if that task can be made so automatic that it requires little or no conscious effort, then on the day that the diver needs to act under stress, the task may get performed successfully in spite of the buildup of panic.

For a task to be automated, the necessary components must flow readily from the memory system to whatever mechanism controls actions. In order to say more about the relation between practice and automatization, we need to know more about the way that information is represented within memory. Thus, we postpone further discussion of this issue until

we have completed our survey of memory. Chapter 9 is devoted to this problem: the relationship between practice, automatization, and memory.

Attention is a central concept in human behavior. It is a general, global aspect of human cognition, intimately connnected to one's state of self-awareness, of consciousness. We see that highly learned, automated activities may require little or no conscious attention to be performed. These highly automated activities can be disrupted, and when that happens they require conscious control, often to the detriment of their normal performance. In cases of mental pathologies and everyday stress some learned skills appear to break down, again requiring conscious control, thereby disrupting a person's normal life. In situations of high arousal, activities do not get performed properly, even when the failure to perform a simple act can lead to death. Thus, conscious attention to activities is a critically important aspect of the human mind, yet such conscious control can disrupt performance. The whole situation is well summarized by the old doggerel:

> A centipede was happy quite until a frog in fun,
> Said, "Pray which leg comes after which?"
> This raised her mind to such a pitch,
> She lay distracted in the ditch,
> Considering how to run.
>
> [Anonymous]

A CAPACITY MODEL OF ATTENTION

Now it is time to combine our studies of pattern recognition with our understanding of the phenomena of the attentional processes. In our study of attention, we saw that attention to one task appeared to limit the performance on other tasks, and the natural quesion to ask was, "at what stage in the processing does this limit occur?" But the very nature of the question implies a certain picture of the stages of information processing, one that we now see may not be an accurate characterization of the processing. When we ask about processing stages, we are thinking primarily of a bottom-up analysis of stimulus information, going from initial stages of sensory transduction, through the extraction of critical features, to the recognition of the input. But the various papers we have examined on pattern recognition show us that the process is more complex than that. Different sources of knowledge converge to aid the process of acquiring information. The system seems to be not a simple sequence of stages, but rather a set of interacting mechanisms, all working together to produce the final result.

The limit on attentional capacity appears to be a general limit on resources, not necessarily a blockage at any particular stage in the processing. Daniel Kahneman, an Israeli psychologist working at the Hebrew University in Jerusalem, suggests that attention and mental effort are intimately correlated, and that the major limitation on processing is in fact a limit of resources. Kahneman suggests that arousal can increase the amount of resources available to the subject. Here is his description of the processes:

*Attention and Effort**

DANIEL KAHNEMAN

The completion of a mental activity requires two types of input to the corresponding structure: an information input specific to that structure, and a nonspecific input which may be variously labeled "effort," "capacity," or "attention." To explain man's limited ability to carry out multiple activities at the same time, a capacity theory assumes that the total amount of attention which can be deployed at any time is limited.

— — —

A model of the allocation of capacity to mental activity is shown in Fig. 4.1. The model should be read beginning with the boxes labeled Possible Activities. These boxes correspond to structures that have received an information input (not shown in the model). Each such structure can now be "activated," i.e., each of the possible activities can be made to occur, by an additional input of attention or effort from the limited capacity. Unless this additional input is supplied, the activity cannot be carried out. Any type of activity that demands attention would be represented in the model, since all such activities compete for the limited capacity. Activities that can be triggered by an information input alone are not considered in the model.

Different mental activities impose different demands on the limited capacity. An easy task demands little effort, and a difficult task demands much. When the supply of attention does not meet the demands, performance falters,

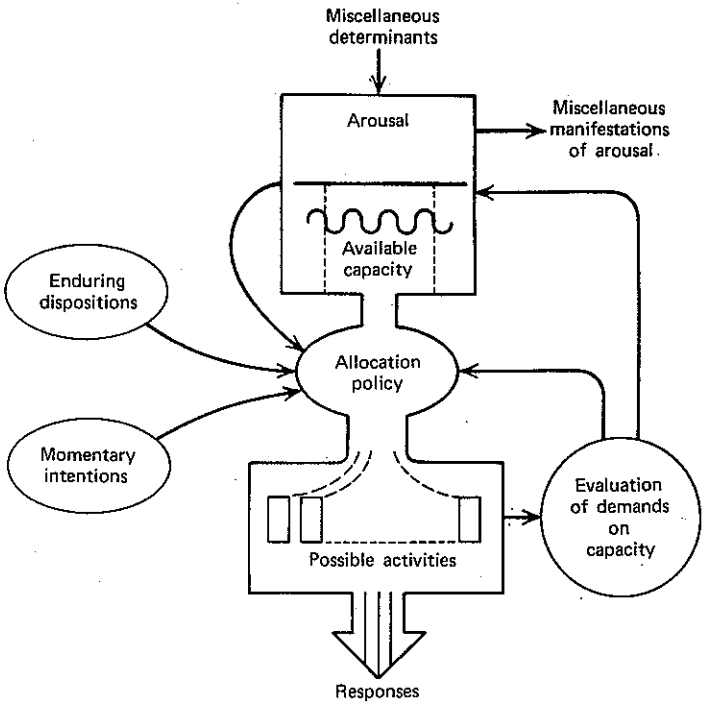

Fig. 4.1. A capacity model for attention.

or fails entirely. According to the model, an activity can fail, either because there is altogether not enough capacity to meet its demands or because the allocation policy channels available capacity to other activities. In addition, of course, an action can fail because the input of relevant information was insufficient. Thus, we may fail to detect or recognize a signal because we were not paying attention to it. But there are signals so faint that no amount of attention can make them plain.

A capacity theory must deal with three central questions: (1) What makes an activity more or less demanding? (2) What factors control the total amount of capacity available at any time? (3) What are the rules of the allocation policy?

— — —

The key observation that variations of physiological arousal accompany variations of effort shows that the limited capacity and the arousal system must be closely related. In Fig. 4.1, a wavy line suggests that capacity and arousal vary together in the low range of arousal levels. In addition, arousal and capacity both increase or decrease according to the changing demands of current activities.

The two central elements of the model are the allocation policy and the evaluation of demands on the limited capacity. The evaluation of demands is the governor system that causes capacity (or effort) to be supplied, as needed by the activities that the allocation policy has selected. The policy itself is controlled by four factors: (1) Enduring dispositions which reflect the rules of involuntary attention (e.g., allocate capacity to any novel signal; to any object in sudden motion; to any conversation in which one's name is mentioned); (2) Momentary intentions (e.g., listen to the voice on the right earphone; look for a redheaded man with a scar); (3) The evaluation of demands: there appears to be a rule that when two activities demand more capacity than is available, one is completed; (4) Effects of arousal.

— — —

The present chapter has illustrated two types of attention theories, which respectively emphasize the structural limitations of the mental system and its capacity limitations. Both types of theory predict that concurrent activities are likely to be mutually interfering, but they ascribe the interference to different causes. In a structural model, interference occurs when the same mechanism is required to carry out two incompatible operations at the same time. In a capacity model, interference occurs when the demands of two activities exceed available capacity. Thus, a structural model implies that interference between tasks is *specific,* and depends on the degree to which the tasks call for the same mechanisms. In a capacity model, interference is *nonspecific,* and it depends only on the demands of both tasks. Both types of interference occur. Studies of selective and divided attention indicate that the deployment of attention is more flexible than is expected under the assumption of a structural bottleneck, but it is more constrained than is expected under the assumption of free allocation of capacity. A comprehensive treatment of attention must therefore incorporate considerations of both structure and capacity.

The aim of cognitive processes is to form a meaningful interpretation of the world. Sensory information at any moment must be gathered together and interpreted in terms of a coherent framework. Past experience has created a vast repertoire of knowledge. Assume that this knowledge is organized into structural *frames* or *schemas* that can be used to characterize any experience. The problem of the perceptual processes is to determine the appropriate schema to match the present occurrences. When there are discrepancies, either a new schema must be selected or the current one must be reorganized.

Assume that when sensory information enters through the sensory system, the processes operating on it do so automatically, up through the extraction of features. Then as a result of these processes, the sensory memory is active, with different regions representing the different feature sets. Imagine, if you will, a memory space with regions of

activity flourishing here, fading away there. Each new sensory input starts up new activity, and the system must attempt to organize the structures that have been activated into some meaningful schema. This is a bottom-up, data driven analysis: analysis driven by the sensory input.

There are other ways to analyze things. Consider a schema that has been activiated because it is suggested by an input or from context. What else does this schema require? Use the requirements as a guide in the search for the feature space. Does the schema require a contour to the left? Ask if any procedure can provide data about one. Does the system postulate that it is perceiving a room? Then look for corners, walls, a ceiling. Ask if the feature space is consistent with the interpretation. These are top-down, conceptually driven analyses: analyses driven by the conceptual organization.

Kahneman's suggestion that the total resources available to a system are limited is a valuable one. It can be extended to a detailed analysis of processing mechanisms. To see this, follow the analysis presented by Norman and Bobrow:

On Data-Limited and Resource-Limited Processes*

DONALD A. NORMAN AND DANIEL G. BOBROW

RESOURCE-LIMITED PROCESSES

Consider the problem of performing a complex cognitive task. Up to some limit, one expects performance to be related to the amount of resources (such as psychological effort) exerted on the task. If too little of some processing resource is applied (perhaps because processing resources are limited by competition from other tasks being performed at the same time) then one would expect poor performance. As more resources are applied to the task, then presumably better and better performance will result. Whenever an increase in the amount of processing resources can result in improved performance, we say that the task (or performance on that task) is *resource-limited.*

The principle of continually available output allows an increased use of computational resources to be reflected in an improvement in performance. If a process using a fixed strategy did not provide an output until it was finished, then increasing resources would simply shorten the time required to get some output. But, if the process continually makes available its preliminary results, higher-level processes can continually be making use of them. As increased resources allow the process to upgrade the quality of its output, the improvement can be immediately used by any other processes for which the output is relevant. In a similar fashion, processing overloads need not cause calamitous failure, but simply a decrease in performance.

**D. A. Norman and D. G. Bobrow. On data-limited and resource-limited processes.* Cognitive Psychology. *1975, 7, 44–64. Copyright © 1975 by Academic Press, Inc. Reprinted by permission.*

DATA-LIMITED PROCESSES

Consider the task of detecting a superthreshold sound: for example, the sound made by striking a piano key in a quiet room. The detection task is straightforward: the processing is limited by the simplicity of the data structure. Consider now the task of determining whether or not a particular signal has occurred within a background of noise. Suppose the recognition mechanism uses all the most powerful techniques at its disposal—matched filters, correlated techniques, and so on. In either of these two tasks, once all the processing that can be done has been completed, performance is dependent solely on the quality of the data. Increasing the allocation of processing resources can have no further effect on performance. Whenever performance is independent of processing resources, we say that the task is *data-limited.*

In general, most tasks will be resource-limited up to the point where all the processing that can be done has been done, and data-limited from there on.

The implications of this argument are easy to derive, with one digression. First, we must consider how any process is affected by the resources given to it. In general, if a process is resource-limited, then the more resources allocated to it, the better the performance. If the process is data-limited, then its performance is not affected by resources: The graph detailing the relationships between performance and resource is shown in Fig. 4.2.

Most processes have both data-limited regions and resource-limited regions. Whenever some medium range of resources is allocated to a group of processes, some will appear to be always data-limited, others will appear to be always resource-limited, and still others will seem to be in the transition stage between resource-limited and data-limited, the exact status depending on how much processing resource is allocated.

Norman and Bobrow discuss central processing:

LIMITED CAPACITY CENTRAL PROCESSING*

When several active processes compete for the same limited resource, then the performance-resource functions of these processes become critically

* Norman and Bobrow. *op. cit.* 1975. Pages 49–50; 57–59.

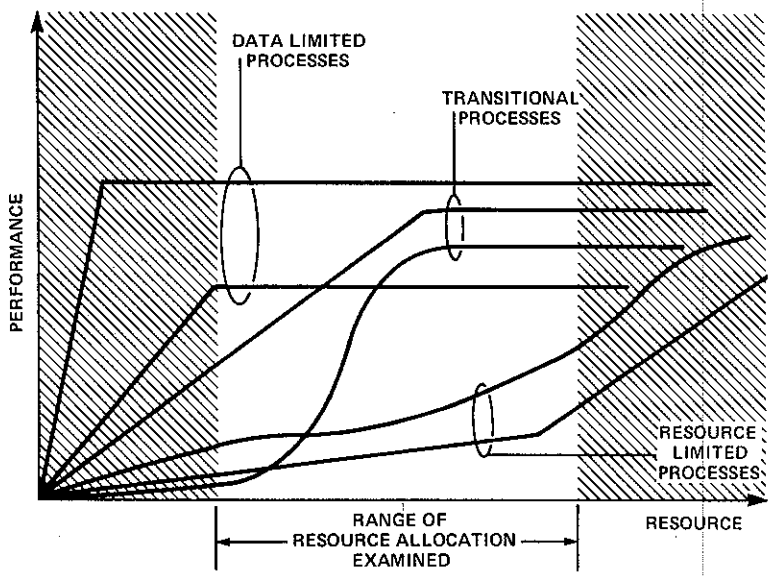

Fig. 4.2. Observable classes of performance-resource functions. When processes are examined only over a limited range of resource allocation, some will appear to be independent of resources (because they are data-limited in the region under consideration), others will appear to require indefinite amounts of resources (because they are resource-limited within this region), and others will be in a transition between data- and resource-limited operation.

important in determining just what effects will be observed. We assume that there is a fixed upper limit on available processing resources: Let the limit be signified by L. Operations which share the same limited capacity mechanism will not interfere with one another until the total processing resources required by all exceeds L. Moreover, in any given range of resource allocation, one process may interfere with others, but the others need not interfere with it. Just what kind of interference effects are found depends upon the particular form of the performance-resource function for each process. Interference can only be observed when a process is operating within its resource-limited region. Note, therefore, that the effects of interference need not be symmetrical. If task A interferes with task B, but not the reverse, then it would be incorrect to conclude that one of these tasks does not require processing capacity from the same central pool as the other. On the contrary, interference in either direction implies that both tasks draw resources from the same common pool. The asymmetry in effect results when one task is data-limited while the other is resource-limited. The symmetry or asymmetry of interference between two tasks is likely to depend in large part upon task instructions and subject strategy—upon which of the competing tasks receives first priority. The high-priority task will tend to be data-limited, and the low-priority task resource-limited.

Selective Attention

The literature on selective attention provides a rich set of data to be analyzed. Consider the experiment in which a subject is presented with two channels of spoken information by having two voices played to him over earphones, one voice to each ear. He is asked to repeat aloud the words that he hears on one channel (the procedure is called "shadowing") while the experimenter manipulates what happens on the other channel. In the literature the channel that is to be shadowed is called the *primary* channel and the other the *secondary* one.

In one such experiment, performed by Treisman and Geffen (1967), special target words were inserted into either the primary or secondary channels and subjects were instructed to tap the desk with a ruler whenever they detected a target word, no matter on which channel the target had occurred.

For the purposes of this experiment, we must compare the recognition of words on the primary channel with the recognition of words on the secondary channel. Processing resource is divided into two parts, with R_P going to the primary channel and $R_s = L - R_p$ going to the secondary channel. The relevant performance-resource functions are shown in Fig. 4.3. The perform-

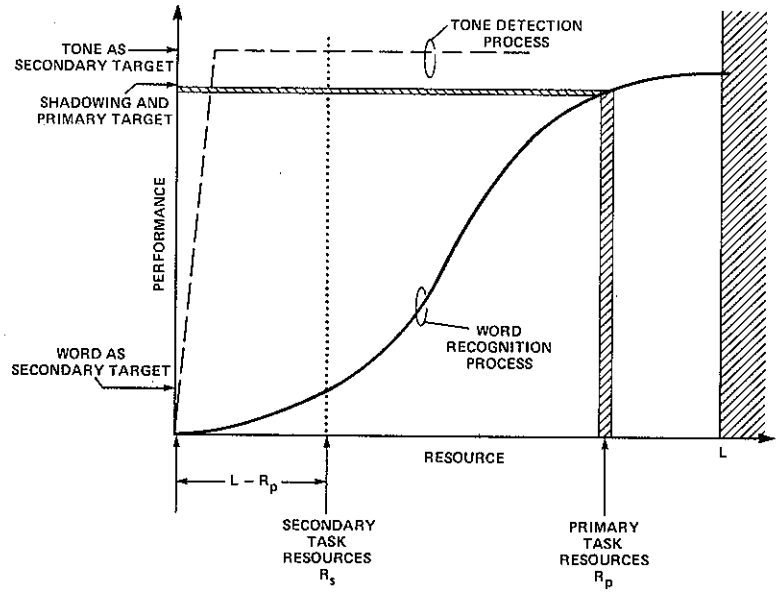

Fig. 4.3. The performance-resource function for a shadowing task. When the primary task is to shadow, sufficient resources must be allocated to keep performance relatively high. As a result, performance on a secondary task that is governed by a similar performance-resource function must appear to be "attenuated" (most secondary tasks that are verbal meet this criterion). A task operating on a much simpler performance-resource function will still be performed at a high level despite the simultaneous performance of the primary task. (This is illustrated with a data-limited function, corresponding to the task of detecting an auditory tone.)

ance on the primary task—shadowing—is determined by the instructions to the subject. Its level is high, yielding a reasonably high accuracy. The shadowing level tells us something of the performance level and thereby allows us to determine the primary processing resource, R_p. That being known, the maximum value that the secondary processing resource, R_s, can take on is $L-R_p$. Figure 4.3 shows that this guarantees a low level of performance for the detection of words on the secondary channel.

Suppose the target signal were an auditory tone instead of a word. Presumably the process necessary to detect a superthreshold tone and to discriminate it from speech sounds is rather simple, implying that it becomes data-limited at low resource values. Thus, as shown in Fig. 4.3, a tone should always be detectable on the secondary channel, even with very high performance levels on the primary channel. This is essentially the result found by Lawson (1966).

The analysis presented here is quite consistent with the idea that unattended inputs are "attenuated." The difference is that we are stating how a division of processing resource might force a process lower down on its resource-limited function, thereby essentially "attenuating" its analysis. The notion of "attenuation" is thus seen to depend critically upon the principle of continually available output in the operation of the relevant processes.

A THEORY OF ATTENTION

The picture we now have of attentional processes is the same as our earlier one, and yet it is quite different. If there really are simultaneous processes working, performing bottom-up, data driven analyses as well as top-down, conceptually driven ones, then the major limitation on what we can perform depends on how many resources are used.

Now we can begin to make sense out of a lot of the literature on attention. Since we no longer need ask the question about the location of a critical attentional bottleneck in the stages of processing, the earlier debate about the locus of attention is no longer relevant. Instead, we ask exactly what resources are demanded by a task and see if the demand exceeds the supply. If we assume that well-learned tasks are those that have become "automated," thus requiring little conscious control and few resources, then we can see why we can perform several well-learned tasks simultaneously, but only one or just a few poorly learned ones. When we first learn to drive a car, we cannot both drive and talk, and the task of driving seems difficult and mentally tiring. Later on, driving experience reduces the resources required, and we can drive, talk, sing, and still have excess capacity.

A good example of the nature of resource limits comes from the results of an informal experiment. While visiting the laboratory of William Johnston at the University of Utah, in Salt Lake City, I acted as a subject

in Johnston's experiment on attention. Johnston presented two different passages of text to his subjects over earphones. Both passages were spoken in the same male voice, and both were about scientific topics (one was about psychology, the other about some other, unrelated topic). Both passages were presented to both earphones, so both appeared to be localized in the middle of the head. The only way to tell the passages apart was by topic. One passage started slightly before the other, and the subject was asked to attend to it, for he would be asked questions about it at the end. This task is very difficult. When both passages occur at the same location in space, read by the same voice, it is very difficult to keep the two separated, but it can be done.

Now, in addition, a light in front of the subject occasionally increased in intensity for a brief period. Whenever the subject saw the light flash, he was to push a button. The amount of time that it took to respond to the light flash was taken as a measure of the attentional load on the subject.

The interesting part of the experiment came when one of the two passages was a familiar one. One passage was made familiar by playing it to the subject many times prior to the start of the experiment. After each playing, the subject was asked a question about its contents. When the experiment was performed, if the passage that was to be attended to was the familiar one, it was easy to follow the proper text, and the amount of time needed to respond to the light flash was reasonably short. When the task was reversed, so that the passage to be followed was unfamiliar while the passage to be ignored was familiar, then things were much more difficult. When I acted as a subject, I found it very difficult to concentrate on the proper passage, for phrases of the other one kept breaking through, automatically, despite my attempts to prevent it.

The most interesting result, however, came when the light flashed. Even when I had managed to get going properly and was listening to the unfamiliar passage, each time the light flashed and I made my response, the effort of making the response seemed to stop the listening process. I would lose my place and would have to start over from the beginning. The task seemed only barely possible, and even the very slight amount of processing necessary to respond to the light was sufficient to drive me close to the limit of my resources for performing the listening task. Needless to say, Johnston found that the reaction time to respond to the light in this condition was increased for all his subjects (Johnston and Heinz, 1974).

The word "attention" refers to a variety of concepts, each differing in meaning, but each overlapping the other. The consideration of mental resources adds a unitary concept to the nature of attention. A person can direct processing resources in many different ways, concentrating

sometimes on some aspect of the sensory input through the sense organs, sometimes on deep processing of internally generated ideas, and sometimes on preparing for a forthcoming activity. Moreover, the processing system is continually attempting to combine all sources of information at its disposal into a unified, understandable picture. Many different knowledge sources all interact. Some processing proceeds from the input signals, a bottom-up, data driven sequence of processing. Other processing proceeds from internally generated hypotheses or conceptualizations, a top-down, conceptually driven sequence of processing. Above all, there is some limit on how much processing can be performed at any one time.

In Chapter 2 we saw that the phenomena of attention studied in the experimental laboratory seemed to lead toward two conflicting models of the attentional process. Moreover, although each model had its virtues, each was also unable to explain some experimental results. From this chapter, we begin to see that perhaps both models were equally correct, but both were incomplete. Models of attention that postulate early selection can probably be compared with models of top-down processing. Both kinds of processing occur, and both models are correct.

If different knowledge sources interact with one another, and if the system can drive itself by both data driven and conceptually driven processes, then we can see how things fit together to make sense of the original efforts to piece together a theory of attention. When attention is concentrated on incoming sensory information, then the system is one that emphasizes data-driven, bottom-up analyses; it therefore looks as if it were an early-selection device. With other tasks, the system can be one that emphasizes conceptually driven, top-down processing and therefore can look as if it were a late-selection device. But other modes of operation are also possible, and these give rise to some of the other aspects of attentional phenomena.

The trail of the theoretical understanding of attention has been long and tortuous. You are forgiven if you think it to be muddled, full of confusions and contradictions. Such is the way of science. The early theories are absolutely necessary, for they combine the existing phenomena in useful ways, giving insights and helping to generate intelligent new experiments. These experiments lead to reconsideration of the theoretical picture. New, competing theories often emerge, but because each theory is constructed intelligently, each is correct for some things, wrong for others. Eventually, some new approach provides a way of incorporating all the previous results (and theories), and a new step of progress has been completed. You might think of a scientific theory as providing conceptually driven guidance to the field and of experiments as providing data-driven guidance. Both are necessary.

Beware. The resolution of the attentional processes presented here is not the final word. In the coming years we will learn more about the human information processing system. Some of the new evidence will present puzzles, strange anomalies that refuse to fit comfortably into our understanding of processing structures. But eventually some new formulation will come along, and the new data will then fit nicely with the old. A new stage will have been reached.