Behavioral experimentation

ALEXANDER POLLATSEK AND KEITH RAYNER

How might one study the complex processes of the mind? The method favored by early philosophers and psychologists was introspection. While introspection is still used today, perhaps the major source of evidence used by cognitive scientists to understand cognition is data collected from experiments in which subjects are engaged in some type of relevant task. While these data come from some type of experiment, the methods differ widely, and, as we shall see, the type of method is strongly influenced by the area of inquiry.

The establishment of the first major psychological laboratory in the late 1800s by Wilhelm Wundt led to numerous experimental investigations of basic processes associated with, among other things, cognition, memory, reading, speech perception, and attention. However, the onset of the behaviorist revolution in American psychology in the early 1900s resulted in psychologists not studying cognition for a long period of time. During the height of the behaviorist movement (about 1920 to 1960), rigorous experimental methods were devised to study various topics of interest to behaviorists. However, studies of mental processes were pretty much taboo during this period, as evidenced by the fact that most experimental psychology used animals such as white rats, rather than humans, as subjects.

The cognitive revolution, which occurred during the mid-1960s, led to a resurrection of interest in basic cognitive processes. While most of the theoretical tenets of the behaviorists were rejected by cognitive psychology, the rigorous experimental methods that they employed were adopted by the field. Cognitive scientists from disciplines other than cognitive psychology (such as computer science, philosophy, and neuroscience) now have a tendency to refer to research in cognitive psychology as behavioral experimentation (and the resulting methods as behavioral methods). Our own preference is to refer to the work as experimental cognitive research (and the methods as experimental methods).

In this chapter, we will first describe some basic experimental methodologies and then point out how they are used to address specific research questions. It is impossible for us to review thoroughly all the experimental methods in existence. Thus, we will selectively discuss methods which we believe are used most frequently. For the most part, we will describe methodologies used by cognitive psychologists. However, many of the methodologies have also been adopted by other cognitive scientists, particularly in interdisciplinary research. We will start with methodologies that were developed for studying low-level cognitive processes such as sensation and perception and then move to those used for studying higher-level cognitive processes such as memory and problem solving. One reason for this is that many of the tools that were developed for studying sensation and perception have been adapted for studying higher-level processes as well.

Basic experimental methodologies

Psychophysical methods

Psychophysical methods are among the oldest experimental techniques in psychology. Originally, these methods were developed to measure the psychological correlates of physical quantities, such as the subjective perception of the loudness of a tone as a function of the physical intensity of the stimulus. The goal was to understand the psychophysical function: the functional relationship between the physical and psychological quantities (Fechner: 1860). Basically, psychophysical methods fall into two categories: subjective judgment methods and discrimination methods. In the former, the subject reports his or her perception either by verbal report or by adjusting a comparison stimulus so that it bears some desired relation to the experimental stimulus. In the latter, subjects are asked to discriminate between two or more stimuli, and the percentage or pattern of errors is used as the basic data.

Subjective judgment methods As indicated above, these methods often rely on verbal reports: either a numerical judgment for a quantitative variable such as loudness or a categorical judgment, such as the color of a stimulus or which way a Necker (reversible) cube appears to be facing (see figure 27.1a). In the early research, it was assumed that (a) people’s introspections were synonymous with perception, (b) people could reliably report their introspections, and (c) the numerical values which subjects reported were linearly related to the actual internal states that were giving rise to these perceptions. It is still generally agreed that there are a reasonable number of circumstances in which these subjective judgments are reliable and may be the only way to get at the question of interest (e.g., what color does the stimulus appear to be?); but there is general agreement that assuming that a numerical judgment literally measures an internal perceptual intensity is problematic. Among other things, these numerical judgments have been shown to be sensitive to details of the instructions and to the composition of the actual set of stimuli presented in the experiment.

As a result, these methods have been refined by getting people to judge the relationship between the experimental or test stimulus and a standard stimulus. For example, in studying the Müller–Lyer illusion, experimenters might measure the strength of the illusion by giving subjects one of the two arrowhead stimuli in figure 27.1b and then have them adjust a standard (a horizontal line segment) so that it is the same subjective length as the test stimulus. This kind of procedure has two different variants. In one, the subject has control of the length of the standard and adjusts it until subjective equality is reached. In the other, the experimenter presents a sequence of standards, and the subject has to make a forced choice between the standard and the test stimulus (in this example, saying whether the standard is longer or shorter than the test stimulus). The typical method is to present a sequence of standard stimuli either increasing or decreasing in length by a fixed amount, and when the subject’s judgment changes, the experimenter reverses the direction of change and also the size of the increment. In the example above, the experimenter might start out with a standard appreciably bigger than the test stimulus and then present standard stimuli that are successively 1 cm smaller than the prior standard until the subject judges the standard to be smaller than the test stimulus. Then the experimenter increases the standard
stimuli by (say) 0.5 cm increments until the subject judges the standard to be bigger, at which time the standards would start to decrease in size by 0.25 cm, and so on, until the subject appears to be judging longer or shorter essentially at random. This is commonly called the staircase method. (Much the same procedure is used in the typical eye examination.)

In these comparison methods, as in the example above, the subject is usually asked to make a judgment of equality on a particular dimension. Other examples would be to judge whether two sources of light had the same amount of red in them or whether two tones were subjectively the same pitch. However, these methods were expanded in such a way as to try to reproduce the psychophysical functions without requiring absolute judgments (i.e., requiring numbers). In these methods, subjects would be asked to adjust a standard so that a certain relationship existed between two stimuli (e.g., adjust the standard tone so that it sounds half as loud as the test tone). Psychophysical functions can also be constructed from these data.

Discrimination data Almost all the above methods were widely criticized as being too subjective when the behavioristic revolution came to psychology. Hence, the emphasis changed to more objective methods, which were usually variations of the same—different judgment. Some of the time, these methods weren’t too different from those mentioned in the previous two paragraphs, although the usual question was the more objective one of whether a subject could discriminate between two stimuli, rather than the more subjective one of having the subject judge whether two stimuli were equal on a particular dimension. The purpose of these methods was originally to determine thresholds, or stimulus values below which subjects perceived nothing. Thus, on some trials, subjects were given a stimulus, such as a faint tone, and on other, catch trials, they were given no stimulus but were asked if they detected the tone. The basic data would be the hit rate, or the percent of times a subject would say that he or she detected the tone when it was presented, and the false alarm rate, or the percent of times a subject detected a tone on the catch trials. If the hit rate is no higher than the false alarm rate for a given intensity stimulation, the stimulus intensity is deemed as being below the threshold. Above the threshold, the difference between the hit rate and the false alarm rate served as another index of the subjective intensity of a stimulus. There is considerable controversy, however, about how to compare the hit and false alarm rates. The most commonly used value is called $d'$, which is the difference between the normal transforms of the two probabilities and comes from signal detection theory (Swets et al., 1961).

A second, and perhaps more general, paradigm entails measuring discrimination thresholds, or assessing how well people can discriminate stimuli that are similar to each other. Originally, the method most widely used was a simple same—different discrimination, in which two stimuli are presented (e.g., tones that differ by a small amount in either pitch or volume), and the subject simply decides if they are the same or different. Because both stimuli differ in how often they choose to respond same, most modern experiments use more sophisticated variants. Perhaps the most common is called the "ABX" paradigm. For example, the subject gets tones A and B successively (let’s say $A = 1,000$ Hz, $B = 1,005$ Hz), followed by tone X, which is the same as either tone A or tone B. The subject's task is to judge whether the third tone is the same as the first tone or the same as the second. (X is randomly varied to be the same as A or B from trial to trial.)

In the limit, if the subject is at chance on the task, one would say that his or her discrimination was below threshold. However, these methods are used more widely to get at discrimination functions. That is, the probability of discriminating A from B is used as a measurement of how far apart the two stimuli are subjectively. As a result, these methods have also been used to create psychophysical functions, using the discrimination measure to index the psychological distance on a continuum. One commonly used term here is the just noticeable difference, or JND, which (by convention) is the point at which the subject can correctly discriminate between two stimuli 75 percent of the time. To create a psychophysical function from these data, people often make the assumption that all JNDS are subjectively equal. That is, if someone can discriminate $A = 1,000$ Hz and $B = 1,010$ Hz tones correctly 75 percent of the time and can correctly discriminate $C = 2,000$ Hz and $D = 2,015$ Hz tones correctly 75 percent of the time, then the subjective difference between A and B is assumed to be the same as the subjective difference between C and D. This assumption, of course, need not be true and has been the source of considerable controversy.

Tachistoscopic methods As indicated above, the classic uses of psychophysical methods typically involved presenting stimuli for reasonably extended periods of time, with no time pressure on the subject’s response. Thus they presumably measured the subject’s perception of a stimulus after a reasonably extended period (often several seconds). However, a major issue that arose in perception was how percepts developed over time.
of the discrimination threshold method introduced above. The former method suffers from the problem that, in many cases, subjects rarely make errors in identification unless stimuli are not degraded in some way; hence discrimination thresholds can only be used to scale the psychological distance of normal stimuli that are highly confusable (such as tones differing only slightly in pitch or loudness). On the other hand, the problem with degrading stimuli to uncover perceptual similarities is that masks are not neutral (e.g., the pattern of confusions for speech stimuli could easily depend on the particular form of noise used).

Reaction time methods

Perhaps the most commonly used method in current human experimental psychology is measuring reaction time (RT), or the time which elapses between the onset of the stimulus presentation and the initiation of the subject's response. Reaction time is measured precisely, usually to the nearest millisecond. Moreover, it involves averaging over many trials; usually somewhere between 10 and 50 trials per condition need to be run on a single subject. One use of RT is as a measure of the similarity of two stimuli: the longer it takes to judge that two stimuli are different, the more similar they are. This is sometimes used as an alternative to threshold methods, because RT is often sensitive to differences in stimuli even when the subject can discriminate them without error.

The major use of reaction time methods, however, is to measure the time course of mental events in order to understand the components of mental processing. Let us illustrate with examples from what is one of the most popular uses of RT methods: visual word identification. One obvious question is how long the process of word identification takes. One could have subjects either name the word (i.e., read it aloud), perform a lexical decision on the word (i.e., judge whether it is a word or not), or perform a semantic categorization task (e.g., decide whether it represents a living thing). While the RTs in the three tasks can all be used to make inferences about the time of word identification, it is clear that the total time between presentation of the stimulus and the response in any of the tasks could not be taken as the time to identify a word. This is because the total RT includes not only the time to identify the word but also subsequent processes relevant to formulating a decision and response relevant to the specific task the experimenter sets after the word has been identified. Moreover, there are concerns in the naming and lexical decision tasks about whether subjects truly have to identify the stimulus in order to either name it or judge whether it is a word or not. However, the absolute RTs in some tasks can be used as upper bounds on the total time of the operation of interest.

More commonly, RTs are used to assess the relative time course of a process. Thus, for example, one could contrast the RTs to high-frequency (common) words with those to low-frequency (less common) words in one of the tasks mentioned in the previous paragraph. If the frequency of a word affects the word identification process, it should have an effect on the time to name the word (it does). Moreover, one can interpret the difference in RTs as reflecting the difference in identification times for high- and low-frequency words if one assumes that word frequency doesn’t affect stages of processing other than word identification (see Article 53, STAGE THEORIES REJECTED). Often
experimenter seek converging evidence by assessing whether the pattern of RTs is the same across different tasks.

Four major paradigms involving RTs need to be mentioned. The first is the use of RT to a stimulus as a probe of ongoing mental activity. For example, subjects could be cued to attend to a region of space (while holding their eyes fixed). If that region of space is in fact attended to, subjects should process stimuli in that region of space better than stimuli that appear in other regions. Hence one could present a probe and ask subjects to respond to it; if RTs are faster when the stimulus is in that region of space than when it is outside it, one would have evidence that subjects can in fact attend to regions of space.

The second paradigm is often known as priming (Meyer and Schvaneveldt, 1971). A typical priming experiment involves words (the prime and the target, respectively) which are presented in succession, and subjects are asked to respond to the target, making a lexical decision about it. The key manipulation is the relation between the prime and the target; in the original experiments, the prime and the target were either semantically related (e.g., CAT–DOG) or unrelated (e.g., HAT–DOG). The major result was that the RT to the target DOG was faster when preceded by the related prime CAT, indicating that (a) a memory trace of the prime lingers beyond its presentation, and (b) the relationship between target and prime is influencing processing of the target. One can thus view the above priming paradigm as a probe paradigm, where the target is the probe indicating (a) whether the prime is active and (b) which codes of the prime are active. Priming methods can also be used in more complex situations. For example, one can have a passage of text in which the word it appears, referring to a previously mentioned dog. One can then probe a subject immediately after reading it with CAT (e.g., requiring the subject to name it) to determine whether the concept dog has been reactivated. (The naming time would be compared to that of an unrelated control word.) One could also investigate the time course of this phenomenon by probing at different points in the text.

A third general paradigm using RTs is search (Neisser, 1964). For example, in visual search, an array of display elements (e.g., letters) is presented, and subjects press one key when a target (e.g., a B) is found and a second key when the target is not present in the display. One of the key manipulations is the number of distractor elements (i.e., nontarget), which is controlled by the number of distractor elements. If subjects need to search serially through the elements of the array to find the target, one would expect RT to be a linear function of the number of distractor elements. On the other hand, if subjects could process all the elements at the same time (in parallel), then one might even expect RTs to be unaffected by the number of distractor elements. In this research, it is often found that when discrimination of the target element depends on a single feature (e.g., a slanted line target with vertical line distractors), the search appears to be parallel; whereas when the discrimination of the target relies on integrating several features (e.g., a B target with various letter distractors), the search appears to be serial. (The pattern of data, however, has turned out to be far from simple.)

A fourth basic paradigm using RTs is the dual task paradigm (Posner and Boies, 1971). This paradigm is used to study attentional processes, but the basic logic is somewhat different from the probe paradigms above. The basic assumption of the method is that people have a finite processing capacity or pool of cognitive resources. If so, then one would expect that doing two tasks at once would be more difficult than doing either one alone if both took up processing resources, because the two tasks would have to share these resources. Thus, for example, one could have a subject name a visually presented word while simultaneously have them make a manual response to a tone (e.g., respond with the right hand if it is high in pitch and with the left hand if it is low in pitch). The key question is whether the response times for these tasks are slower when the subject must do the tasks simultaneously than when they are done in isolation. If so, one has evidence that both the tasks are using some shared processing resources. Moreover, the extent of the slowdown can be taken as a measure of the amount of resources that a task takes.

Another example of the dual task logic is the phoneme monitoring task, used to study speech perception. Here, the subject has two tasks to perform, both on the same stimulus. The first is to comprehend the meaning of the orally presented sentences. The second is to detect the presence of the phoneme /b/ and to press a key as soon as a /b/ is detected. The logic of the paradigm is that when lexical, semantic, or syntactic processing is easy, the target phoneme will be easy to detect. However, if the processing is difficult, it should take longer to detect the target phoneme.

We should close this section with a brief discussion of two general methods for using RTs to uncover basic components of mental processing. The earlier, dating from the nineteenth century, is the subtractive method. The original research by Donders (1868), for example, contrasted RTs in a simple RT task (e.g., the subject presses a single key when a stimulus appears, regardless of its form) with those in a choice RT task (e.g., the subject presses one of two keys, depending on whether it is a square or a circle). Assuming that the same stimuli were used in the two tasks, Donders assumed that the difference in RT between the two (~250 msec) reflected the time course of mental components involved in the choice RT task that were not entailed by the simple RT task (e.g., discrimination of the two stimuli, deciding which of two keys to press).

The subtractive method has been criticized because it assumes pure insertion of processing stages (i.e., that the mental processes in the choice RT task are identical to those in the simple RT task except that certain additional processing stages have been added). This assumption is harder to justify, leading to the development of a more widely used additive factors method (Sternberg, 1969). The key assumption that this method rests on is that there are discrete mental stages of processing between stimulus and response (i.e., that one stage must be finished before another begins). If there are discrete stages, and if variable A affects one stage of processing and variable B affects a different stage of processing, then one should expect the effects of A and B to be additive; by contrast, if two variables affect the same stage of processing, then it is likely that the variables will interact.

An example of the use of the additive factors logic would be to answer the following question: In processing a printed word in visual noise (i.e., random dots added to the word), is there a prior cleanup stage of perceptual processing before the word is identified? Variable A, which presumably influences the cleanup stage could be whether the stimulus is intact (with no visual noise) or degraded (with visual noise), and variable B, which presumably influences word recognition, could be the frequency of the word in the language. If the effects are additive, this means that the addition of noise has the same effect on high- and low-frequency words, which would be evidence that the
cleanup process precedes word identification. In fact, the noise affects recognition of lower-frequency words more, indicating that the cleanup process is not a separate stage, but intimately connected with the process of word recognition.

**Processing time methods**

Processing time methods are closely related to reaction time methods and tachistoscopic methods, but the distinction that we make in this chapter is that the unit of analysis in processing time methods is much larger than is typically the case in either reaction time or tachistoscopic experiments. As indicated above, most reaction time or tachistoscopic experiments require subjects to respond to a single stimulus or event, such as a single word which they must name. By contrast, with processing time methods, subjects either (a) may be asked to read a sentence or short paragraph of text, and the amount of time that it takes them to do so is measured; or (b) they may be given a fixed amount of time to read something, and their accuracy is measured.

Processing time methods have been used quite frequently in studies of language processing. Numerous studies have been conducted in which subjects are asked to read a sentence, and the amount of time that it takes to do so is recorded. By varying characteristics of target sentences, it is possible to infer something about the mental processes occurring during language processing as a function of differences between target sentences. For example, holding the number of words constant but varying the number of semantic propositions that are contained in the sentence has revealed that reading time is longer when there are more propositions.

While the **reading time** method just discussed provides information about how long it takes to read a sentence or a passage, many experimenters desire more accuracy in terms of how long it takes to process a given segment of text. Thus, the **self-paced reading** paradigm was developed to obtain greater precision about the time course of processing. In this paradigm, segments of text are presented on a screen, and subjects push a button to obtain the next segment of text. The size of the segment typically varies as a function of the issue being investigated: sometimes only a single word is presented as a segment, while at other times three words or an entire phrase or clause may be presented. In some versions of the self-paced reading paradigm, the new segment of text is always presented in the center of the video monitor from which the subject is reading. In another version (called the **self-paced moving window** paradigm) the text is presented sequentially across the screen, and usually dashes are presented as place-markers for the yet-to-be-presented words.

A major problem with the self-paced reading paradigm is that reading is slowed significantly when subjects must push a button to obtain more text. This is undoubtedly because the reaction time of the hand is slower than the reaction time of the eye. (In normal reading, new text is obtained by moving the eyes; see below.) The question then arises as to whether subjects engage in different strategies than usual when reading is slowed down.

Another attempt to study language processing involves holding the amount of time that a segment of text is presented constant. In the **RSP** (rapid serial visual presentation) paradigm, successive segments of text are presented for a set period of time (such as 200 msec) in the same spatial location on the video monitor (Potter et al., 1980). The presentation rate may be varied, so that single words are presented for 50, 100, 150, 200, or 250 msec, for example, and the rate of comprehension (measured as the number of words from a sentence that can be recalled) is then examined. This method is a generalization of the tachistoscopic method.

An analogous method for studying listening is the **time-compressed speech** paradigm, in which the rate of speech is increased by eliminating silent periods in the speech record to determine if listeners can comprehend speech when it is presented at fast rates. Finally, in the **auditory moving window** paradigm (which is analogous to the self-paced moving window paradigm discussed above), listeners push a button to get each new segment of speech.

These methods, of course, are not limited to studying reading or language processing. The time taken in other extended tasks, such as problem solving, are also used to make inferences about problem solving. Like the reading experiments, such experiments also attempt to investigate substages of processing by presenting only part of the problem and then having subjects press a key when they are ready for the next piece of information relevant to it.

**Eye-movement methods**

As we noted above, most of the processing time methods have an inherent problem wherein normal processing is slowed by the requirement, for example, to push a button to advance the next segment of text. The recording of eye movements allows researchers information about fine-grained processing without slowing down the process. For example, while the amount of time that it takes to read a sentence can be ascertained via processing time methods, recording of eye movements enables researchers to know how long it took to process individual words in the sentence. Furthermore, subjects can read at their own rate, so there is much less chance that specialized strategies will be adopted in experiments in which eye movements are recorded. Like the self-paced reading paradigm, the unit of analysis can be either a word or a series of words (usually referred to as regions of analysis). However, exactly where subjects look when eye movements are recorded is under their own control (unlike the self-paced method, where subjects are not free to look back in the text).

Eye movements have been utilized very effectively to study questions about reading, language processing (such as how readers process syntactically ambiguous sentences or sentences containing lexically or phonologically ambiguous words), visual search, scene perception, face perception, and so on. In many experiments dealing with reading and language processing, for example, readers are asked to read text, and their eye movements are recorded. Where readers look and how long they look at individual words have been documented to be related to the ease or difficulty associated with processing those words.

One particularly informative eye-movement method has been the **eye-contingent display change** paradigm (McConkie and Rayner, 1975). In this paradigm, changes in the text or scene that a subject is viewing are made contingent upon the position of the eye. For example, in the **eye-contingent moving window** paradigm, a window of text around a reader's fixation point contains normal text, but outside the window the text is perturbed in some way (all of the letters outside the window could be replaced with x's). In this paradigm, an eye-movement recording system is interfaced with a computer, which in turn is interfaced with a video monitor from which the subject reads.
An example of the preferential looking paradigm that is more extended in time is studying children watching television. In these experiments, the percentage of time that children look at the television (rather than something else in the room) is used as an index of their processing and their interest in the television message. For example, it has been shown that even relatively young children (two–three years old) will look at the television less often when the program “doesn’t make sense” (e.g., when segments of a cartoon are randomly sequenced) than when it does. It should be pointed out that these methods with children are measuring attention in a grosser manner than those with adults; the children are not only moving their eyes—they are usually moving their heads and bodies as well.

"Physiological" methods

In the last 20 years, there has been an increasing emphasis on studying mental processes by obtaining various physiological measures of processing. In all these methods, a recording device placed on the subject’s skin records some electric or magnetic activity that is thought to be a measure of some bodily process indicative of mental activity. Sometimes such records are used as indirect measures of mental activity. For example, measures of heart rate are thought to index effort (among other things) and thus are used in studying attention (especially with infants); measures of lip muscle activity may be used to index covert speech; palm sweating may also be used to index attention. However, the current focus is largely on measures thought to index actual brain activity. These methods will be discussed more fully elsewhere in this volume, so our discussion will be brief.

The most common method used is event-related potentials (ERPs), which are recorded from electrodes placed on the scalp. Although the electrical signals that are observed are small, by averaging the signal over many trials per condition (time-locked to the presentation of the stimulus), one can obtain a fairly reliable pattern of voltages that indexes brain activity. The voltages vary continuously both in polarity and in amplitude, resulting in a series of peaks and valleys that are reasonably interpretable. The following illustrates the method of interpretation. There is a peak (P200) that varies in amplitude with whether a stimulus is attended to or not, but not with whether or not it is the target in a detection experiment. Thus, it is claimed that this peak indexes the process of identifying a stimulus, but precedes the process of deciding what response to make to the stimulus.

While the logic is uncomfortably close to being circular, the method does offer some advantages over RT methods. First, if the ERP peak or valley does reflect the actual brain activity corresponding to identification of the stimulus, then this method might allow one to record the time course of mental events. In this case, the latency of the brain component (less than 200 msec) is appreciably less than the RTs measured in the previous section for identification of stimuli, so they give a tighter upper bound on when stimuli are identified. However, there is no guarantee that the ERP wave is reflecting the mental processing of interest (rather than being a later reflection of mental processing). Indeed, some ERP effects can occur after the response is made.

Another advantage of ERPs is that they have the potential to locate the activity in the brain. Currently, when spatial localization is attempted, arrays of 64 or more electrodes are placed on the scalp. With these methods, one can get quite precise spatial
patterns of electrical activity on the scalp. However, there is considerable controversy as to how well the pattern of brain activity can be inferred from the pattern of scalp activity. Fortunately, the inferences made from ERP recordings appear to be reasonably consistent with those obtained from methods that are better spatial indices of brain activity. (A method related to ERPs uses magnetic signals and gives better spatial localization measures, but is far more expensive.)

The two most widely used measures for localizing brain activity during mental operations are functional magnetic resonance imaging (fMRI) and positron-emission tomography (PET). Both may be somewhat risky to the subject (unlike ERPs): the former involves subjecting the head to strong magnetic fields, and the latter involves ingestion of small amounts of radioactive material. They also require access to extremely expensive equipment. They both index physiological processes in the brain relating to ongoing neural activity (in one case, blood flow), and one can obtain a fairly precise record of which areas in the brain are the most active.

There are two major problems with these latter two methods. The first is that they record the total activity over quite extended temporal intervals (often minutes). Thus, at present, they can index only the location of brain activity, not the time course. The second is that the brain undergoes all sorts of activity unrelated to the task at hand. Thus, a variant of the subtractive method (discussed earlier) is often used to interpret the patterns of activities. For example, in one condition a subject passively looks at words and doesn’t respond to them, whereas in another condition the subject decides whether each word represents a living thing. The difference in activity between the two patterns is then taken as indicating the brain regions used for processing the meaning of words (assuming that such processing is not going on automatically during the passive condition.) A similar subtractive logic is often used in ERP experiments as well.

The evolving methodology is to use both ERPs and either PET or fMRI on the same subjects, combining the temporal precision of ERPs with the spatial precision of PET or fMRI. Together, these methods hold out the promise of providing patterns of brain activity that are indices of mental processes. However, we should emphasize that, even in the best cases, these are merely indices of mental activity. For example, if one finds that region X of the brain is reliably more active at time t when subjects perform a semantic classification of a word than when they look passively at the same word, one cannot assume that that region of the brain is performing the classification at that time. The increased activity might merely reflect (for example) the fact that some other region of the brain has performed the computation of classification, and that region X is maintaining the result some time after the original computation for the purposes of responding or storing the information in memory. Hence, it is unlikely that these methods will replace the methods described elsewhere in the chapter. Instead, they can provide a useful supplement, enabling us to make some educated guesses about the relation between mental processing and brain activity.

Memory methods

Most experiments designed to assess what the memory representation is after someone has learned something are variations of a few basic paradigms. In these paradigms, subjects are typically presented with some stimuli in a learning phase of an experiment and then tested to see what they remember. For example, subjects may be presented with a list of 25 unrelated words and then asked to write down as many of the words as they can remember. Such a task is typically referred to as a recall task—subjects are asked to recall as many of the words as they can. The number of words recalled is used as a measure of memory. Another measure of interest is the pattern of errors. (This is related to the stimulus confusion methods described earlier.) For example, if errors tend to be semantically related to the test words, one might infer that the memory that they are being retrieved from is semantically organized.

A second basic memory paradigm involves a recognition judgment regarding the words. In such an experiment, subjects may see a study list of words (say 50 words) and then be presented with a test list containing 100 words, 50 of which were in the original list and 50 of which were not. They are then asked to judge which of the words in the test list are old (i.e., were in the study list) and which are new (i.e., not in the study list). The recognition memory paradigm is conceptually similar to several of the psychophysical methods we discussed earlier; in the former, the question is whether the sensory representation is above threshold, and if so, how strong it is; whereas in the latter, one is asking analogous questions about a memory representation.

A third method for testing memory is savings. In such a paradigm, subjects may be shown a list of words to remember and be given a large number of trials until they can remember them all correctly. They would then be brought back several days later, and their memory tested by examining the number of trials needed to learn the same list correctly. If the number of trials needed to relearn the list is less than the number of trials in the original learning, then one can infer that the subjects must have some memory of the original list. Of course, controls must be run to ensure that the improvement isn’t due to some improvement in the ability to memorize.) The savings method or variants thereof have often been shown to be the most sensitive measures of memory.

Modern variants of the savings method are variations of the probe methods described earlier. One example would be to present subjects with a study list of 50 words and then later present a series of 100 words tachistoscopically and have subjects identify them. Note that this is not a recognition task; the test phase merely asks subjects to identify the test words. What is usually found is that subjects identify test words that were in the study list more accurately and/or more rapidly than test words that were not in the study list. That is, the memory of prior learning is tested indirectly, or implicitly.

Interestingly, there is evidence of amnesic patients with damage to the limbic system in the brain (the hippocampus and related regions) who show relatively intact memory when tested on such implicit memory tasks, but extremely impaired memory on both recognition and recall tasks. This has led to theorizing that there are two fundamentally different kinds of memory systems, usually called procedural and declarative, in which these implicit memory tasks tap procedural memory, and recall and recognition tasks tap declarative memory. However, there is considerable controversy on this issue, and it is by no means clear that any method of testing memory is a pure index of one type of memory.

Question-answering methods

These methods are similar to memory methods, but the focus is somewhat different, in that these methods are directed more to nonliteral use of material. Two common
examples are text comprehension and problem solving. In the former, subjects are given text to read and then questions to answer based on their understanding of the text. In the latter, subjects are given a problem (such as a mathematics story problem) to solve. As indicated above, the focus in text comprehension is not on literal memory of the text, but rather on what the subject has done to the text (e.g., what inferences he or she has made, how the text has been organized in memory). Subsequent sections will discuss more introspective, process-oriented types of data obtained from these activities (such as having subjects talk out loud while trying to solve problems), but at present, we will discuss more objective, product-oriented types of data.

As with the simpler, perceptual tasks discussed earlier, one can use simple accuracy or latency measures to get at mental activity. For example, one could index comprehension of a passage of text by some combination of speed in answering questions about the passage and number of errors. One or could use speed of solution of a problem or the probability that subjects solve it as an index of the difficulty of the problem. These could be used to assess the effects of certain variables (e.g., comparing comprehension of the same ideas expressed in prose in two different ways).

In addition to measuring the number of errors, these methods often rely on examining the pattern of errors. For example, in text comprehension, a common question is whether the subject has made inferences of certain types. One method of probing this is to ask subjects whether they have seen a certain phrase in the text literally. If they falsely recognize inferences with a high probability, one may infer that the inference was drawn and stored as part of the process of reading the text. Conversely, one can examine the pattern of responses to the more usual pattern of comprehension questions (which ask the subject to answer on the basis of both the actual text and everything that can reasonably be inferred from it). For example, it might be hypothesized that children do not draw certain types of causal inferences from text (especially those relating to motivations of people in a story). Accordingly, one can draw up a battery of questions that probe different kinds of inferences and examine the pattern of results on this battery of questions.

We should point out that recent work has used probe methods related to priming to get at similar issues. One simple example is the following. If a text has a sentence in it, such as “The assailant stabbed the woman with his weapon,” but knife is never explicitly mentioned in the paragraph, one might want to know whether the reader inferred that the weapon was a knife. In such an experiment, the subject would read the passage, and immediately afterwards, they would name a target word aloud (knife in this case). The RT to name the target word would be a measure of the prior activation of knife. (This RT would have to be compared with other conditions, one of which didn’t involve stabbing, and one which explicitly mentioned knife.)

A problem with all the above methods which attempt to assess processing of a complex stimulus such as a text is that they involve responding to a question (or some other probe). Thus, one often cannot be sure whether the inference drawn from the material occurred during the original processing or only at the time at which the question or probe was presented. As a result, there is considerable controversy about how to interpret such results. However, the use of probe tasks such as naming a single word minimizes the chances that the inferential processes occur merely at the time of test.

The above controversy, however, does not extend to the use of these methods for studying problem solving and inference, because in these cases, one is examining thought processes explicitly provoked by the question. We will give two brief examples of using the pattern of responses to make inferences about mental operations. First, a common paradigm used to study inferential reasoning involves the presentation of an assertion like “Robins can get edomosis” (a fictitious disease); the subject is then to judge whether some other animal (e.g., a bat or an ostrich) is likely to get the disease. Commonly, subjects are asked to estimate the likelihood that the inference is true (scaling it from 0 to 100). Presumably, this method can be used to infer which aspects of similarity are used to support inference (e.g., superficial visual similarity versus deeper biological similarity). Some research has used this method cross-culturally to determine whether different cultures use similarity differently in making such judgments.

The second example uses a question-answering technique to examine the mental processes underlying probabilistic judgments. For example, when given the question, "The mean IQ of the population of eighth-graders in a city is known to be 100. You have selected a random sample of 10 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample?", a majority of subjects answer 100 (even though the correct answer is 105). This indicates that subjects are not using correct statistical logic in answering the question. Moreover, the answer is consistent with a theory called representativeness, which asserts that many people believe that all samples should be representative of the population (and hence have the same mean). Thus, these data are taken as evidence for the theory. Needless to say, there could be other explanations for this single piece of information. The logic used by good research in this area is to develop a set of substantive questions in which representativeness (or whatever theory is being examined) predicts a specific pattern of wrong and right answers, and then to examine whether the observed pattern of answers is consistent with the predicted pattern.

A problem with this research is to know what consistent means. Clearly, it would be unrealistic to expect that all subjects (or even a sizable subset of subjects) exhibit a pattern exactly like the predicted one. A subject might misread a question, make a slip of the pen in answering a question, or otherwise fail to attend; moreover, subjects may not consistently use the same logic or heuristic on every problem. Yet the theory might be of value in uncovering a mental process that a reasonable number of subjects use reasonably often in thinking about a particular domain. The question of how consistency should be measured, however, is still largely unresolved.

Observational methods

All of the methods described above typically involve subjects being tested in a laboratory setting. Thus, stimuli can be presented for precise periods of time; exact measurements can be taken of how long it takes subjects to respond or process the material; the number of errors made can be counted; and so on. Observational methods, on the other hand, typically involve observing behavior in a natural setting, which could be outside a laboratory setting (though it need not be).

Observational methods often utilize videotaping of subjects, and data analysis is based on trained raters analyzing the behavior of interest. Videotaping is not always done, and sometimes the ratings take place directly on the behavior (though obviously videotaping is preferred because when the raters disagree, the behavior can be reanalyzed). Interviews are also frequently used in observational studies.
These methods can be considered generalizations of the preferential looking studies described earlier to examine face perception and television viewing. The distinction here is that these observational methods often use significantly less objective coding schemes for characterizing what the subject has done. For example, the record of someone engaged in problem solving or in a social situation might be coded in terms of the types of activity they are engaged in and how much time they are involved in each activity. In many cases, these behavioral categories are not nearly as clear cut as deciding whether a child is looking at a TV set or not. Hence, it is usually important to have different people independently code the record, to determine whether such a coding scheme is reasonably reliable.

Often, the analysis rests heavily on a content analysis of the subject’s verbalizations during the activity. Two methods for provoking verbalization — both developed to study problem solving and reasoning — are worth special mention. In one, the subject is encouraged to think out loud almost continuously as he or she is attempting to solve a problem. A key aspect of this method is that the experimenter does not interact with the subject while he or she is verbalizing.

A second way of provoking verbalization is an interview. This could be characterized as an idealized Socratic tutoring session. In the interview, a subject is again encouraged to think out loud, but not necessarily continuously. However, the experimenter/interviewer periodically gives probes (either explicit questions or subtle hints to the subject that further elaboration of a thought is needed) that presumably help to draw out the mental processes that subjects bring to bear on questions. Obviously, the fact that the experimenter is probing the subject’s response makes it hard to characterize what the effective stimulus is to the subject.

A key controversy that runs through this research is how objective or scientific it is. Usually, there is some variable being manipulated in the situation, such as the type of problem presented, some detail of the problem, or some aspect of a social situation. One then examines whether this manipulation produces any differences in behavior; as examined by either an objective coding scheme or a more subjective or intuitive evaluation scheme. An experiment in which the subject behaves naturally in the situation and in which the coding is objective is to be preferred as long as one has confidence that the behavior itself and the coding scheme are really indicative of mental processes that are going on. However, in examining activities like problem solving, objective indices, such as how often the subject looks up at the ceiling, draws pictures to represent the problem, etc., are informative, even though they are unlikely to reveal a great deal about what mental structures the subjects are bringing to bear on the problem. This is why many researchers have used verbalization as a key part of the data base. But there is considerable controversy at present about the best way to collect and interpret such a record.

As indicated above, a widely used method involves subjects thinking out loud (with no intervention by the experimenter). The verbalizations obtained in this manner are usually assumed to be a relatively direct record of the subject’s actual thought processes. There are two principal problems with this method. The first, obviously, is that the verbal protocol may in fact be only tangentially related to the subject’s actual thought processes. The second is that thinking out loud is unnatural for many subjects. Spontaneously, subjects often stop talking out loud quite soon after starting to solve a problem. As a result, they often need to have preliminary training in thinking out loud (i.e., to keep talking throughout the problem-solving activity). This raises the concern that the act of continuous verbalization is causing subjects to solve the problem in a substantially different way than they would if left to their own devices. That is, this procedure may be influencing how subjects think by identifying thinking with verbal report.

The interview technique also asks subjects to verbalize, but not in a continuous stream. Instead, the emphasis is on having subjects give justifications for their answers or explanations of their reasoning. Unlike the thinking out loud paradigm, where verbalizations are assumed to reveal steps in thinking, the interview technique generally treats verbal reports as products of the thought process (i.e., a thought may be verbalizable only after the subject has solved a problem or a part of a problem). Probing by the experimenter is used, because it is felt that subjects’ spontaneous verbalizations may be quite ambiguous when it comes to revealing actual mental processes. Of course, this results in a loss of objectivity and reproducibility in method. As a result, these methods are often used in conjunction with the more objective question-answering methods described above, where the objectively formulated question evolves from the interview situation.

Summary

The methods we have discussed have progressed from more objective methods, in which the experimenter has complete control of the stimulus, and the measurement of the subject’s response is quite objective, to methods in which both control of the stimulus and characterization of the response are quite loose. This progression is largely a result of the problem area being studied.

The initial methods evolved for studying the perception of simple stimuli that are easy to characterize objectively (simple tones and visual stimuli) and are not extended in time. We then moved to methods that are used to study more complex ecological stimuli, such as reading text or processing extended speech or television, but in which the task was still largely one of processing the stimulus. However, such tasks soon involve memory and higher-level processes, as processing text or television involves having not only a literal memory, but also imperfect memory and interpretation and condensation. We then moved to methods used for explicitly studying memory, and then to methods for studying problem solving and thinking.

The questions then move from ones that are easy to characterize, such as “Did the subject perceive the stimulus correctly?” and “What did the subject remember about the stimulus?” to more complex questions, such as “What mental processes did the subject use when solving this problem?” We have indicated that somewhat less objective methods have often been used to study these complex mental processes and that these less objective methods may be justified. This latter characterization is somewhat controversial, however, and there are many psychologists who object to the use of such methods. For instance, in our example of people’s views of random sampling, many researchers believe that only objective (i.e., reproducible) question-answering methods are valid. These involve giving people several variants of a problem and looking at the pattern of responses across these variants.

There are several problems with these more objective methods, however. The major one is that the responses are often difficult to interpret. One symptom of this is that it is
often the case that minor differences in wording may cause large differences in
patterns of response. Thus, for example, it is often unclear whether the subject’s dif-
ficulties with a particular problem are a superficial function of the wording or due to the logic of
the problem (the latter is presumably what is of interest to the experimenter). One
possible solution here is to present many different wordings and to look at the pattern
of responses over all these wordings. On the other hand, this is likely to lead to a com-
binatorial explosion of different problems and/or different versions of a single problem
needing to be presented. This may require either an impractical number of subjects or
asking a subject to solve an impractically large number of problems. As a result, there
appears to be a real need for less objective methods (i.e., ones which may not be so
easily reproducible) as well for studying complex mental processes. In our opinion,
research in these areas should attempt to combine both more and less objective methods.

References and recommended reading

Donders, F. C. 1868: Over de snelheid van psychische processen. Onderzoekingen gedaan in het
Psychologisch Laboratorium der Utrechtse Hoogeschool. Tweede reeks, 2, 92–120. Translated
by W. G. Koster for Attention and Performance, vol. 2, which was published in Acta
Psychologia, 30, 412–30.

McConkie, G. W. and Rayner, K. 1975: The span of the effective stimulus during a fixation in
reading. Perception and Psychophysics, 17, 578–86.

of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227–
34.

408.

Potter, M. C., Kroll, J. F. and Harris, C. 1980: Comprehension and memory in rapid,
Erlbaum, 395–418.

Sperling, G. 1960: The information available in brief visual presentations. Psychological Mon-
ographs, 74 (no. 498).

Koster (ed.), Attention and Performance, vol. 2, also published in Acta Psychologica, 30, 276–
315.

Swets, J. A., Tanner, W. P. Jr. and Birdsall, T. G. 1961: Decision processes in perception. Psycholog-
ical Review, 68, 301–40.