HUMAN EXPERIMENTAL PSYCHOLOGY

Joan Gay Snodgrass
Gail Levy-Berger
Martin Haydon

New York Oxford
OXFORD UNIVERSITY PRESS
1985

5

Mental Chronometry: Measuring the Speed of Mental Events

Donders' subtraction method 89
Sternberg's additive factors method 91
Memory-scanning experiments 92
Parallel self-terminating search 93
Serial self-terminating search 95
Serial exhaustive search 96
Factors that affect other stages in memory scanning 98
Posner's Same/Different classification task 98
The letter-matching task 99
Methodological issues in reaction time 101
What is the minimum reaction time? 101
The problem of very long reaction times (outliers) 102
Error rates and the speed-accuracy trade-off function 103
Summary 105

> Would thought also not have the infinite speed usually associated with it, and would it not be possible to determine the time required for shaping a concept or expressing one's will? For years this question has intrigued me.
>
> (Donders, 1868/1969, p. 417)

With this comment, F. C. Donders, a professor of physiology at the University of Utrecht, introduced the field of mental chronometry to psychology in a paper published in 1868 entitled "On the Speed of Mental Processes."

Donders' method of measuring the speed of mental processes, usually referred to as the *subtraction method,* fell into disfavor around the turn of the century until it was revived 60 years later, in a different form, by a group of modern-day psychologists. The most influential of these has been Saul Sternberg, whose revision of the subtraction method, called the *additive factors method,* we describe in detail later in this chapter. However, it is important to understand the history of these early attempts to measure the duration of mental processes, why these initial attempts failed, and why modern attempts to use the subtraction method have succeeded.

DONDERS' SUBTRACTION METHOD

Donders was influenced in his attempts to measure the speed of mental processes by the work of Helmholtz. In 1850, Helmholtz attempted to measure the speed of nerve transmission in the frog by measuring the time between the stimulation of a part of the frog's body and the resulting muscular contraction. He later applied this technique to humans by measuring the time to respond to a mild electric shock delivered at points of varying distance from the brain. Helmholtz used the method of subtraction to do this: He measured the difference in time between stimulation of the elbow and stimulation of the hand. Then, knowing the approximate length of nerve fibers between the hand and elbow, he was able to use the difference in reaction time as a measure of nerve transmission time.

Donders sought to measure processes vastly more complex than nerve transmission time, namely, the sum of all the processes that must intervene between presentation of a stimulus and activation of a voluntary response. When there is a single stimulus to which the subject makes a single simple response, such as moving the hand or foot, the time elapsing between stimulus presentation and completion of the motor response is known as the *simple reaction time.* Donders lists no fewer than 12 mental events that must take place between presentation of a stimulus and its motor response in simple reaction time. Present-day reaction-time theorists have whittled the number down to three subprocesses: stimulus input time, central processing or decision time, and motor response time.

Donders saw no way to disentangle the three processes and separately measure each component—in fact, we are in much the same state today. However, he proposed that by complicating the simple reaction time to make it what we now call *choice reaction time,* he could "insert" one or more steps in the simple reaction time mental process chain and, by subtraction, measure the time of those added steps.

In *choice reaction time,* two or more stimuli are presented, and the subject must indicate which stimulus has been presented by producing one of two or more responses, a different response for each stimulus. The choice reaction time must include both the time to discriminate one stimulus from another *(discrimination time)* and the time to select one of the several motor responses *(motor choice time).* Thus, Donders reasoned that the difference between simple reaction time (which he called the *a-reaction*) and choice reaction time (which he called the *b-reaction*) must represent the sum of discrimination time and motor choice time.

Next, to get rid of motor choice time, he invented the *c-reaction.* (Today, we call it

Donders' c-reaction.) In the c-reaction, the subject is presented with two or more stimuli, just as in the b-reaction, but makes only a single response, to one of the stimuli, and omits that response to all others. The c-reaction might be called a go/no-go response: the subject responds by making a single motor response to one stimulus (the go response) but omits the response to all others (no-go). This procedure, Donders reasoned, should eliminate motor-choice time and leave only discrimination time. The responses thus increase in complexity from a to c to b (the order of the letters reflects the order in which they were developed rather than their complexity). By subtracting c from b, Donders was able to measure motor-choice time, and by subtracting a from c, he measured discrimination time.

To summarize:

a-reaction = stimulus input time
+ decision time
+ motor-response time

b-reaction = stimulus input time
+ decision time
+ *discrimination time*
+ *motor-choice time*
+ motor-response time

c-reaction = stimulus input time
+ decision time
+ *discrimination time*
+ motor-response time

Therefore

and $c - a =$ *discrimination time,*

$b - c =$ *motor-choice time.*

Donders used a variety of methods to test the subtraction method, but the one he reports on most fully in his paper (1868/1969) used the experimenter's pronunciation of one of the syllables* *koo, ko, kah, cay, key, queue* as the stimulus and the subject's repetition of that syllable as the response. In the a-reaction, only one syllable is presented, and the subject always repeats just that syllable. In the b-reaction, any of two or any of six of the syllables are spoken and the subject repeats whichever syllable has been spoken. In the c-reaction, one of the syllables is designated beforehand as the target, or to-be-responded-to stimulus, and the subject repeats that syllable when it is presented and remains silent for all others.

From this experiment, Donders found that c − a, the discrimination time, was about 36 ms (millisecond or 1/1000 of a second) whereas b − c, motor-choice time, was 47 ms. These times, even by today's standards, are exceedingly short, and, as Donders points out, probably represent the minimum values for these mental processes, since repeating what someone says is a highly overlearned task (or, as we would describe it today, there is high stimulus-response compatibility). For other, less well-learned stimulus-response links, such as pressing a right key to a red light and a left key to a white light, the differences are much larger and represent the time of those processes when no strong compatibility exists.

The subtraction method was soon adopted in other psychological laboratories of the day, but with disappointing results. First, the c-reaction was not always found to be shorter than the b-reaction, which it must be if the method is to be valid. Recall that in the c-reaction, the subject responds to one stimulus and refrains from responding to a second. This reaction was developed by Donders to eliminate the stage of motor choice, which is assumed to be included in the b or choice reaction. However, experimental psychologists of that day, Wundt among them, pointed out that the c-reaction *does* involve a motor choice—a choice between making and not making a response. To eliminate motor choice entirely, Wundt invented an alternative reaction-time response, the *d-reaction*, or "discrimination" reaction. Wundt's d-reaction is like Donders' b-reaction in that several different stimuli are presented. It is like the a-reaction in that only a single motor response is made, and it is made to all of the

*Approximate English translation as given by Koster (Donders, 1868/1969, p. 410).

stimuli. The difference between the d- and a-reaction is that for the d-reaction, the subject is instructed to recognize or identify the stimulus before responding—in short, to discriminate it from other possible stimuli. Wundt thus assumed that discrimination time could be measured by subtracting a- from d-reaction times.

As with the c-reaction, however, d-reaction times were unreliable, sometimes being as fast as a-reactions, and sometimes slower than b-reactions. The problem is that there is no way for the experimenter (or, for that matter, the subject) to know that the stimulus has been identified before the response is made, since the same response is made to all of the stimuli.

Another criticism of the method came from introspections of the subjects (who were usually the experimenters themselves). Subjects observed that their internal mental operations differed in the simple and choice reaction-time tasks. For simple reactions, the response was evoked by the stimulus as if it were a prepared reflex, with little in the way of voluntary decision involved. For choice reactions, in contrast, the subjects were aware of a variety of cognitive processes that intervened between stimulus and response. In addition, motor readiness seemed to be much higher in the simple than in the choice reaction time situation, so the motor response time component of both was unlikely to be equal. This is a particularly devastating criticism because the subtraction method is based on the assumption that inserting the processes of discrimination and motor choice into the a-reaction to produce the b-reaction does not affect the common stages of stimulus input and motor-response time. If this assumption is not true, the whole method is invalidated. Because of these problems, the subtraction method was abandoned as a way of timing mental events (although the use of reaction time as a dependent measure in a variety of tasks continued).

In the next section, we consider a reinterpretation and restatement of Donders' position proposed by Saul Sternberg. Sternberg's contribution was to reinstate the study of timing of mental processes by showing how substages of mental processes could be studied by subtraction. Sternberg's subtraction method differs from Donders' in that it does not involve inserting or deleting a whole stage of processing, but rather is based on manipulating variables to affect the amount of time each stage requires. This method is called the *additive factors method.*

After discussing Sternberg's theoretical contribution, we consider a number of different questions about the mind's functioning that have been asked using the additive factors method, and finally we discuss some methodological issues on the use of reaction time as a dependent variable.

STERNBERG'S ADDITIVE FACTORS METHOD

The approach of the additive factors method is to manipulate the task of the subject in such a way that a complete stage (such as discrimination or motor choice) is not deleted but rather is simply affected—either lengthened or shortened—by the experimenter's manipulation. Use of the method can indicate how many stages there are, how long a particular stage, or combination of stages, must take, and which variables affect which stages.

The best way to illustrate the additive factors method is with a concrete example, but before we describe the particular example, we need to introduce some new terminology. All of the experiments that test the additive factors method use a binary or two-choice reaction time paradigm. By that, we mean that the subject chooses one of two responses (usually the press of one of two keys) in response to the presentation of a stimulus. In *binary classification* experiments, there are more stimuli than responses, and the subject's task is to partition the set of stimuli into two exhaustive and mutually exclusive categories by responding to one set with one response and to the second set with the second response.

In binary classification experiments, then, there is a many-to-one relationship between stimuli and responses.

Memory-Scanning Experiments

The particular experiment we use to illustrate the additive factors method is known as a *memory-scanning* experiment or paradigm (sometimes also referred to as the Sternberg paradigm). In this paradigm, subjects are given a short list of items, such as digits, letters, or words, to memorize. The length of the list is varied within the limits of short-term memory (see Chapter 8), so that no more than five or six items are presented for memorization. The length of the memory list is called the *memory set size*, and the memory set is either varied from trial to trial, in what is called the *varied-set procedure*, or remains constant across a blocked series of trials, in what is known as the *fixed-set procedure*. The results from both procedures are quite similar, and in a laboratory that is not fully automated, the fixed-set procedure is more convenient. However, for purposes of illustration, we will use the varied-set procedure as a model.

After the memory set has been presented to the subject (usually with a visual display), a trial begins with the presentation of the test item, or *probe*. The probe can be selected from either the memory or *positive set* or its complement, the *negative set*. If the item is from the positive set, the subject responds by pressing a YES button, and if it is not, the subject responds by pressing a NO button.

Results from some typical trials in a memory-scanning experiment are shown in Table 5-1. The first column in Table 5-1 gives the trial number; the second column shows the actual memory set presented; the third column shows the set size (the number of items in the memory set); the fourth column shows the probe; the fifth column shows the correct response; and the sixth column shows a typical reaction time.

It is apparent that the memory demands on the subject are not great (that is, forget-

Table 5-1 Typical Trials in a Memory-Scanning Experiment with a Varied Set Procedure, Showing the Memory Set in Use on Each Trial, the Probe Presented, the Correct Response, and a Typical Reaction Time

Trial number	Memory set (visually displayed)	Set size (n)	Stimulus probe	Correct response	Reaction time (RT)
1	4,6,1	3	1	YES	470
2	2	1	6	NO	440
3	5	1	5	YES	390
4	7,2,8,9	4	0	NO	560
5	1,3,2,9,7,4	6	9	YES	590
6	4,5	2	9	NO	480
7	9,5,0	3	0	YES	470
8	0,4,2	3	3	NO	520
9	0,7,9,6	4	8	NO	560
10	4,2,0,6,8	5	7	NO	600
11	4,0,3	3	3	YES	470
12	0,3	2	8	NO	480
13	1,0,3,7,4	5	1	YES	550
14	3,7,1,8,0,4	6	0	YES	590
15	2,1,9,3	4	2	YES	510
16	6,2,1,8,3	5	7	NO	600

Note. The RTs are predicted by the following equations:

RT (YES) = $350 + 40n$

RT (NO) = $400 + 40n$

The RT depends on both the set size and the response types.

ting which items are in the positive set is unlikely), so few errors occur in this task and the major variable of interest is response time. The substantive question of interest in this experiment is how the subject accesses items in short-term memory (that limited-capacity memory system holding information we are immediately aware of) to make the yes/no decision. Two questions are relevant to asking how a subject searches short-term memory. First, does it take longer to decide an item is in short-term memory if there are more items to choose from (that is, when the memory set size is increased)? And second, if reaction times do increase with memory set size, does the way they increase tell us anything about how the subject searches his or her memory? The answer to both of these questions is "Yes." The larger the memory set, the longer it takes the subject to make both positive and negative decisions. Furthermore, positive and negative reaction times increase linearly, and at the same rate, with memory set size. This second result provides us with a way of determining how the subject searches short-term memory.

Figure 5-1 shows some hypothetical data from an experiment in which memory set size was varied from one to six items, positive and negative items were presented equally often, and the RT functions for positive and negative responses are plotted as a function of memory set size. The positive and negative responses are indicated by different lines, and memory set size increases along the abscissa. The equations of the best-fitting straight lines (using the linear regression techniques described in Chapter 15) are shown on the graph.

Before we reveal the secret of how we can discover the subject's search strategy by the way his or her reaction times increase with memory set size, let us consider some possible strategies subjects might use in this task. First, to determine that a stimulus probe item *is* a member of the positive set, the subject need only find the item in the memory set that matches

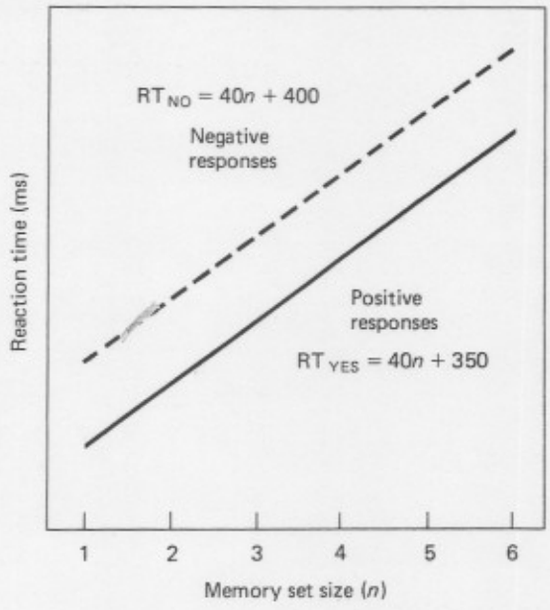


Fig. 5-1 Hypothetical data from a memory-scanning experiment.

the probe. Logically, subjects could terminate their search as soon as they had located the positive item in the set, so we call the positive decision a logically self-terminating one. As an analogy, imagine looking for a book of matches in your purse or pocket. Once you find the matches, you could terminate the search.

Consider, however, the problem of determining that a probe item *is not* a member of the positive set. In order to do this, the subject needs to search through the entire set of items in memory in order to verify that the item is not there. The analogy to a missing book of matches in purse or pocket should be clear. Logically, the negative decision is exhaustive in that all items must be checked before deciding the probe is not among the set. With this in mind, let us consider some possible strategies that might be used in this task, and what they predict about the relationship between RT and set size. These are shown symbolically in Figure 5-2.

Parallel Self-Terminating Search. In parallel self-terminating search, subjects search the items in their short-term memory by considering all of them simultaneously, and terminating the search once

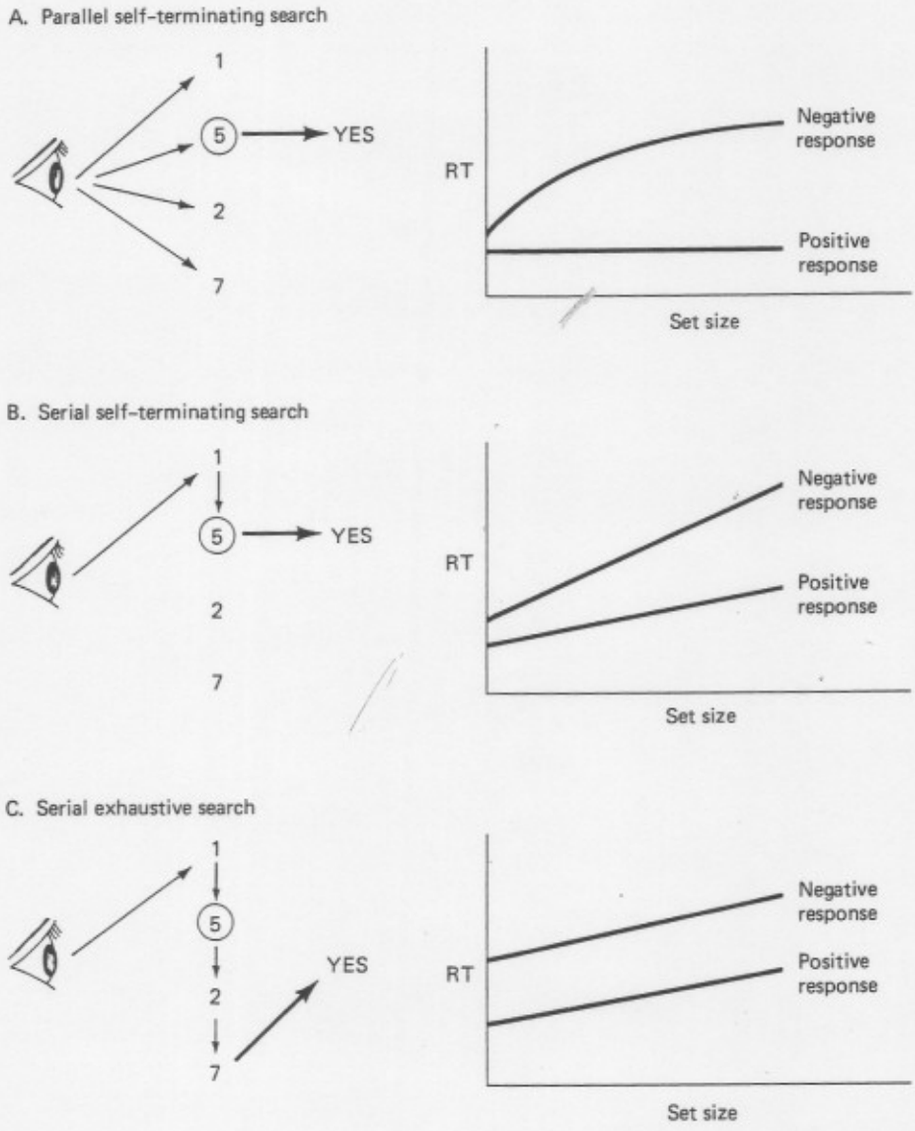


Fig. 5-2 Some possible strategies, and their predicted RT functions, for the memory-scanning task. The memory set consists of the four items (1,5,2,7), and on the trial illustrated, the probe is 5 so a positive response is correct.

the probe item is found. This process is illustrated in Figure 5-2A by the arrows from the mental eye pointing to each member of the entire set simultaneously. If subjects could do this with no loss of efficiency as memory set size increases, then YES responses should be unaffected by how many items are in the positive set. Thus, the function relating positive responses to memory set size should be flat, as shown in the graph to the right.

For negative decisions, the situation is somewhat different. Even though subjects can access all items simultaneously, if there is some variability in access times across items *and* if they need to wait until all have been accessed before they can decide NO (the logically exhaustive criterion for negative decisions), then the more items there are, the longer it should take for NO responses.

If this is not clear, think of a horse race in which we vary the number of horses running in the race. The number of horses corresponds to memory set size. You, the observer, are seated at the finish line and can only tell a particular horse ran the race by whether or not the horse crosses the finish line. A positive decision that a particular horse was in the race can be made whenever that horse crosses the finish line, and that horse's speed will be unaffected by

how many other horses are running in the race (disregarding all the obvious limiting physical and psychological factors). On the other hand, a negative decision can only be made when the *slowest* horse has crossed the finish line. The more horses there are in the race, the slower the slowest one will be, on the average. Another example is illustrated by the problem of collecting a group of people for a committee meeting or for dinner at a restaurant. The larger the group of people who must assemble together before the event can begin, the longer it takes for the last person to arrive. Both of these are examples of the general rule that the larger the sample, the more likely it is, on purely statistical grounds, to have a very large or small value in that sample.

Although the way in which the largest or maximum value in a distribution increases with n depends on the theoretical distribution assumed (see Gumbel, 1958), the value of the maximum generally increased more slowly than n. Thus, the parallel self-terminating search model predicts that negative responses will increase with set size, but at a negatively accelerated rate, as shown in the negative response function of Figure 5-2A. Because we observe a very different pattern of experimental results, we will have to conclude that this *parallel self-terminating model* is not a good description of a subject's strategy in this task. (Here we consider only a simple parallel model for this task. There are more complex parallel models that give other patterns of results, but their discussion lies beyond the scope of this chapter. These models are discussed in Snodgrass & Townsend, 1980.)

Serial Self-Terminating Search. A second possibility is that subjects search the items in their short-term memory by considering the items one by one. This is known as *serial processing* or *serial search.* Suppose that they follow a self-terminating rule and on positive trials search through the list of memory items until they find the one that matches the probe. On some trials, the matching item will be found immediately, with the first comparison; on other trials, it will be found with the second comparison; and on still others, it will be found on the last comparison. On the average, subjects will need to search through $(n + 1)/2$ items to find the matching item on positive trials. So for a memory set of three, they will need to search through an average of two, for a memory set size of five, they will need to search through an average of three, and so on.

The way this works is as follows: Imagine there are three items in memory, and the subject searches them in the order they were presented on a trial. Because items are probed across serial positions an equal number of times, on one third of the trials, the target will be found with the first comparison; on one third of the trials the target will be found with the second comparison; and on one third of the trials the target will be found with the third comparison. The average number of comparisons needed for a memory set size of three is $(1 + 2 + 3)/3$, or an average of two. In general, the average number of comparisons that need to be carried out for a memory set size of n is given by summing the digits $1 \ldots n$ and dividing by n. This formula can be simplified to the form $(n + 1)/2$ given above.

On negative trials, on the other hand, the subject will need to search through all the memory list items before concluding that the probe item is not a member of the positive set. Therefore, the subject always needs to search through n items before deciding the item was not there. The combination of these two strategies—a terminating rule on positive trials and an exhaustive rule on negative trials—is called a *serial self-terminating search.*

The crucial prediction for this type of search is that the time for positive responses will increase at a *slower* rate than the time for negative responses as memory set size is increased. Increasing the memory set size by two items, from three to five, will only increase the average number of items searched by one (from two to three) for positive trials, but will increase the number of items searched by two (from

three to five) for negative trials. In fact, the slope of the positive response function should be half the slope of the negative response function, as shown in Figure 5-2B. Because the experimental RT functions do not show this pattern, we also reject the serial self-terminating strategy as a reasonable description of what subjects do in this experiment.

Serial Exhaustive Search. A third possibility, which is consistent with the pattern of empirical results, is that on both positive and negative trials subjects search through the entire list of items. This is termed *serial exhaustive search* because on positive trials, the subject exhaustively examines all items in the set, even though an exhaustive search is not logically required. In serial exhaustive search, increasing the number of items in the memory set will have exactly the same effects on both positive and negative reaction times (RTs), so the two RT functions will be parallel, as shown in Figure 5-2C. This is consistent with the pattern of observed data, so we conclude the subject must be using the serial exhaustive search strategy for this task.

Later we consider why subjects might adopt the strategy of serial exhaustive search, which, on the surface, appears to be inefficient. However, for now, let us introduce the stages that Sternberg has inferred must exist in this task (Sternberg, 1966, 1969a, 1969b, 1975). These are shown in Figure 5-3.

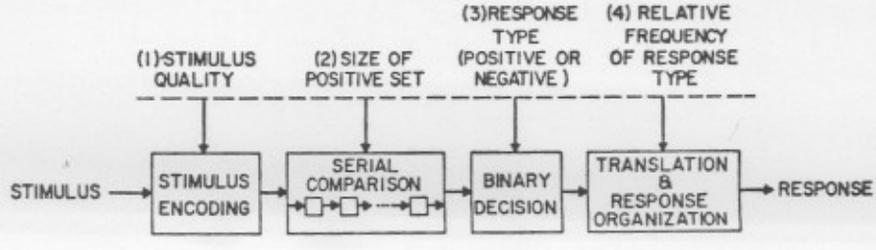
Figure 5-3 shows the four stages that are hypothesized to intervene between the presentation of the stimulus (probe item) and the recording of the response (key press). Stage 1 is *stimulus encoding*, by which is meant the process of perceiving the stimulus and representing it in some way so that it can be compared with the items in memory. The nature of this representation is unclear, although it is probably not in an acoustic or articulatory form. Stage 2 is *serial comparison* of the encoded representation of the probe with the items in memory. As noted previously, this is assumed to be a serial exhaustive search, so the number of comparisons for both positive and negative probes is identical and equal to n. Stage 3 is *binary decision*—the decision about whether or not the probe was in the memory set. This decision is fed to Stage 4, which is *response organization and execution.*

It should be noted that these stages themselves are serial and additive, in that one stage does not begin until the previous one has finished. Also, no stages are ever experimentally deleted in the additive factors method. Rather, variables are manipulated which are assumed to selectively affect one or more stages.

Let us now try to interpret the data graphed in Figure 5-1 in terms of the stages shown in Figure 5-3. First, note that for this hypothetical experiment, there were only two independent variables: (1) the number of items in the memory set, which varies from one to six, and (2) whether the probe was a member of the positive set or not, which occurred with equal probability.

Disregarding for the moment the distinction between positive and negative responses, the increase in RT with set size is assumed to occur entirely in stage 2, the serial comparison stage. As the number of items is varied from one to six, the number of serial comparisons that are required increases from one to six. Therefore, it is

Fig. 5-3 The four stages proposed by Sternberg for the memory-scanning task.



possible to use the *slopes* of the RT functions to estimate the time for each serial comparison (recall that the slope of a straight line represents the increase in the *Y* variable, or RT, for each unit increase in the *X* variable, or set size). For the data in Figure 5-1, we conclude that the time it takes to compare the encoded version of the probe to a single item in memory takes approximately 40 ms, because the slopes for both RT functions are 40 ms. Stated another way, the number of comparisons that can be made in a second is equal to 1000/40 or 25.

This is an extremely rapid rate of search, and its very rapidity gives us clues to two puzzles that we were unable to resolve earlier in the chapter: (1) why should the search be exhaustive and (2) what is the form of the encoded probe? The search is apparently exhaustive because the comparison process itself is so fast that it is more efficient to complete the search through all the items and then determine whether the probe item has matched any of the items than to stop after each comparison and make a decision. That is, taking the time to decide whether or not the probe matches each item may take much longer than making the comparison on which the decision is based (see Sternberg, 1975).

Before leaving the topic of the exhaustiveness of the short-term memory search, we point out here that evidence for serial exhaustive search is not universally found in all search tasks. For example, Egeth, Jonides, and Wall (1972) and Neisser (1963; Neisser, Novick, & Lazar, 1963) found evidence for parallel search through visual displays, Atkinson and Juola (1973) found evidence for serial self-terminating search through long-term memory, and Theios, Smith, Haviland, Traupmann, and Moy (1973) have argued in favor of a serial self-terminating model for memory scanning. Schneider and Shiffrin (1977) and Fisk and Schneider (1983) have shown that extensive practice with memory scanning when the positive and negative items remain constant produces parallel search through both memory and visual display sets. In contrast, they found that changing the assignment of items to positive and negative sets produces serial self-terminating search regardless of practice.

The second puzzle, how the probe is encoded, is also clarified somewhat by the rate of comparison. We are fairly sure the form of encoding is not acoustic or articulatory (e.g., pronouncing each member of the memory set to yourself to see whether it matches the probe) because we know from other studies that it takes much longer than the scan rate of 40 ms to subvocalize well-learned sequences, such as the letters of the alphabet. Landauer (1962) found that subvocalization rates were about 170–200 ms per item for these tasks. Accordingly, the form of encoding must be different from implicit speech, although whether it is a visual image or some more abstract form is not known.

What can we say about the intercept of the RT functions in Figure 5-1? Recall that the intercept of a linear function is that value of *Y* when *X* is zero. Thus, for the RT functions in Figure 5-1, the intercept corresponds to the RT when the set size is zero—that is, when there are *zero* items in memory to be searched. We cannot observe the intercept RT-value empirically—that is, we cannot include a condition in the experiment in which the subject never searches memory. Rather, we infer the value of the zero-intercept statistically. In the memory scanning paradigm, the intercept of the RT function must represent the sum of the times of stage 1, stage 3, and stage 4, since it theoretically represents a case in which there are zero serial comparisons to be made (no stage 2). Now we consider the fact that the intercepts of the positive and negative functions are not the same but differ by 50 ms. (This advantage of positive over negative reaction times when their probabilities are the same is a pervasive finding, although the size of the difference is not as constant as the size of the slope.) This difference between the intercept of positive and negative responses is attributed by Sternberg to the binary decision process of stage 3; it takes longer to decide in favor of a negative than a positive decision.

Factors That Affect Other Stages in Memory Scanning

So far, we have considered the effects of memory set size and type of response (positive versus negative) on two of the four stages. Here we consider the effects of two additional variables—stimulus probe degradation and response probability—on the remaining two stages of stimulus encoding (stage 1), and response organization and execution (stage 4). By *stimulus probe degradation* we simply mean some manipulation that makes the probe more difficult to see.

We might expect that stimulus degradation would affect the stage of stimulus encoding by making it longer. We might also wonder whether it would affect the second stage, stimulus comparison or search as well. Sternberg (1967) compared reaction times for intact visual probes with degraded probes, in which the degradation was accomplished by superimposing a checkerboard pattern on the probe. He found that after some practice, subjects showed a higher intercept for both positive and negative responses for a degraded than for an intact probe (the average difference was about 65 ms), but no change in slope. He interpreted this as meaning that the degradation affected only stage 1 (stimulus encoding) but not stage 2 (comparison). He reasoned that if the comparison process had been slowed by the degradation of the probe, then the slopes of the RT functions would have increased. It is interesting that degrading the visual quality of the probe did not increase the comparison times. This suggests that if comparisons of visual images underlie the serial comparison stage, the encoded version of the probe must be processed to make it equivalent to a nondegraded probe.

We expect response probability to affect the last stage, that of response organization and execution. Many RT experiments show that when response probability in a choice reaction time situation is varied by varying stimulus probability, the more probable response is executed faster and the less probable response is executed more slowly. Sternberg (1969a) manipulated the presentation probability of positive probes (and hence the complementary presentation probability of negative probes) from .25 to .75. He found that presentation probability affected only the intercept of the RT functions, and not their slopes, with the more probable response having a lower intercept (overall faster times regardless of memory set size), and the less probable response having a higher intercept (overall slower RTs regardless of memory set size). This effect is located by Sternberg in stage 4 because response probability does not interact with response type (YES versus NO), and the lack of an interaction between two variables when they are applied together is used as an indication that these two variables do not affect the same stage.

In summary, Sternberg's additive factors method is a type of subtraction method. Here, however, rather than deleting whole stages, variables are manipulated so that differences in RT between different levels of the same independent variable are used as measures of the duration of substages of the major stages. Thus, we use the subtraction method in additive factors to measure substage duration, rather than stage duration.

POSNER'S SAME/DIFFERENT CLASSIFICATION TASK

Michael Posner and his colleagues have used a same/different classification task to isolate and measure components of comparison times. In the Posner paradigm, subjects are asked to classify pairs of stimuli as SAME or DIFFERENT on the basis of some criterion. The classification criteria may vary in abstractness from physical identity, to name identity, to the most abstract level of category or rule identity. As abstractness increases, so do the number of stimuli that are to be considered identical. Posner's method allows us to study the classification problem under experimentally rigorous conditions by using stimuli

that are simple and well-learned, yet can be classified by a variety of criteria.

The Letter-Matching Task

The basic paradigm is a letter-matching task, in which the subject is presented with a pair of identical or similar stimuli, such as letters of the alphabet, and is required to judge as quickly as possible whether the pair is the same or different. The basic data are the reaction times required to carry out the task.

In a typical letter-matching task (Posner, 1969; Posner & Keele, 1968), a name-identity criterion is used to classify letters as same. Two letters are shown either simultaneously or successively. The letters may be physically identical (AA, aa), have the same name but different physical forms (Aa, aA), or be different (AB, ab). The subject's task is to classify all the letters with the same name as same and those with different names as different. Thus, both physically identical and name-identical pairs are to be classified as SAME. Figure 5-4 illustrates the name-identical criterion for same judgments under simultaneous and successive presentation conditions.

Well-practiced subjects responding to simultaneously presented letter pairs show a 70- to 100-ms advantage in matching physically identical over name-identical pairs. This result is taken as evidence that subjects match physically identical stimuli on the basis of visual rather than name characteristics even for such familiar stimuli as letters of the alphabet.

Posner (1969) points to other lines of converging evidence for this conclusion. These include the fact that letter-like stimuli without names are matched as fast as physically identical letters and that inverted letters are matched as quickly as upright letters as long as the inverted letters are close to one another in the visual field. The difference in time between physical and name matches also suggests that it takes between 70 and 100 ms to convert a pair of letters into a name code (or alternatively, to convert one of the letters into the opposite case).

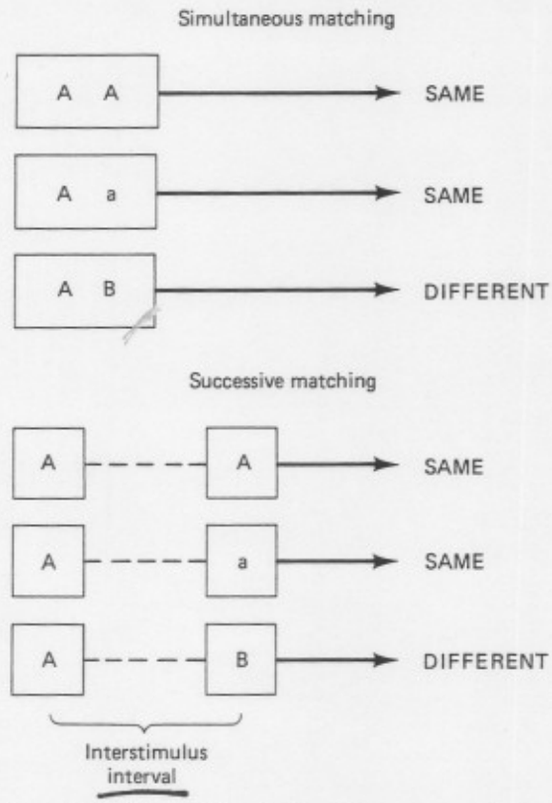


Fig. 5-4 Examples of simultaneous and successive matching in a Posner task with a name-identical criterion.

Another line of converging evidence for the visual basis of physical matches comes from results of experiments studying successive matching (Posner, Boies, Eichelman, & Taylor, 1969). What happens to the advantage of physical matching when there is an interval between the first and second letter in a pair, so that the matching decision must be based on memory for the first letter? Is that memory based on the visual appearance of the first letter or on its name? Figure 5-5 summarizes the results from two experiments in Posner et al. (1969) by plotting differences in matching-time between physically identical and name-identical letter pairs as a function of the interstimulus interval. Time differences rather than absolute times are shown because the absolute times vary widely between the two experiments because of differences in viewing conditions. The zero condition represents the simultaneous presentation condition, and results in about a 90 ms advantage of physically identical over name-identical pairs. However, as the

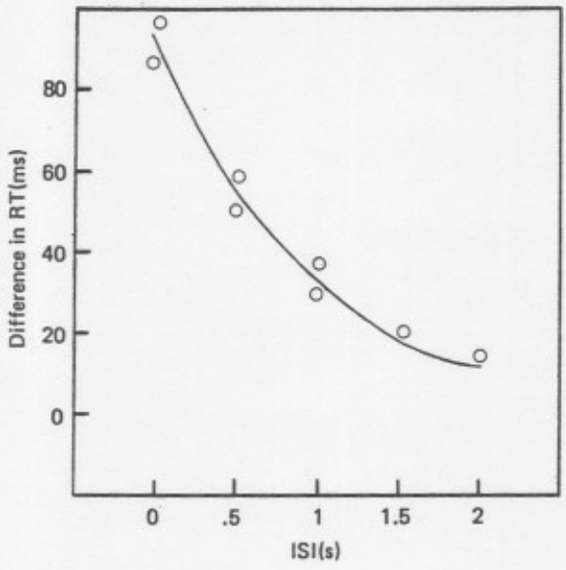


Fig. 5-5 Difference in RT between name and physical identity SAME responses as a function of *ISI* between two successive letters. The open and solid circles represent two different experiments. (From "Retention of Visual and Name Codes of Single Letters," by M. I. Posner, S. J. Boies, W. H. Eichelman, and R. L. Taylor. In *Journal of Experimental Psychology,* 1969, *79* (Monograph Suppl. 1), 1–16. Copyright 1969 by the American Psychological Association. Reprinted by permission.)

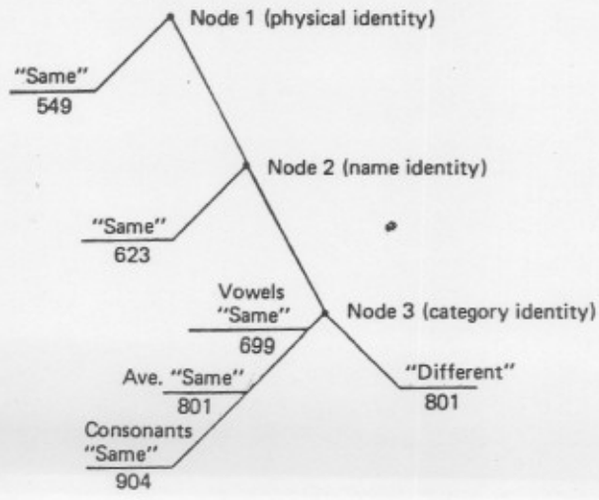
interstimulus interval increases, the advantage gradually disappears and is completely absent after an interval of about 2 seconds.

This suggests that subjects after that interval are matching letters on the basis of their names, since a match like AA is made no faster than a match like Aa. Thus, the duration of an efficient visual code for matching appears to be short. Other experiments by Posner and his colleagues have shown that the visual memory code is affected by other attentional demands, since interpolation of a cognitive activity such as addition interferes with the visual matching process but not with the name matching process (Posner, 1969).

Posner and Mitchell (1967) extended the letter-matching paradigm to a matching task involving category (or rule) identity. Subjects were presented with pairs of letters presented simultaneously and asked to classify them as same if they were both vowels or both consonants and as different otherwise. Pairs of same letters could also have the same name or could be physically identical. In this task, letters that were physically identical retained their advantage over those with the same name, but same name pairs were matched faster than those with the same rule of classification (i.e., letters that were both vowels or both consonants).

The times for the four types of same judgments are shown in Figure 5-6, along with the times for different responses. Physical identity matches are about 70 ms faster than name identity matches, which in turn are about 80 ms faster than vowel matches. However, vowel matches are more than 200 ms faster than consonant matches. The difficulty subjects have in deciding whether two letters are both consonants presumably is because there are many more consonants than vowels. In fact, Posner and Mitchell present evidence that the consonant match decisions were

Fig. 5-6 Hypothetical series of stages in Posner and Mitchell's study in which subjects classified letter pairs as same category (both vowels or both consonants) or different category (one of each). Numbers represent time in milliseconds. Node 1 refers to physically identical pairs (AA); node 2 to name identical pairs (Aa), and node 3 to category identical pairs (Ae for vowels and BD for consonants). (From "Chronometric Analysis of Classification," by M. I. Posner and R. F. Mitchell. In *Psychological Review,* 1967, *74,* 393–409. Copyright 1967 by the American Psychological Association. Adapted by permission.)



made by exclusion—that is, by checking that neither letter was a vowel.

One way of interpreting the results presented in Figure 5-6 is by assuming that subjects must proceed through all the processing "nodes" in serial order to make the most complex decision. Thus, to decide that B and D are both consonants, a subject checks the physical identity node (node 1), next the name identity node (node 2), next the vowel identity node (node 3), and finally the consonant identity node. If the processing is serial, differences between times for complex decisions and those for simpler decisions reflect times of the interpolated processes. Posner and Mitchell (1967) acknowledge that processes other than serial ones can be invoked to account for their results. Nonetheless, their simple and elegant experiments provide a clear example of another application of the subtraction method in mental chronometry.

Although the use of reaction time to measure the duration of mental events still has a long way to go, the introduction of the additive factors and classification methodologies into the arsenal of the cognitive psychologist has brought the ultimate goal of unraveling mental processes much closer.

METHODOLOGICAL ISSUES IN REACTION TIME

There are a number of methodological issues in the use of reaction time as a dependent variable. Here we survey some of them.

What Is the Minimum Reaction Time?

RT investigators generally assume that there is some minimum RT in simple and choice tasks below which subjects cannot respond. For simple reaction time, this represents the minimum time for stimulus input, decision, and motor response time and has been termed the *irreducible minimum reaction time*. Estimates of this time vary with conditions and subjects, but for auditory stimuli, which produce faster reaction times than visual stimuli, the irreducible minimum is generally estimated to be between 90 and 100 ms. (Snodgrass, 1969; Woodworth, 1938).* Reaction times shorter than the irreducible minimum are called *anticipations*. These may occur if the subject knows with some certainty when the reaction stimulus will occur. Usually, subjects are warned about the presentation of a stimulus by a warning signal that occurs some time before the reaction stimulus.

The first stimulus, the *warning signal*, indicates to the subject that the second stimulus, the *reaction signal*, will occur after some time interval. This time interval is known as the *foreperiod*, and may be fixed in duration or may vary randomly from trial to trial. When the foreperiod is fixed, the subject very quickly comes to know its value. For fixed foreperiods, the optimum duration is between 1 and 2 seconds (Karlin, 1959; Woodrow, 1914). Shorter foreperiods do not permit the subject to prepare sufficiently, and longer ones are too long for subjects to maintain optimum readiness. However, a disadvantage of fixed foreperiods is the large temptation for the subject to anticipate the signal and try to respond just after the signal. Although there are experimental methods for controlling anticipations (Snodgrass, 1969; Snodgrass, Luce, & Galanter, 1967), these are quite time-consuming, so most investigators solve the problem of anticipations by discarding those RTs that are shorter than the irreducible minimum. This minimum is usually taken to be about 100 ms for simple reaction time experiments.

The anticipation problem is less critical for choice reaction time, because subjects cannot respond correctly with more than chance performance until the stimulus has been identified. Thus, investigators who use choice reaction time do not usually need to discard very short RTs.

*In New York City, the irreducible minimum RT to a light stimulus may be defined as the time between a traffic light's turning green and the honk of the taxicab behind you. However, this particular situation has not been investigated in detail.

The Problem of Very Long RTs (Outliers)

The problem of long RTs, or *outliers*, is characteristic of both simple and choice RTs. Very short RTs (or anticipations) are also outliers, but long outliers are more of a problem because they can affect the mean RT more than spuriously short RTs. Furthermore, there are reasons for expecting subjects to produce occasional long RTs. Often subjects report that on that particular trial their attention wandered or they momentarily forgot which key to press for their response. This resulted in a long RT that was outside the normal range.

Accordingly, investigators adopt one of three strategies for dealing with outliers: (a) throw out or (b) replace those that are unreasonably long (we will come back in a moment to the definition of "unreasonable"), or (c) use a measure of central tendency that is less sensitive to outlying observations than the arithmetic mean, such as the median.

When discarding or replacing outliers, investigators may use either a relative criterion, so that an outlier is defined for each subject with respect to the other RT values in that subject's distribution, or an absolute criterion, so that an outlier is defined as any response greater than k ms. The value of k may be as low as 500 ms for simple reaction time or as high as 3000 ms for choice reaction time.

A commonly used relative criterion is to discard any response that lies more than three standard deviations from that subject's mean for that condition. This criterion is based on a model for RT that assumes RT values are normally distributed. Another widely used criterion is the Dixon test, which is a statistical method for determining the likelihood that a suspected outlier comes from another distribution of RTs. In the Dixon test, we compute the difference between the suspected outlier and the next largest observation, and compare that difference with the entire range of RTs. The Dixon test is described more fully in Chapter 15 on descriptive statistics. Discarding observations from a distribution is known as *trimming*. An alternative to discarding outliers is to replace them with the next largest or smallest value. This procedure is known as *Winsorizing* and it, too, is described in Chapter 15.

The most commonly used alternative measures of central tendency are the median and the geometric mean. The median or 50th percentile is the middle value of a distribution of RTs, and as such is completely insensitive to extreme observations. For example, the median for the set of observations 140, 150, 160, 170, 180 is the same—160—as the median for the set of observations 140, 150, 160, 170, 400. The median is insensitive to the replacement of the value of 180 in the first set by the value of 400 in the second set; in contrast the arithmetic mean is very sensitive to the outlier, being 160 for the first set and 204 for the second.

The geometric mean is the nth root of the product of n scores. It, like the median, has the property that it is insensitive to extremely long scores. For example, the geometric mean of the series 10, 100, 1000 is 100, while the arithmetic mean is 370. Using the geometric mean as a measure of central tendency is the same as using a logarithmic transformation on the data and then computing the arithmetic mean on the logarithmic transforms. So, for example, the logs to the base 10 of the series 10, 100, 1000 are 1, 2, and 3, and the arithmetic mean of the log transforms, 2, represents the log of the geometric mean. The usual way to calculate a geometric mean is to convert each score into its logarithm, compute the arithmetic mean of the logarithms of the scores, and then find the antilogarithm of that mean.

The use of medians or geometric means is not appropriate when an experimenter is interested in testing an additive type of model, such as Sternberg's, for RT data. The reason these transformations are not appropriate is that the stages that are hypothesized to underly the total RT are additive, and these transformations do not preserve their additivity.

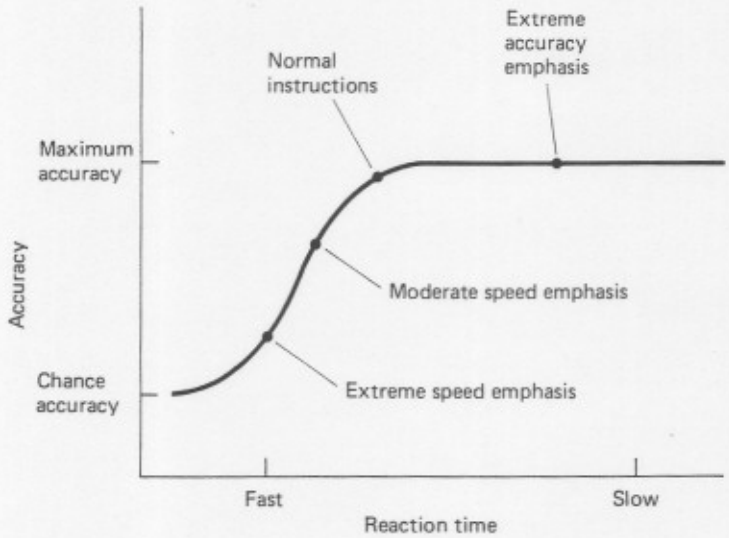


Fig. 5-7 An idealized speed-accuracy trade-off function.

Error Rates and the Speed-Accuracy Trade-Off Function

When RTs are used as the dependent variable, subjects are usually instructed to respond as quickly as possible, consistent with an extremely low (ideally 0%) error rate. Thus, the average RT should represent an ideal performance for that subject for that condition.

However, subjects do make errors. In simple RT experiments, these errors may be anticipations, extremely slow RTs, or omissions. In choice RT experiments, an additional source of error is a response of type B to stimulus A or a response of type A to stimulus B. Regardless of how well practiced subjects are on a choice RT task, there are invariably some small percentage of trials (1 to 2% in the best cases) on which subjects respond to a stimulus with the wrong response. Subjects often explain their errors by saying something like, "I wasn't paying attention on that trial—my mind wandered" or "I forgot momentarily that the right hand response was the correct one." Most investigators merely omit the error RTs from the analysis and report the error rates along with the correct average RTs.

However, it is clear that subjects are capable of manipulating their speed of response by manipulating how many errors they make. For example, if they take a long time to respond, they can approach perfect performance. On the other hand, if they respond too quickly to fully process the stimulus, they can produce very fast RTs but at the cost of lower accuracy. The relationship between RT and accuracy (with increases in RT leading to increases in accuracy) is called the *speed-accuracy trade-off function*, and an idealized one is presented in Figure 5-7.

Here, accuracy of performance is plotted along the ordinate (y-axis) against average RT along the abscissa (x-axis). If subjects are encouraged to produce very fast RTs (with instructions emphasizing extreme speed), they will also produce low accuracy or a large percentage of errors. If they are given normal instructions, their RTs will increase along with their accuracy. Subjects may differ on where they place themselves along the speed-accuracy operating function, depending on how they interpret the instructions, their own personal biases about which is more important—speed or accuracy—and so on. We might use as an analogy taking a timed achievement test. Some people will race through such tests, preferring to answer all the questions quickly at the expense of some accuracy. Others will take their time and may not answer all questions but will be more accurate on the questions they do answer.

Some investigators (e.g., Pachella, 1974) study the form of the speed-accuracy trade-off function explicitly, by keeping the experimental situation constant and varying payoffs to selectively reward ac-

curacy at the expense of speed (so that subjects adopt a conservative criterion), or to reward speed at the expense of accuracy (so that subjects adopt a liberal criterion). More usually, however, investigators who are interested in the effect of one or more independent variables on RT adopt a standard set of instructions that they hope will produce similar criteria for speed versus accuracy across subjects and conditions.

Thus, most investigators do not worry about the form of the speed-accuracy trade-off function as long as the results from their conditions do not show such a trade-off. When experimenters compare RTs across a number of different conditions, some conditions are usually hypothesized to be more difficult in that they require more processing steps or more complex processing than other conditions. As long as error rates are positively correlated with RT, so that low error rates accompany fast RTs and high error rates accompany slow RTs, the investigator may feel fairly confident that the pattern of RTs cannot be accounted for by a speed-accuracy trade-off. On the other hand, if error rates are negatively correlated with RT, then a speed-accuracy trade-off is a possible alternative explanation to the investigator's preferred hypothesis.

Table 5-2 shows RT and error results from an experiment in which the pattern of RT results *cannot* be explained by speed-accuracy trade-off, and Table 5-3 shows a pattern that *can* be explained by speed-accuracy trade-off.

Both experiments show hypothetical results from a same-different RT experiment in which subjects were presented with letter pairs selected either from a familiar, well-learned class (English letters) or from an unfamiliar, poorly learned class (Greek letters). Within each class, letter pairs could be the same by being physically identical—capital A's (AA) and capital gammas (ΓΓ)—or could be the same by having the same name—a capital and lower-case A (Aa) and a capital and lower-case gamma (Γγ). The investigator hypothesized that physically identical matches would be faster than name-identical matches (e.g., Posner, 1969), and that familiar letter pairs would be faster than unfamiliar letter pairs. (Note that subjects must be familiar enough with the unfamiliar class of Greek letters to know when the letters have the same names.)

Table 5-2 shows just the expected pattern of results: familiar physical matches are faster than unfamiliar physical matches, and unfamiliar name matches are longest of all. The rates of error responses (calling a same pair DIFFERENT) correlate positively with RT in the sense that fast RTs are accompanied by low error rates and slow RTs are accompanied by high error rates, so the RT results cannot be the result of a speed-accuracy trade-off.

Table 5-3 shows the same RT data, but now error rates correlate negatively with RT—fast RTs are accompanied by high error rates and slow RTs are accompanied by low error rates. This pattern of RTs could be explained by a speed-accuracy trade-off. In this particular case, it would be explained by assuming that in the hypothetically "easy" conditions, subjects

Table 5-2 Correct RTs and Error Rates for Familiar and Unfamiliar Letter Pairs in a Same/Different RT Experiment for which no Speed-Accuracy Trade-Off Is Obtained (Only Same RTs Are Shown)

	Mean RT	Error %
Familiar letters (English)		
Physically identical (AA, aa)	400	2
Name identical (Aa, aA)	550	5
Unfamiliar letters (Greek)		
Physically identical (ΓΓ, γγ)	500	3
Name identical (Γγ, γΓ)	700	10

Table 5-3 Correct RTs and Error Rates for Familiar and Unfamiliar Letter Pairs in a Same/Different RT Experiment for which a Speed-Accuracy Trade-Off *Is* Obtained

	Mean RT	Error %
Familiar letters (English)		
Physically identical (AA, aa)	400	10
Name identical (Aa, aA)	550	3
Unfamiliar letters (Greek)		
Physically identical (ΓΓ, γγ)	500	5
Name identical (Γγ, γΓ)	700	2

had a bias toward responding as quickly as possible, thereby producing both fast RTs and high error rates, whereas the opposite occurred for the hypothetically "difficult" conditions. On the basis of Table 5-3, it is impossible to reject the notion that the RTs are really the same across the "easy" and "difficult" conditions, or even that the RTs might run in the opposite direction if error rates were kept constant across the conditions.

SUMMARY

Mental chronometry is the measurement of mental processes by the use of reaction time (RT). Donders pioneered the use of RT to measure mental processes with his subtraction method. However, various problems with Donders' method led to the ultimate rejection of his use of the subtraction method, in part because it involved changing the experimental task so that entire stages were omitted or added. Sternberg's additive factors method revived the subtraction method by manipulating variables that affect various hypothetical stages assumed to underlie the RT, without either adding or omitting stages. We have presented one widely used application of the additive factors method here, namely, its use in short-term memory scanning. By use of this method, Sternberg discovered (a) that scanning through short-term memory was serial and exhaustive, (b) that degradation of the stimulus probe affects the encoding stage but not the search stage of the process, and (c) that factors such as response probability affect only the last stage—response organization and execution—and not the binary decision stage.

Posner's experiments, which apply mental chronometry to classification of letters, showed that physical matching is faster than name matching, which in turn is faster than rule-based matching of vowels and consonants. There is evidence that the extraction of these three levels of information proceeds in a serial fashion, and thus that differences between the RTs for the three matching conditions is a measure of the time for such processing. The advantage of physical matching disappears when the two letters are separated by an interval of two seconds or longer, suggesting that the visual code that is the basis for matching at a physical level is of short duration.

A number of methodological issues in reaction time were discussed to illustrate the correct treatment of RT data. These include the issue of outlying observations, both those that are very short and thus may be produced by anticipating the reaction signal, and those that are very long and thus may be produced by a different process than that under study, such as inattention or response forgetting. Finally, the problem of speed-accuracy trade-offs in RT was discussed.

Reaction time as a method of measuring the speed of mental processes has become increasingly important in recent years and is the dependent variable of choice in many areas. It is important to understand the underlying logic in using reaction time to dissect the workings of the human mind and to be alert to possible confounding variables in reaction time research.