Sperling, G. A unified theory of attention and signal detection. In R. Parasuraman and D. R. Davies (Eds.), *Varieties of Attention.* New York, N. Y.: Academic Press, 1984. Pp. 103-181.

4

A Unified Theory of Attention and Signal Detection[1]

George Sperling

Introduction

The plan of this chapter is to review some experiments on attention, mostly from my laboratory, and to show how all these experiments and experiments on signal detection are subsumed under the same theoretical description.[2] The kind of attention tasks under consideration are exemplified by the concurrent tasks of driving an automobile while listening to a news broadcast on the radio. Can a driver do both tasks simultaneously without loss? Or does driving suffer when too much attention is paid to the news? Or is memory for some news events lost because of momentary concentration on a traffic obstacle?

The concurrent tasks of driving and listening are prototypical of the ones under consideration here. But driving–listening is complicated because the difficulty of each task varies from moment to moment and because difficulty depends not only on the present stimuli but also on previous stimuli and on previous responses to previous stimuli. Before a good understanding of such a complex situation is possible, it is necessary to understand some simpler ones. Considered first are some simpler (but not simple!) visual concurrent tasks in which both tasks involve visual stimuli and in which the relevant stimuli are condensed into essentially an instantaneous flash. The interest here is in examining how visual attention is distributed over the visual field in a single instant of time and in how long it takes to shift attention within the visual field. In the course of this examination, it will become useful to discard "attention" as an explanatory concept (while retaining it as a description of the situation) and to replace it with "processing resource." A calculus for processing resources is developed.

[1] Preparation of this article was made possible by United States Air Force, Life Sciences Directorate, Grant AFOSR 80–0279.

[2] The theory of attention (concurrent–compound tasks, operating characteristics, optimization) was first published in Sperling and Melchner (1976a, 1978b) and elaborated in two hitherto unpublished talks (Sperling, Note 1, Note 2).

ISBN 0-12-544970-4

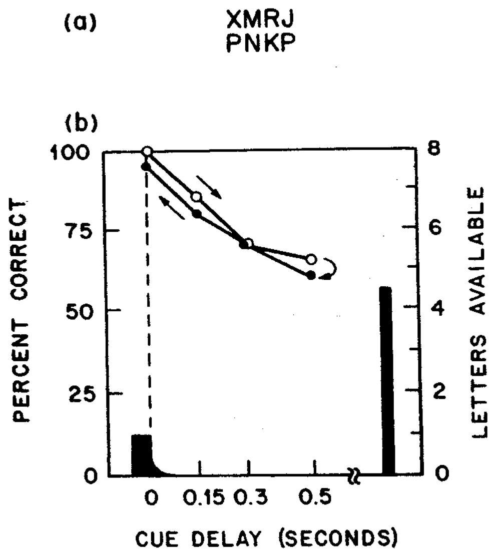


Figure 4.1 (a) Stimulus configuration for a two-alternative partial-report procedure, and (b) partial-report accuracy as a function of cue delay. The right-hand ordinate, letters available, is the mean fraction of letters correct times the number of letters in the stimulus. The mean number of letters in a whole report is shown in the bar on the right. The stimulus is indicated on the abscissa, beginning at cue delay −0.05 seconds and ending, except for a very brief decay, at 0.0 seconds. Arrows indicate the sequence in which blocks of trials were conducted. Data are shown for one subject, ROR. (From Sperling, 1960.)

Experiments

Three Experiments in Visual Attention

Partial-Report Procedure

Estimating the Capacity of Visual Information Storage (VIS) The partial-report task was originally used by Sperling (1959, 1960) to demonstrate and measure the duration of short-term visual information storage—subsequently called *iconic memory* by Neisser (1966). A subject views an array, for example, two rows of four letters, exposed very briefly so that he or she cannot make an eye movement during the exposure. After the exposure, a cue (for example, a high- or low-pitched tone) directs the subject to report the letters of either the upper or the lower row (partial report). In control conditions, all letters of both rows must be reported (whole report).

In the whole-report condition, subjects typically report less than 5 letters correctly. In the partial-report condition, they can report about 3.5 of the 4 letters in the cued row, provided that the tonal cue follows the exposure within a few tenths of a second (Figure 4.1). Because the choice of tonal cue is random from trial to trial and unknown to the subject, we infer that the subject must have available 7 of the 8 stimulus letters at the time of the cue in order to maintain such a high accuracy of partial report. These available letters constitute visual information storage (VIS). In similar experiments involving larger arrays with more letters (and more tonal signals), estimates of the capacity of VIS have been as high as 17 (of 18) letters (Sperling, 1963). Estimates of VIS could be made even higher if there were some reason to construct still larger stimulus arrays and corresponding sets of cue signals. Typically, VIS decays within a few tenths of a second, but may not do so for seconds under conditions that produce long-lasting afterimages.

The Role of Attention in Partial Report Procedures In traditional terms, the cue directs the subject to attend to a particular row and to memorize the letters of just that row for later recall. A subject cannot memorize more than four or five letters from a brief exposure, but as long as VIS exceeds the subject's short-term memory (STM) capacity, the subject has a choice of what to memorize; attention determines that choice.

In contemporary terms, it might be said that the subject has several *information processing resources*[3] available: VIS, STM, and a transfer process from VIS to STM. The transfer process is under voluntary control—the cue to attend to (and to report) a particular row being translated into control instructions for the transfer process. The *limiting resource* is the limited five-item capacity of STM. At this point, I have simply introduced novel words for familiar processes. But there is a lot more to the analysis of the partial report-procedure, and this example is considered again later.

Visual Search

In their classic experiments, Neisser and his collaborators (Neisser, 1963; Neisser, Novick, & Lazar, 1963) studied the ability of subjects to find a particular target character or characters embedded in long lists of randomly chosen characters. Subjects searched lists from top to bottom and made a manual response when they detected the target. Neisser *et al.*'s fastest reported search times were on the order of 20 msec per distractor (nontarget) character. For example, if the target were the thousand and first character on the list, it would take the subject about 20 seconds longer to discover the target than if it were the first character on the list. Unfortunately, Neisser *et al.*'s calculated search times per character were not consistent between lists having different spatial arrangements of characters.

Eye Movements in Visual Search To investigate the conjecture that eye movements might have been a limiting factor in Neisser *et al.*'s visual search, a computer-driven display was devised to enable visual search to proceed without

[3] The term information processing resource was first introduced by Norman and Bobrow (1975).

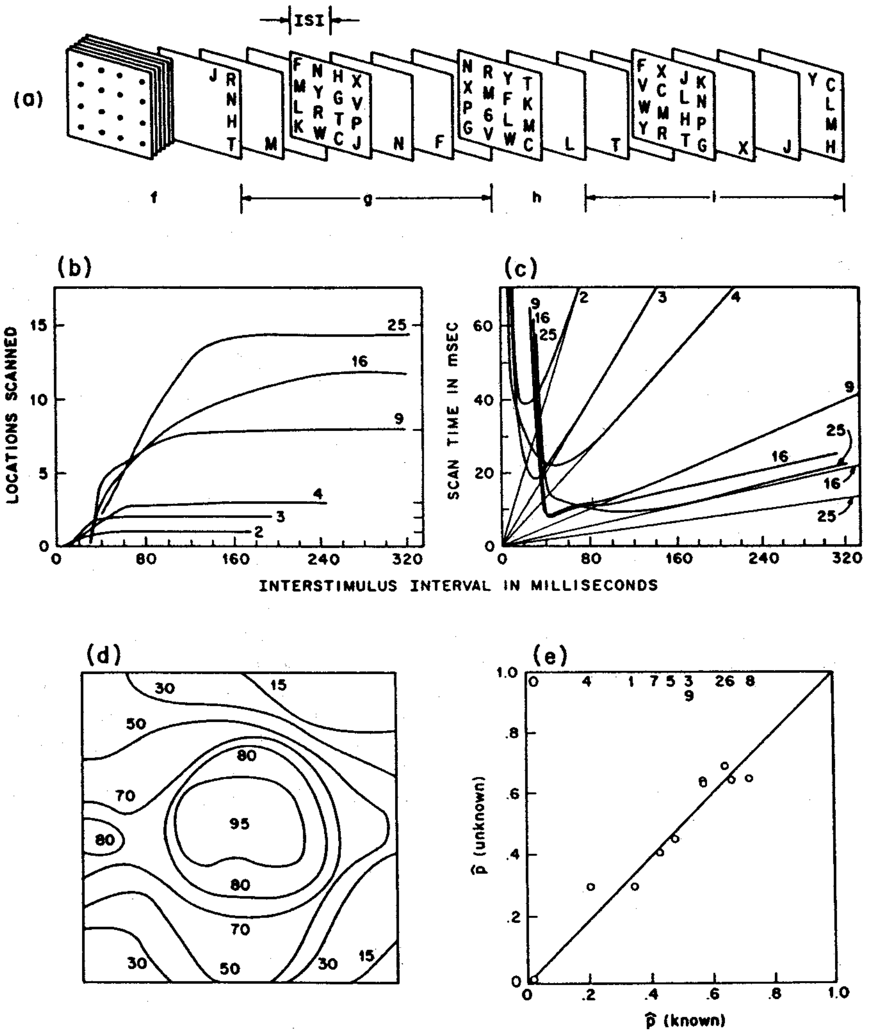


Figure 4.2 Visual search without eye movements in computer-generated displays. (a) The stimulus—fixation point (f), random number (6–12) of displays containing only letters (g), critical display containing a numeral target (h), and 12 more nontarget displays (i). (b) Number of locations (ℓ) searched as a function of interstimulus interval (ISI), the parameter is the number of letters in the display, ℓ is corrected for guessing. (c) The scan time T derived from (b): T = ISI/ℓ. (d) Search-field contours in a 7 × 7 array. The parameter shown is search accuracy at the contour. (e) Comparison of search accuracy with known numeral targets [$\hat{p}$ (known)] to search with unknown numeral targets [$\hat{p}$ (unknown)]. Numerals at the top indicate the identity of the target points plotted, and the line through the data has a slope of 1.0, accounting for 97% of the variance. (Panel d based on Sperling and Melchner, 1978b; Panels a–c, and e based on Sperling, Budiansky, Spivak and Johnson, 1971.)

eye movements (Budiansky & Sperling, Note 3). In the sequential search procedure, a sequence of briefly flashed letter arrays is presented on a cathode-ray tube (CRT) display screen, with each new array falling on top of its predecessor. A critical array containing a lone numeral target is embedded somewhere in the middle sequence. The target's spatial location (within the array) and its identity are chosen randomly on each trial. The task of the subject is to detect the location and to identify the target (see Figure 4.2).

In rapid, natural, visual search through simple material, the eyes make about four saccadic eye movements per second, each movement lasting a few tens of milliseconds (depending on the distance transversed), with the eyes relatively motionless between saccades (Woodworth & Schlosberg, 1954). To approximate this natural search mode, the computer-generated arrays are exposed for durations of 200 msec with brief 40-msec blank periods between arrays. Such a stimulus sequence to the stationary eye approximates the stimulus sequence produced by saccadic eye movements. In fact, data obtained with 200-msec exposures followed by 40-msec blank periods are not different from data obtained with 10-msec exposures and 230-msec blank periods (Sperling, 1973; Sperling & Melchner, 1978b, p. 676). The computer-generated sequence has many information processing advantages over the natural sequence. For example, in natural search, when the eyes do not move quite far enough between fixations, some of the same material falls within the eyes' search area in successive fixations and is searched twice, which is wasteful. Even when redundant material on the retina is ignored, the redundant material still usurps space within the search area that could have been occupied by new material. If the eyes move too far between fixations, they leave unsearched lacunae in the stimulus.

In natural search, there are two unknown factors: (1) the eye movement strategy and (2) the attentional field around the eye fixations. Eye movement strategy must be known to determine the attentional factors. In the computer-generated sequence, eye movements are effectively eliminated[4] so that the attentional field around fixation can be determined.

Experimental Investigations of Visual Search in Computer-Driven Visual Displays Visual search was studied with many different presentation rates in addition to those that most closely approximated natural search (Sperling 1970a;

[4] Subjects reported no difficulty in maintaining fixation in the center of the display sequences. In briefly flashed arrays, eye movements during the array are not a problem because subjects would require several tenths of a second to initiate a movement and the array is exposed only for 0.01 second (10 msec) or less. That the stimulus is radially symmetric around the intended fixation in the center of the display negates the utility of altering the fixation point, and the rapid succession of displays encourages oculomotor passivity. When Sperling and Reeves (1980) observed and Murphy, Kowler, & Steinman (1975) and Murphy (1978) carefully measured fixation stability in similar situations, it has been found that subjects can and do maintain stable fixation.

Sperling *et al.*, 1971). The most rapid visual search actually occurred when new arrays were presented every 40 msec—five times faster than the fastest possible saccade rate (Figure 4.2c). At these artifically high presentation rates, search proceeded at a rate of one background character per 10 msec, about twice as fast as Neisser's (1963) maximum rate and twice as fast as in the 240-msec presentation rate that simulated Neisser's conditions. In fact, there was only a small difference in detection accuracy between interarray times of 120 and 240/sec (Figure 4.2b); suggesting that in some natural searches, the motor control of the eye is the limiting factor. In Neisser *et al.*'s (1963) search task, if their subjects' eyes had executed saccades every 120 msec, search rate might have doubled with little loss of accuracy. The second half of many fixation pauses seems to have been wasted waiting for the eyes to move.

In contrast to Neisser *et al.*'s lists, the computer-generated arrays of different sizes are searched at similar rates (characters/sec). Further, there is a considerable trade-off possible between scanning characters in one array or in several; thus, almost as many background characters can be scanned in 1 array presented for 120 msec (12 characters) as in 3 arrays each presented for 40 msec (4 characters/array). This is best seen by looking at the scan times per character (Figure 4.2c), which dip just below 10 msec/character throughout the 40- to 120-msec interval.

The attentional search field around fixation is defined by the proportion of targets detected at various points within it, as shown in Figure 4.2d for search of 7 × 7 letter arrays. The search field is approximately concentric, centered slightly above fixation. However, locations with fewer neighbors or with adjacent blank space are easier to search (Bouma, 1978; Harris, Shaw, & Bates, 1979; Shaw, 1969), so that the measured search field is distorted by the boundaries of the 7 × 7 stimulus. The subject in Figure 4.2d tends to concentrate search more in the left than in the right half of the stimulus.

The search field depends on the stimuli used to measure it; extremely rapid presentations or extremely small-sized characters shrink the search field. However, these parameter variations do not necessarily alter the shape of the search field. That means, except for a task-dependent monotonic transformation, the search field would appear to be an invariant property of the visual system. Attention is distributed gracefully, like Fujiyama, high in the center and tapering gradually towards the periphery. Attractive though this picture of attention may be, in the next section it is shown to be false by evidence that the spatial distribution of attention can be voluntarily altered.

The Role of Attention in the Visual Search for Multiple Targets Among the most interesting questions concerning this aspect of attention (see also Schneider, Dumais & Shiffrin, Chapter 1, this volume) is: Can a subject search as quickly for a known target, for example say a *5*, as for an unknown numeral that is a member of a set of potential targets, for example say *0, 1, . . 9*? Neisser (1963) conjectured that the answer was yes, but he did not test the hypothesis correctly. The correct test requires comparing performance for the *same* target in known and unknown conditions.

In the sequential-search procedure, comparing detection accuracy for known and unknown targets requires comparing accuracy of the location judgments (*Where* in the critical array did the target occur?) in the two conditions: A typical *numeral-known* condition is a block of 100 trials in which only the target 5 occurs. The corresponding *numeral-unknown* condition is a mixed list of 1000 trials in which the numeral targets *0, 1, . . . 9* occur with equal probability. From this mixed list, the subset of 100 trials (on each of which the target *5* occurred) is extracted for comparison with the known condition. Obviously it would have made no sense to compare identification responses between two conditions because the subject knew in advance the identity of the target in the numeral-known condition. Location judgments are used to compare numeral-known and numeral-unknown conditions. Sperling *et al.* (1971) found that accuracies of location judgments for each of the numerals *0, 1, . . . 9* were nearly the same for target-known and target-unknown conditions and were highly correlated (0.97; see Figure 4.2e), providing strong evidence that the same search processes were executed in the two conditions. Any difference in search processes would suggest that a numeral that was relatively easy in one condition (e.g., known) would be relatively more difficult in the other condition (e.g., unknown), but this was not observed.

One can conceive of the search for a particular target numeral, for example, the numeral *1*, as a search task, for example, Task 1. One can conceive of search for the numeral 2 as Search Task 2. It is known that a subject who has to execute both searches simultaneously (i.e., either target may occur), does so with negligible loss in either search and is thus able (in classical terminology) to *attend* to 2 (or even 10) numerals at once. This is a case of apparently lossless division of attention, or better, of multiplication of attention. To describe this situation in contemporary jargon, the subject can execute 10 searches (for the 10 numerals) in parallel.

There is a technical problem in the interpretation of simultaneous search for two possible targets as the carrying out of two simultaneous tasks. Recall the prototypical concurrent tasks mentioned earlier: driving and listening to the radio. The search for Target 1 is analogous to driving the car, and the search for Target 2 is analogous to listening to the radio. Sometimes a driver encounters a road obstacle and a radio news item simultaneously, but Sperling *et al.*'s (1971) subjects never encountered Target 1 and Target 2 simultaneously in the same display. This is symptomatic of an important, real difference. Before resolving this difference, Example 3, in which two search tasks occur simultaneously, is considered.

Concurrent Search for Two Targets

Target Size Matched to Information-Processing Capacity The visual search experiments of Sperling *et al.* (1971) just described were directed at the question of finding optimal stimuli for visual search. How should characters in an array be arranged so that search can proceed as efficiently as possible? In the course of subsequent experiments with arrays of characters composed of different sizes, it quickly became apparent that it was inefficient to compose arrays of characters of just that size. For example, although it was efficient to image many small characters in the foveal area where acuity is good, these characters were below the acuity limit of peripheral vision, and thus most of the peripheral visual field was wasted. Conversely, composing an array of large characters that are resolvable in peripheral vision caused central acuity to be squandered because the fovea was fully occupied by a mere fragment of a character. The obvious solution seemed to be to compose an array of characters of different sizes, ranging from small characters in the center to large characters in the periphery, in which each size of character was matched to the information-processing capacity of the retinal area on which it was imaged. Anstis (1974) independently developed remarkably similar displays, which he used for demonstrating letters that are equally above their acuity threshold in different areas of the retina.

The investigation of spatially matched arrays was begun by the author in collaboration with Melvin Melchner. As before, test sequences were constructed in which only one target numeral occurred in a critical array that was otherwise composed entirely of letters (see preceding description). This target might occur at peripheral locations that received large-sized targets or central locations that received smaller targets. Figure 4.3 shows one of several array configurations that were tested.[5] To our astonishment, Melchner was unable to search arrays simultaneously for large and small targets (e.g., a large 9 or a small 9). This was especially astonishing because in earlier experiments, he had been able to search simultaneously for ten numeral targets (0, 1, . . . 9) when they were all the same size. Was a large 9 more different from a small 9 than from a large 3 or 4? We set about to devise an appropriate search task to test this possibility.

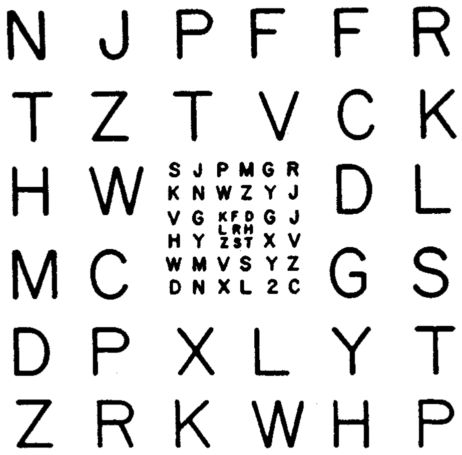


Figure 4.3 Search array matched to the information-processing capacity of the visual system. There is one numeral target.

Concurrent Search for Large and Small Targets To investigate how well subjects can search for targets of two different sizes concurrently, arrays were constructed as follows: an outer frame contained 16 letters of the same large size as the earlier Sperling *et al.* (1971) experiments. A central interior contained 4 small letters. A long sequence of briefly flashed arrays was presented at a rate of 4/sec. A critical array embedded in the middle of the sequence contained a randomly chosen numeral target at one of the 16 outside locations and another randomly chosen numeral at one of the 4 inside locations.

In the main experimental conditions, the subjects' task was to detect both targets: the subjects had to state the identity, location, and their confidence level for each of the two targets. In some blocks of trials, they were told to give 90% of their attention to the inside target and 10% to the outside target; in other blocks, the instructions were reversed; and in still other blocks, subjects were instructed to give equal attention to both classes of targets.

Some useful methodological innovations were incorporated in the analysis of these data. Responses on which the lowest confidence was used (zero, "guessing") were found by chi-square tests to indeed be statistically independent of the stimuli. This means the subjects really knew when they really didn't know (Sperling and Melchner, 1976b, p. 209). Further, analysis of verifiable location errors showed that more than 95% of the time when a target was mislocated, it was mislocated at an adjacent horizontal or vertical (not diagonal) position. Thus, a more rigorous criterion for true identifications could be used, namely, correct identification response *and* confidence greater than zero *and* mislocation not greater than one adjacent position. The data are displayed in Figure 4.4a.

The abscissa and ordinate represent the percent of correct identifications of outside and inside targets, respectively, and each data point represents the average of 70–150 trials. The data fall along a line of slope approximately −1, indicating that as probability of identifying one class of target increases (according to the instructional demand), it is compensated by an almost exactly equivalent decrease in identification probability for the other class of targets. The locus of all achievable joint performances on the two tasks (approximated by the straight-line segments connecting the data points) is called an *attention operating*

[5] In all these experiments, stimuli are flashed very briefly and are disposed symmetrically around the fixation point to induce subjects to maintain eye fixation and move attention rather than move their eyes. See also footnote 4.

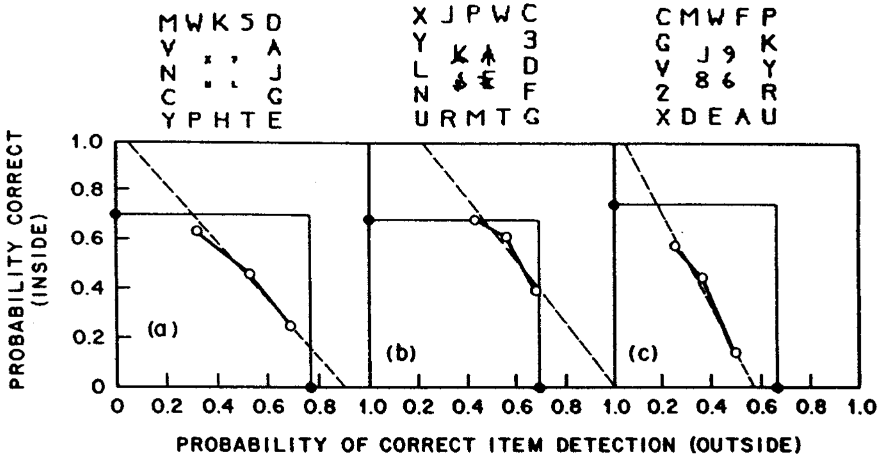


Figure 4.4 Attention operating characteristics for three pairs of concurrent tasks. In (a), concurrent detection of large and small numeral targets is shown. The abscissa indicates the percentage of correct identifications of the outside target; the ordinate indicates the percentage of correct identifications of the inside target. Isolated control conditions are indicated by darkened circles on coordinate axes; the independence point is defined by the meeting of the perpendiculars drawn through these control points. Concurrent performance is indicated by open circles. Data are shown for one subject (MJM) with two (or occasionally three) blocks of trials averaged for each data point. Attention conditions, ordered from upper left to lower right, respectively, are 90% to the inside, equal, and 90% to the outside. The heavy line connecting the data points is the AOC. The broken line represents the best-fitting straight line to the data. In (b), coordinates are the same as in (a), the outside task is the same as in (a). The inside task is detection of a noise-obscured numeral target of the same size as the outside target. In (c), the coordinates and the outside task are the same as in (a). The concurrent inside task is detection of a letter target among three numeral distractors, instead of vice verse as in (a) and (b). (Subject MJM, from Sperling and Melchner, 1978a.)

characteristic (AOC; Sperling & Melchner, 1976, 1978a),[6] following the terminology of signal detection theory. (For the history of receiver operating characteristics [ROCs] see Swets, 1964.) The AOC is a particular instance of the performance operating characteristic (POC) proposed by Norman & Bobrow (1975).

Control Conditions Two control conditions are especially important.

1. Control for memory overload: A series of trials was run in which the letter distractors were replaced by dots. This made target identification trivially easy, and subjects never failed to report both targets correctly. Thus, any errors subjects may make in experimental conditions are due to their inability to detect the targets among distractors, not to their inability to report targets once detected.
2. Isolation condition, report of only one class of targets: In some blocks of trials, subjects were instructed to report only outside targets and to ignore inside targets; and in other blocks, they received the reverse instruction. These control data are graphed directly on the coordinate axes in Figure 4.4a. That the probability of report is nearly equal in the two control tasks (inside, outside) is not a coincidence; the character sizes and array sizes were chosen to match the tasks in difficulty.

Can Subjects Search Simultaneously for Large and Small Targets? If subjects could search for both targets concurrently without loss, their performance in all experimental conditions would fall on the *independence point*, the point at which subjects identify both large and small targets concurrently as accurately as they do in the corresponding control condition. (This is the upper right point of the square in Figure 4.4a). Clearly, this point is not achieved; there always is some loss.

Three Concurrent Pairs of Search Tasks In order to gauge the amount of loss, it is informative to investigate other, related pairs of concurrent tasks. In all, three pairs of tasks were studied. One task in each pair remained precisely the same throughout: detection of a numeral among the outside letters. Three different inside tasks were matched to this task in difficulty: (1) detection–identification of a small inside target, (2) detection–identification of a normally-sized inside numeral (where every inside character was obscured by a randomly chosen "noise squiggle"), and (3) detection–identification of a single target *letter* among inside numerals.

The same control and experimental conditions as before were conducted with these stimuli. Figure 4.4 shows the AOCs generated by these three pairs of tasks. The distance of the AOC from the independence point is a measure of the incompatibility of two tasks. *Perfectly compatible* tasks—performed as well concurrently as in isolation—would fall on the independence point. Perfectly incompatible tasks would fall somewhere on the straight line connecting the isolated control performances.

The percent of isolated performance achieved by concurrent performance provides an index of compatibility between two tasks. Concurrent performance is averaged over the component tasks under conditions of equal attention, that is, at the point where the AOC curve—or surface in higher dimensional space—crosses the line connecting the origin and the independence point. (Area under the AOC is a better, but more complex, measure considered later in this chapter.)

[6] The first to use the term "Attention Operating Characteristic" appears to have been Kinchla in 1969 in the title of an unpublished talk in rural Netherlands and in a privately circulated document (cited in Kinchla, 1980). Curiously, his subsequent publications (Kinchla, 1977; Kinchla and Collyer, 1974) argued against an AOC, proposing instead that attentional manipulations had no effect on the subjects' allocation of their information-processing resources, only on their decision criteria. The attention allocation data of Figure 4.4, the AOC and its relation to the ROC, the attention-switching model, and attendance theory were first reported at the Psychonomic Society, St. Louis, November 1975 (Sperling, 1975) and subsequently more completely at the Seventh Conference on Attention and Performance, Gordes, France, and the International Congress of Physiology, Leningrad, USSR, November 1976. Because of unanticipated publication delays, the data actually were published in the Soviet Union (Sperling & Melchner, 1976a) 2 years earlier than in the United States (Sperling & Melchner, 1978a, 1978b). These are the first published AOCs.

For subject MJM, the most incompatible pair of tasks consists of (1) searching for a numeral among letters concurrently with (2) searching for a letter among numerals. These tasks are almost mutually exclusive, average concurrent performance being about 54% of isolated performance. (By doing only one task or the other, never both—even under concurrent instructions—50% of isolated performance would be achieved, by definition.) The most compatible tasks are searching for a numeral (on the outside) concurrently with searching for a numeral of the same size obscured by noise (on the inside). Concurrent performance is about 82% of isolated performance. The original pair of tasks (searching for a large numeral concurrently with searching for a small numeral) falls between with a concurrent performance of 66% of isolated performance.

We have distinguished implicitly between the process by which performance moves from one AOC to another (changing the component tasks) and the process by which performance moves along a given AOC (increasing the amount of attention allocated to one member of a concurrent task pair at the expense of the other). To compare the compatibility of two pairs of tasks, we generally need two AOCs (not just one point on each AOC). The situation is analogous to signal detection theory in which, to compare the detectability of two signals, in general two ROCs are needed, not just one point on each. The relation of AOC to ROC is developed fully in subsequent sections of this chapter.

Concurrent versus Compound Tasks

Concurrency in Partial-Report Procedures

AOC Graph

Consider the partial report procedure just described in which a subject views a brief flash of a 2 × 4 letter array. The coordinates of an AOC graph are accuracy of report of the top row versus accuracy of report of the bottom row.

Control Conditions

The isolated control condition would consist of a block of trials in which only one particular row of the stimulus is presented. Unfortunately, there is a technical problem with this task. The problem with conducting blocks of trials or cuing the subject too far in advance of the stimulus presentation is that his or her visual fixation will drift involuntarily—if not actually jump—toward the requested row (Kowler & Steinman, 1979, 1981), and such trials are thus not strictly comparable to ones in which the eyes remain fixated between the two rows. An equivalent but better isolated control task is telling the subject just 150 msec in advance of each stimulus presentation which row to report. Either form of the isolated control task leads to nearly perfect performance.

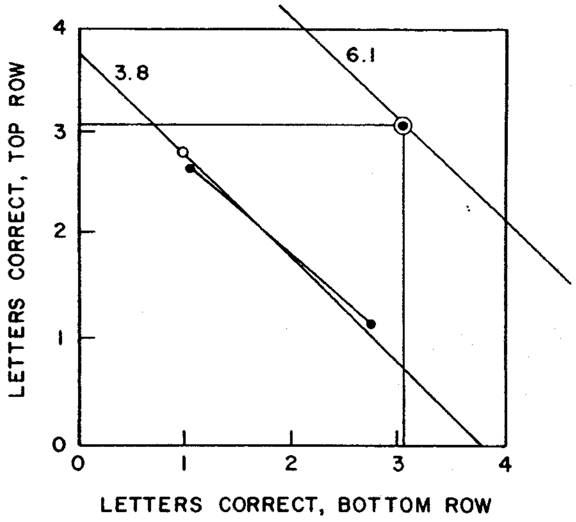


Figure 4.5 Concurrent reports of the top and the bottom row in a whole-report procedure. Filled circles indicate data from one subject for two conditions: (1) give attention to and report the top row first; and (2) give attention to and report the bottom row first. The light diagonal lines indicate equal capacity contours (equal numbers of letters). Open circle indicates data from unconstrained whole reports. Partially filled circle indicates data from partial reports of one row only. The extrapolation of the partial-report data to the coordinate axes represents a lower bound on the performance that might have been observed in an isolated control task. (Subject JC, from Sperling, 1959.)

Concurrent Tasks

The concurrent condition would be a full report of both rows, with the instruction to direct attention primarily to the top row in some conditions and to the bottom row in others. Sperling (1959, 1960) studied a closely related condition that was interpreted by some subjects as the concurrent condition described here. Data are shown in Figure 4.5 for one such subject.[7] This subject has a memory capacity of 3.8 letters in whole reports. In unconstrained whole reports, the subject normally reports the top row first (achieving 2.8 of 4 letters) and then the

[7] The AOC and partial report data of Figure 4.5 are averaged over four cue delays (0, 0.15, 0.30, 0.50 seconds) because performance did not vary between these short delays due to the strategy the subject used. For partial reports (partially filled circle, Figure 4.5), accuracy is assumed to be equally divided between top and bottom rows because the breakdown by row is no longer available. (Data from Sperling, 1959, app. I, Tables 8, 11, 12.)

bottom row (achieving 1.0 of 4). When requested, the subject can, remarkably, allocate most of this capacity either to the top or to the bottom row, with residual capacity falling to the remaining row without any loss. These data define subject's AOC. Data from partial reports (reports of only one row) represent a lower bound on the performance that might have been observed in an isolated control task, and these data are illustrated in Figure 4.5, on the coordinate axes. (Looked at in another way, the independence point, defined by the intersection of the isolated control data on a graph like Figure 4.5, would lie on or beyond the partial report point.) In making whole reports, subjects show enormous loss relative to the independence point because they can remember barely one row of 4 letters and independence would require them to remember 1.5 rows.

The whole report task is a concurrent task requiring concurrent reports of top and bottom rows. In the component (control) task, the subject is prepared for and reports only one row. This control task is equivalent to a partial report task in which the signal indicating which row to report is given long in advance of the stimulus presentation (prestimulus cue.) How are data of partial report tasks with post-stimulus cues interpreted—a task in which the subject must be prepared for both rows?

To represent partial-report data on an AOC graph as in Figure 4.5, the accuracy of partial reports on the top row is graphed as the y coordinate and reports of the bottom row as the x coordinate, even though x and y values represent different trials rather than the same trial as in concurrent tasks. The control tasks (theoretically, blocks of partial reports of only the top row or only the bottom row but, in practice, partial reports with prestimulus cues) are the same as for the concurrent task of reporting both rows.

The partial report data are interpreted as indicating that the subject has a memory capacity of 6.1 letters, almost twice the whole report of 3.8 letters. Unlike the whole report task, in the partial report task there is no trial in which the subject actually reports 6 letters—the inference of a memory capacity of 6 letters is based on trials on which the subject reports only 3. Thus, the interpretation of the partial report task requires a theory about an assumed mental process—visual memory.

Compound Tasks

Partial report is an example of an attention task that is only partially concurrent. Stimuli of *both* component tasks (upper row, lower row) are presented, and only *one* component response (report of only one row) is made. A functionally different type of task is involved in the search for one of two possible targets. The stimulus from just one of the component tasks is presented (one target) and the subject responds with a response chosen from just one of the component tasks (a target name). This is an example of a compound task.

The analysis of compound tasks is fundamentally different from concurrent tasks. To characterize the difference, it is necessary first to define a task. A task is a triple (S, R. U) of two sets (Stimulus, Response) and a Utility function that assigns a real value to every stimulus–response pair (see Sperling, 1983, for details). A concurrent tasks is the *sum* of its component tasks. A stimulus and response are chosen from every component task. In concurrent tasks, such as driving a car and simultaneously listening to a newscast, performance on each concurrent task is compared to performance on the same task in isolation.

A compound task is the *union* of its component tasks. Only a single stimulus selected from any one of all the component tasks is presented on each trial, and the subject responds with any response selected from any of the component tasks. In concurrent tasks, the context that makes a particular task difficult is the other task being performed on the *same* trial. In compound tasks, the context occurs on *other* trials. The detection of Target 1 in the detection task is difficult because, on other trials, Target 2 is presented and the subject must be prepared for both. In the following, it will become clear that, to interpret data from compound tasks, one needs first, a theory of *stimulus uncertainty loss* (loss that a statistically ideal detector with perfect memory and attention would exhibit) and second, a theory of mental processing to deal with any residual, intrinsically human loss.

Concurrency, Compounding, and Processing Stages

Normally, information processing by an organism is regarded as reflecting the operation of a chain of internal processes (a sequence of processing stages.) By definition, concurrent tasks measure the throughput—from stimulus to response—of the whole organism, that is, the whole chain. Compound tasks are usually implicitly interpreted as being concurrent tasks for some subset of the processing chain. Compound tasks are usually interpreted as being concurrent tasks for some subset of the processing chain. For example, compound search for two targets is described (incorrectly) as concurrent within a comparison stage that matches stimuli against remembered targets.

The interpretation of concurrent tasks is easier than that of compound tasks. With concurrent tasks, one observes performance on the task in isolation and then again on precisely the same task in the concurrent context (e.g., driving in isolation and then driving plus listening to the radio). In concurrent tasks, the input and output for all component tasks are observable on every trial.

In compound tasks, only one component task is presented on each trial. Nevertheless, all the compound tasks under consideration here (varieties of signal detection, target localization, stimulus identification, choice reaction times) are interpretable as being concurrent for a subset of the processing chain. This is a favorable circumstance that simplifies interpretation of the performance of the subset, but it is not the whole answer. The input or output of a subset of the processing chain—not both—can be observed because either or both are interior to the organism and therefore unobservable. Ergo, to interpret such a compound task, a theory is needed, no matter how trivial—a theory that (1) asserts which subset is assumed to be concurrently exercised by the compound task and (2) bridges the gap between the exercised subset and the rest of the processing chain. When the compound task is not interpretable as being concurrent within the critical stage, it will be even more cumbersome to relate performance of the compound task to performance of component subtasks; and the theoretical burden will be correspondingly greater.

Concurrency and Compounding in Search Tasks

Simultaneous Search for Different Alternative Targets

Search for Targets 1 and *2 versus Search for 1* or *2* Recall the search experiment in which an observer searched for a numeral target (0, 1, 2, . . . , 9) among letter distractors. This search can be regarded as 10 simultaneous searches. Previously, two of these tasks were considered: (1) search for a *1* and (2) search for a *2*. If performance of Task 1 versus Task 2 is plotted on an AOC graph, the same problem arises as with partial report: searching for *1 or 2* is not the same as searching for *1 and 2*. In searching for *1 or 2*, there are no trials on which both Targets 1 and 2 are presented. Furthermore, the distractor items for Target 1 are also the distractors for Target 2. Contrast this to Sperling and Melchner's (1978a) concurrent search for large and small numerals in which each class of targets had its own distractor—large distractor letters for the large numeral targets and small ones for the small targets.

A Model for Search The purpose of these examples is to illustrate first, that a compound task and a concurrent task require basically different processing mechanisms even when much of the same processing is involved; and second, that a compound task is *necessarily* more difficult than the isolated control, whereas performing two tasks concurrently is not necessarily more difficult than performing an isolated control task.

Consideration is restricted here to the most favorable case for equivalence in the compound–concurrent comparison; namely, to a perception component for which the compound task, search for *1 or 2*, is concurrent (i.e., a component that performs the same operations whether confronted with searching for *1 or 2* or *1 and 2*. Consider, for example, a "perception" component that executes comparisons between a feature list that represents a stimulus item (distractor or target) and the feature lists in memory that represent the eligible target items (0, 1, 2, . . . , 9). It might sensibly program the same sequence of comparisons whether it is confronted with *1 or 2* or *1 and 2* when the task ultimately is to discover the location of *1 and/or 2*.

In order to analyze a subsequent decision component that acts on the output of the perception component, this output must be defined. The required response in both the concurrent and compound tasks is a location judgment. Therefore, it is assumed that the output of the perception component consists of the best estimate of the location of each possible target (*1*, *2*), for example, *1* lower left and *2* upper right. In the concurrent task (search for *1 and 2*), this is exactly what is needed for a response, and the decision component merely needs to transmit the location information for the eligible targets (*1 and 2*) to a response mechanism. In the compound task, this is inadequate information because there are two possible responses (e.g., lower left, upper right). Thus, a sensible decision component would ask the perception component to rate its outputs in terms of confidence. Then, the decision component could choose the one among the alternatives that maximized the expected payoff. Only if the a priori probabilities of target occurrence were equal for all targets, and the payoffs were equal for all correct responses and equal for all errors, would the highest confidence output necessarily be chosen for response. Note that the confidence information—essential to the compound task—is irrelevant to the concurrent task.

Performance in the compound task is inevitably inferior to performance in the isolated control condition—even when there is no attentional loss of information. This occurs because of occasional confusion of a distractor with an unpresented target. For example, in a trial on which Targets 1 and 2 were eligible for presentation, and Target 1 was presented, the subject might think Target 2 occurred somewhere in the search field and respond accordingly. The response, a location judgment, would probably be wrong. This kind of error is avoided in the isolated control condition when the subject knows that only Type 1 targets can appear.

In both examples (partial report and visual search), the concurrent and the compound versions of the tasks were interpreted as making similar demands on an early stage (VIS or sensory memory in one case and perception or recognition in the other) and making enormously different demands on later stages of memory and decision. This is hardly surprising given that these compound tasks were invented in order to study early stages of information processing.

Conclusions The conclusions from the analysis thus far of concurrent and compound tasks are

1. There necessarily is at least some loss in compound search for multiple targets compared to search for a simple target.[8] It requires a detection theory to determine whether or not experimental data exhibit an additional loss due to attention.
2. Concurrent search does not necessarily show a loss compared to an isolated control.
3. In concurrent tasks, a task performed in isolation is the standard against which performance is measured for precisely the same task in different concurrent combinations. Concurrent tasks can be

[8] Duncan (1980) makes similar observations of the performance losses that are inherent in increases in the number of stimulus alternatives, increases that inevitably occur in going from an elementary task to a compound task (of which the elementary task is a component).

analyzed without resort to any information-processing theory. In this respect concurrent tasks are the ideal tool for studying the splitting of attention between competing task demands because concurrent tasks measure the input–output characteristics of the whole organism. There are two cautionary provisions: (1) To avoid degenerate cases, the component tasks of a concurrent task must be completely discriminable—the subject must be able to correctly say which task he or she has performed. (2) Alternatives to attentional loss—such as response incompatibility—must be ruled out in control experiments.

4. Performance in compound tasks cannot be analyzed without a theory about the intermediate stages of information processing. When the assumptions, model, or theory about intermediate stages are not controversial, they may be a small price to pay for learning about intermediate stages. In fact, the focus of compound tasks is inherently on the input–output characteristics of intermediate stages of processing, and cleverly designed compound tasks (together with concurrent tasks) are the ideal tool for unravelling the components of information processing.

Thus, concurrent tasks are appropriate for either macroscopic or microscopic analysis, whereas compound tasks are inherently microscopic in analysis.

Compound and Concurrent Search of Multiple Locations

In this section, these principles are applied to one of the most studied problems of visual attention—the ability (or lack of ability) of observers simultaneously to monitor different locations in the visual field for the occurrence of some event. Over how many locations in the visual field can attention be spread before performance suffers in each local area? This question has already been considered in the numeral detection experiments (Figure 4.2 and 4.4). It is demonstrated here that attempts to answer this question with classical incremental stimuli run afoul of insidious methodological traps.

Mertens One of the first serious experimental attempts to measure the spread of visual attention was by Mertens (1956). He required his observers to maintain fixation faithfully on a central fixation mark. He then presented them with very weak flashes of light to be detected. When subjects detected a flash, they indicated so by pressing a button. In some blocks of trials, the flashes could occur at any of four locations around fixation (northwest, southwest, southeast, or northeast); in others, at only one (e.g., northeast; see Figure 4.6).

Mertens's observers seemed to have slightly lower detection thresholds at an unknown one of four locations than at one predetermined location. He concluded, gamely, that it was more effective for the subject to allow visual attention to spread out over four locations than "to stress himself continually not to look in the direction of attention" (p. 1070). Unfortunately, Mertens was unaware of the rudiments of signal detection theory, and so there were flaws in his procedure. His strange result was replicated once by Schuckman (1963) in an experiment with the same difficulties as Mertens' and by Howarth and Lowe (1966), who found no effect of any kind of uncertainty—not of stimulus location, size, or time of occurrence. Since then, the opposite result has been obtained.

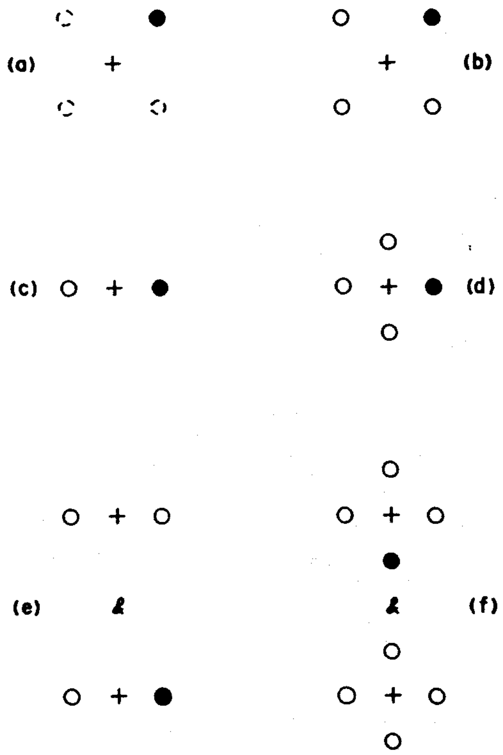


Figure 4.6 Stimulus configurations in spread of attention studies: left column, focused configurations; right column, spread-out configurations. Plus signs indicate fixation marks. Filled circles indicate targets. Open circles indicate other locations at which a target can occur on other trials. (a) Mertens' display for yes–no detection judgments; the dashed circles indicate locations that, in some conditions, contain visible pedestals (baseline illumination) even though no targets will appear there. (b) Mertens' display in which the target may occur in any of four locations; a particular target in the northeast location is indicated. (c and d) Displays for, respectively, two-alternative and four-alternative forced-location judgments. (e and f) Cohn and Lasley's displays for two-interval forced-choice judgments; the ampersand lies between the display that occurs in the first interval (above) and in the second interval (below).

Some contemporary approaches to spatial uncertainty in detection are now considered in order of increasing complexity.

Forced-Location Judgments Consider the following *Gedanken* experiment. A flash can occur in either of two locations[9] (east or west) in some sessions and in any of four locations (Figure 4.6c and d) in other sessions. The accuracy of naming the target location is measured and found to be higher in two-location

[9] See Footnotes 4 and 5.

than in four-location sessions. Unfortunately, it requires a theory of guessing to estimate the attention effect because visual detection accuracy is higher in two-location than four-location stimuli even when the observer's eyes are shut; and for low-intensity stimuli, the guessing effect is predominant. This forced-location judgment, for example, is the one used by Sperling *et al.* (1971) in the search tasks described in a previous section (Figure 4.2).

Two-Interval Forced Choice To obviate guessing analysis, a two-alternative forced-choice paradigm may be used (Cohn & Lasley, 1974). All trials are composed of two intervals, and a target always occurs in one or the other of the intervals. In some sessions, there are two possible locations in which the target may occur; in others, there are four possible locations for the target.[10] Subjects correctly identify the interval containing the target in the two-location trials. Because chance guessing is the same in both kinds of trials, and four locations usually are not monitored as accurately as two, there is an attentional loss in attempting to monitor four locations. Simple? Wrong!

The problem in interpreting the results of the two-versus four-location experiments is that, according to the most widely accepted theory, even a theoretically ideal detector (one without attention or memory loss) would do worse on the four-location trials than on the two-location trials. An attentional deficit is proved only when the human's loss is larger than the ideal detector's loss. Thus, the interpretation of these experiments hinges on the theory of ideal detectors.

An *ideal detector* operates according to a rule that has been proved optimal for some desired outcome; for example, maximizing the percentage of correct responses. An ideal detector does not forget information and is limited only by the quality of data it receives. The ideal detector concept may be extended to the case in which the detector has internal noise, provided the noise does not vary as a function of signals or tasks. How do ideal detectors apply to the present experiment?

Detection Theory Assume that each location (i) being monitored produces an amount of sensory noise n_i on each trial where n_i is a random variable. At the target location (i_T) there is an additional signal (s) producing a net input of $n_i + s$. The ideal detector would choose the location that had the largest net input: If $n_i + s$ were greater than all other n_i ($i \neq T$), a correct detection would occur; if some other n_i ($i \neq T$) happened to be the largest, false detection would occur. A false detection might still, by chance, produce a correct response in a two-interval forced choice if the false detection happened to occur in the same interval as (but at another location than) the target. So stated, the theory is simple. As soon as realistic complications are added—such as, (1) unequal signal probabilities, (2) unequal rewards for the various locations, (3) signal parameters not known exactly, and (4) multiple observations (within an interval) at each location—matters become more complex.[11]

[10] Actually, Cohn and Lasley compared one versus four locations (not two versus four), but this has no significance for the discussion here.

Largest Noise Sample I propose a simple way to grasp the statistical complexity involved in multiple-location paradigms. Consider how an ideal detector comes to make a mistake: the noise at some particular location exceeds the signal plus noise at the target. It is necessary only to consider the location i that produced the largest noise sample (n_i) because this is the one that, if it exceeds $n_T + s$, produces the error. The largest noise sample will be the largest of three (in the four-location monotiring task, Figure 4.6d) or the largest of seven (in the two-interval, four-location monitoring task, Figure 4.6f). This compares to the largest of one (Figure 4.6c) or the largest of three (Figure 4.6e) in the two-location monitoring tasks. The greater the number of equally distributed random variables, the larger the maximum of the sample tends to be, and thus the more likely it is to be the cause of an error. A response based on noise is not necessarily wrong, but it is always less likely to be correct than a response based on signal, which is correct by definition. The more locations there are to be monitored, the more noise samples there will be; the more noise samples there are, the more errors there will be. The particular advantage of looking at the largest noise sample is that, although the distribution of the noise random variable may be unknown, there are only three possible distributions of maxima (Gumbel, 1958; Galambos, 1978). Thus when the number of locations is large, the distribution of the maximum noise sample may be better known than that of any individual sample.

Consequences of Stimulus Uncertainty The point of the preceding discussion is that it is not possible to interpret the loss of accuracy in detection with increases in the number of stimulus locations being monitored unless one has a theory to determine whether the loss is greater than would be exhibited by an ideal detector. This dependence on a theory (ideal detectors) is not surprising; a little reflection shows that the paradigms considered here were exemplars of compound tasks. The complexities of compound tasks can be avoided by concurrent tasks that have different, perhaps more tractable, problems. In the case of N locations being monitored, concurrent means that each location has the same probability of containing a target when it is viewed in the context of the other $N - 1$ locations as it does in isolation. It also means that 0, 1, 2, . . . n targets may occur in a presentation instead of just 0 or 1 as in most compound tasks. Obviously, a large number of targets would pose memory and recognition problems as well as detection problems; fortunately, there are paradigms to provide

[11] Some of these complexities are dealt with, for example, by Swets (this volume), Green and Swets (1966), and Shaw and Shaw (1977).

the data to estimate these sources of interference. Furthermore, the occurrence (or detection) of a target makes the detection of a second target more difficult (Gilliom & Sorkin, 1974; Pohlmann & Sorkin, 1976; Sorkin, Pastore & Pohlmann, 1972; Sorkin, Pohlman & Gilliom, 1973; Sorkin, Pohlmann & Woods, 1976; Schneider & Shiffrin, 1977; Sperling & Melchner, 1978b, p. 681).

Concurrent versus Compound Tasks: Summary and Conclusions

Concurrent and compound are just two categories of tasks. It is certainly easy to create new tasks that are neither concurrent nor compound and that are arbitrarily close to either one or the other category. Clearly the concurrent–compound distinction could be obscured or made unimportant. Nevertheless, the paradigms that are in use today do fall into these categories. What of practical importance can be concluded about the four cases generated by the two paradigms (concurrent and compound) and the two outcomes (loss or no loss relative to the control task)?

No Loss

When there is an insignificant amount of loss in either paradigm, the conclusion is simple: there is no loss. That is, the component tasks of the concurrent combination of tasks can be carried out without loss; and the asserted component processes underlying performance in compound tasks could be carried out without loss.

Loss in Concurrent Tasks

Loss in concurrent tasks means an actual, human performance loss in one or more of the concurrent tasks. We emphasize *human* loss because ideal detectors would not show a loss. In a driving–listening task, the performance loss itself is the datum of interest. In concurrent psychophysical tasks, the fact of a loss often is not in itself of interest—the underlying processes are. Thus we may want to know whether the real performance loss is due to an overburdening of detection, recognition, memory, or response processes. Such questions are not answered in just one experiment. For example, Sperling and Melchner's four variations in targets and distractors greatly helped to eliminate memory or response processes as the determinants of performance loss in their concurrent tasks. Usually, more than just one paradigm is needed; and because compound tasks deal with inferred processes, they can be useful in arriving at answers about component processes.

Loss in Compound Tasks

The null hypothesis for compound tasks is the performance loss shown by an ideal detector. To infer an attention deficit, recognition failure, a memory lapse, response interference, or any intrinsically human loss requires first rejecting the null hypothesis. On the other hand, the null hypothesis for concurrent tasks is that there is no loss; observation of any loss is informative about intrinstic human functions.

To this summary two provisos must be added: (1) Concurrent tasks lose their good properties when the component tasks become indiscriminable from each other (a degenerate case). (2) The parameters of a human sensory system—even when it is behaving like an ideal detector—are intrinsically human and may even interest more people than just the psychophysicists. But the distinction between ideal detectors and theories that postulate additional losses is crucial, as is the relation of these theories to the tasks (concurrent and compound) that give rise to them. The purpose of the metaphors and examples of this section has been to emblazon these distinctions in the mind of the reader so that they may serve as guideposts in his or her encounters with the frequently confusing literature on attention.

Optimization Theory

The Interpretation of Performance Operating Characteristics

Attending Example

Nonoverlapping Classes Consider a student who wishes to attend two classes: a class in Nursing, offered between noon and 2:00 p.m., and a class in Spanish offered between 3:00 and 5:00 p.m. The classes are offered in two different classrooms that are adjacent to each other. The student can run from one class to another in a negligible amount of time, but once the student leaves a classroom, the student may not return. At the end of the semester, the student takes a final examination in each class. Each instructor asks one question about each lecture on the examination. The times within a lecture during which the tested material was discussed are distributed randomly and uniformly over the lecture period. If the student happened to be present at the instant the relevant material was presented, the student will be able to answer the question correctly, otherwise the student will fail that question.

The student's only strategic option is the choice of switching time from one class to the other. Clearly, if the student switches from Class 1 to Class 2 anytime between 2:00 and 3:00 p.m., he or she will score 100% on both examinations.

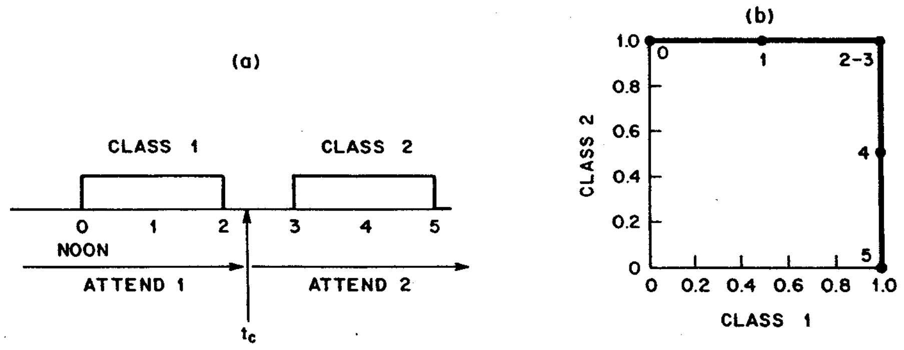


Figure 4.7 Attendance example, nonoverlapping classes. (a) Scheduled times for two classes, (1) and (2). The height of the Class 1 and Class 2 functions represents the instantaneous density of information being offered in each class. The abscissa represents time (in hours), and a particular class-switching time (2:30) is indicated by t_c under the abscissa. (b) Joint performance (performance operating characteristic—POC) for the two classes in (a). The ordinates indicate performance on examinations for each class, 1 and 2, respectively, and the classroom-switching times are indicated near the points on the POC that represent the joint performance on the two examinations.

Figure 4.7a illustrates the two class periods, and Figure 4.7b the attendance operating characteristic, which is another instance of a performance operating characteristic (POC). Joint performance for a switch between 2:00 and 3:00 p.m. is represented at the independence point. The student's expected performance[12] as a function of switch times is illustrated in Figure 4.7b.

The independence point represents the only reasonable strategy for this class schedule. The top and the right hand limbs of the POC represent foolish strategies. The left side and bottom of Figure 4.7b represent perverse strategies. By sitting in Classroom 2 from noon to 2:30 (while nothing is happening there) and then switching over to Classroom 1 and spending the remainder of the time there, the student could achieve a score of exactly zero on both examinations! Having acknowledged that perverse strategies exist, their study is now relegated to another branch of psychology, and attention here is focused on the search for optimal strategies.

Overlapping Classes: Iso-utility Contours Consider now a more perplexing example: Class 1 is scheduled from noon to 3:00 p.m., and Class 2 is scheduled from 2:00 to 4:00 p.m.; the classroom-switching rule and examinations remain as before. Now there is a real scheduling conflict. The attending operating characteristic in Figure 4.8 shows the outcome of the various allowable switching strategies. The student cannot expect to perform perfectly on both examinations. The student can perform perfectly in Class 1 and achieve a score of 50% in Class 2, perform perfectly in Class 2 and achieve a score of 66.7% in Class 1, or achieve something in between. Again, the student can devise strategies, but these are not considered here. How is the student to decide among the reasonable strategies?

To choose a strategy, it is necessary to know the utility of the strategy. For example, suppose that these two classes contribute equally to the student's overall grade point average—the higher the average is, the greater the utility will be. The utility function is

$$u(x_1, x_2) = \frac{100(x_1 + x_2)}{2}, \qquad (4.1)$$

where x_1 is score on Class 1 and $u(x_1, x_2)$ is utility. (Because utility is known only to an arbitrary, strictly increasing monotonic transformation; the concern with scale factors here is only to clarify the example.)

The optimal strategy is to attend all of Class 2 and as much of Class 1 as possible, thereby achieving an average of 83.3%. This and other implications of the particular utility function 4.1 are made intuitively obvious by plotting the utility function together with the operating characteristic[13] as in Figure 4.8. The utility of each strategy can now be computed. That is, utility can now be written as a function of the class-switching time (t_c) by writing the examination scores as a function of t_c. Therefore,

$$u(t_c) = 50\left[\frac{\min(t_c, 3)}{3} + \frac{\min(4 - t_c, 2)}{2}\right], \qquad 0 \le t_c \le 4,$$

where min(a, b) is defined as the smaller (minimum) of a and b, and t_c is measured in hours (with noon taken as zero).

The diagonal lines in Figure 4.8b represent iso-utility contours. Utility (u) can be computed for every point in the joint performance space (x_1, x_2) whether or not that point is achievable. The parameters used to label iso-utility contours

[12] In this example, information is viewed as though it is presented at an instant in time and the stochastic variability in responses is neglected, thereby treating expected outcomes as though they were actual outcomes. These are technical details that would unnecessarily complicate the exposition for those readers who are not fluent in probability theory and that are not essential for those readers who are.

[13] Graphs displaying operating characteristics together with iso-utility contours have long been widely used in economic theory and were introduced to psychology via the study of attention by Navon and Gopher (1979). The concept of utility is central to signal detection theory (e.g., Swets, Tanner, & Birdsall, 1961), and ROCs have been graphed together with various performance criteria (Swets, 1973); but the only prior graph of an ROC together with iso-utility contours is in Metz, Starr, Lusted, and Rossman (1975, Figure 5, p. 420). The attendance example was proposed by Sperling and Melchner (1978b).

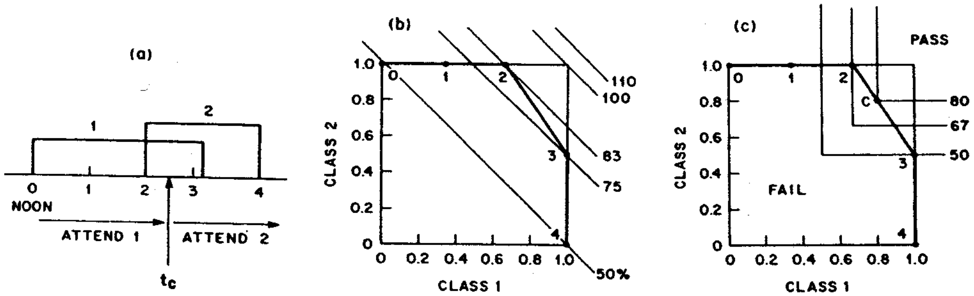


Figure 4.8 Attendance example, overlapping classes: (a) density of information offered in Classes 1 and 2 as a function of time; (b) joint performance (POC) for the classes in (a). The iso-utility contours represent the loci of equal contributions to the cumulative grade point average. (c) Same POC as (b). The contours divide the space into pass–fail regions and the parameter represents the minimum passing grade required in both courses.

indicate their utility. The attendance operating characteristic can be followed as it crosses these contours until it touches the maximum utility contour it can reach. The highest contour reached is 83.3%; this occurs with a class-switching time of 2:00 p.m. and results from a perfect score in Class 2 jointly with 66.7% in Class 1. The reason for the relative neglect of Class 1 is that useful information has higher density per unit time in Class 2 than in Class 1, and therefore the *marginal utility* of attending Class 2 is greater than that of Class 1. The student should exchange time in Class 1 for time in Class 2 whenever possible.

Suppose that what matters is not grade point average but simply passing all the courses. The utility is 1.0 if both courses are passed, and 0 otherwise. Figure 4.8c illustrates utility graphs for three minimum required passing grades (50%, 67%, and 80%). The curves in Figure 4.8c are not iso-utility contours as before, but divisions of the graph into two regions: pass and fail. For convenience, the three boundaries under consideration are represented on one graph. All the reasonable strategies suffice when the minimum passing grade is 50%; two-thirds of the reasonable strategies are adequate with a minimum passing grade of 67%; only one strategy will achieve 80%, which is the highest grade simultaneously achievable in both courses. To achieve a grade of 80%, the student attends 80% of each class;[14] that is, he or she switches from Class 1 to Class 2 after 2.4 hours in Class 1 (at 2:24 p.m.). Ironically, this strategy, which is the only one that will able the student to pass both courses when a passing grade of 80% is required, is the only strategy that would cause the student to *fail* both courses when a passing grade of 80.1% is required.

[14] See Footnote 12.

Signal Detection Theory: Receiver Operating Characteristic

In the classroom example, the assumption that information is transmitted by the instructor uniformly over the whole scheduled class period is unrealistic. More realistic assumptions would be that instructors take a while to "warm up" before they reach their maximum exposition rate; and, having once reached this rate, they begin to tire, at first slowly, and then severely. Figure 4.9a shows estimated instructor information rates for two classes: Nursing from noon to 5:00 p.m. and Spanish from 1:00 p.m. to 4:00 p.m. Figure 4.9b shows the attendance operating characteristic for the student who attempts to take both classes. With these more realistic assumptions, the previously straight-lined POCs now describe smooth curves. The information rate during class period i, $P_i(t)$, is assumed to be zero when class is not in session and to be nonnegative during class. The total amount of information (E_i) presented in a class [$E_i = \int_{\text{start}}^{\text{finish}} p_i(t)\, dt$] is assumed to exist and be bounded. For the attending strategy, in which a student attends Class 1 from its start until time c and then attends Class 2 until its finish, the amount of information E_i accumulated in each class is given by

$$E_1 = \int_{-\infty}^{c} p_1(t)\, dt \quad \text{and} \quad E_2 = \int_{c}^{\infty} p_2(t)\, dt.$$

Information accumulates only from the starting time t_0 of the class; however, because $E_1(t)$ is zero for $t < t_0$, it is convenient to write the integral from $-\infty$ to c rather than from t_0. The same holds true for Class 2. The POC is a graph of E_2 versus E_1 as c varies.

The information rates for Nursing (n) and Spanish (s) shown in Figure 4.9a are analogous to the conditional probability distributions given noise n, $p_n(t)$, and signal plus noise, $s + n$, $p_s(t)$, in signal detection theory (Green & Swets, 1966). Typically, these conditional distributions are assumed to be normal probability density functions, but this is not essential to the theory. On any trial k, the stimulus produces an effect t_k described by a sample from the appropriate distribution (p_s, p_n), and the observer reports "signal" if $t_k > c$ and "noise" otherwise. In signal detection theory, c is called the criterion. The probabilities of hits P_s and of false alarms Q_n are given by

$$P_s = \int_{c}^{\infty} p_s(x) dx, \quad \text{and} \quad Q_n = \int_{c}^{\infty} p_n(x) dx.$$

A graph of P_s versus Q_n (Figure 4.9c) is called a receiver operating characteristic (ROC).

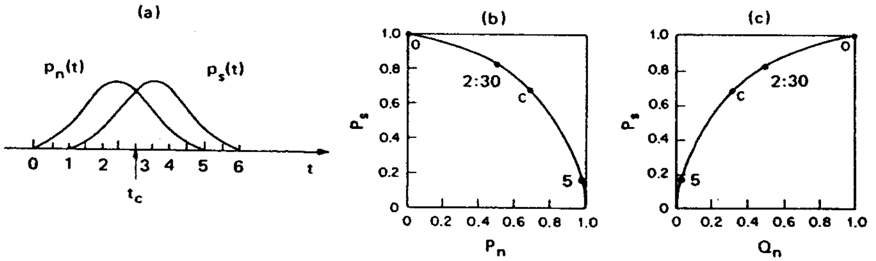


Figure 4.9 (a) Normal assumptions for information density in two classes, Nursing (n) and Spanish (s) as a function of time (t) in hours after noon. (b) Attendance Operating Characteristic for a student who switches from Class *n* to Class *s* at various times. Representative values of switching times are indicated adjacent to the POC, and the abscissa P_n and ordinate P_s indicate performance (examination scores) in the two classes, Nursing (n) and Spanish (s), respectively. (c) A Receiver Operating Characteristic (ROC) for the distributions illustrated in (a). The density functions $p_n(t)$ and $p_s(t)$ are interpreted as the conditional distributions of noise alone and of signal plus noise on the sensory continuum *t*, and the abscissa Q_n and ordinate P_s represent false alarms and hits, respectively. Panel b, the mirror image of Panel c, is now interpreted as a decision operating characteristic, when it is applied to a discrimination or to a signal detection experiment. See text for details.

Decision Operating Characteristic

The ROC and POC graphic conventions produce mirror images of each other (compare Figures 4.9b and 4.9c). In fact, the ROC uses a counter-intuitive convention: it plots good performance on signal trials (hits) versus bad performance on noise trials (false alarms). The mirror image graph of an ROC (1) is mathematically equivalent; (2) plots good performance on signal trials versus good performance on noise trials; (3) follows the usual convention of graphing good performance up and to the right; and, therefore, (4) is psychologically easier to grasp. Let $P_n = 1 - Q_n$. A graph of P_s versus P_n, correct detections (hits) versus correct rejections, is the graph that illustrates good performance in the conventional way and is mathematically isomorphic to the AOC. Although signal detection theory originally was applied to the discrimination of signals from noise, the formalism of the theory applies equally well to other cases. For example, signal detection theory applies well to discrimination experiments in which an observer's task is to discriminate two stimuli (e.g., tones of 1000 Hz and 1001 Hz) as opposed to discriminating one stimulus from zero. Because the general case is discrimination (of which discrimination from zero—detection—is a subcase) the P_s versus P_n representation is appropriately called a discrimination operating characteristic (cf. Sperling & Melchner, 1978b) or a decision operating characteristic (DOC). The AOC, ROC, and DOC are members of a much more general category, that of performance operating characteristics, POCs (Norman & Bobrow, 1975).

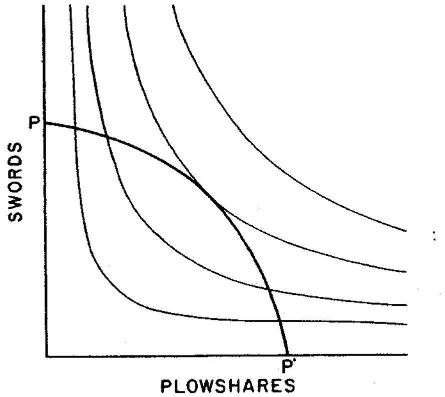


Figure 4.10 Production possibility frontier ($p - p'$) and iso-utility contours for the primitive swords–plowshares economy. The nearly flat utility contours at the extreme right indicate that large changes in (surplus) plowshares can be compensated by small changes in (scarce) swords; similarly, the nearly vertical contours at the upper left indicate that large amounts of (surplus) swords would be traded for just a few (scarce) plowshares.

Economic Analogy

Consider a primitive society in which there are two occupations, agriculture and war. All the farmers could in principle become warriors, and all the warriors could in principle be farmers; although obviously there are some persons who would be better farmers than warriors, and vice versa. The strength of the army is measured in units of "swords" and the productivity of the agricultural sector in units of "plowshares." Consider the joint productivity (swords, plowshares) as the fraction of the labor force that devotes itself to farming varies from 1.0 to 0. The graph of swords versus plowshares has various names; here, it is designated a *production possibility frontier* (Samuelson, 1980). When graphed as in Figure 4.10, the production possibility frontier is concave *toward* the origin. This means, for example, that if the society were to alter the fraction of farmers from 1.0 to 0.95, it would decrease plowshares by less than the increase in swords (see lower right portion of production possibility frontier in Figure 4.10). It is assumed that the first persons to change occupations would be among the worst farmers; that is, persons who were relatively more efficient as warriors than as farmers. Similarly, if the fraction of warriors were to increase from 0.95 to 1.0, it would increase the number of swords only slightly because the very last to join the army would be the least efficient. This tendency to concavity toward the origin is a general property of POCs, which is discussed in more detail later in the chapter.

The production possibility frontier is something that is computed by the systems analysts and economic experts in the society—the technocrats. Insofar as such things can be measured, the production possibility frontier has the status of an objective fact: A particular allocation of resources leads to a corresponding output. On the other hand, the *utility* of any joint combination of swords and plowshares is something that people express through their politicians. In this instance, utility has the status of a preference or an opinion, although it may be formed on the basis of a logical examintion of objective facts. Suppose, for example, the society were devoted entirely to agriculture. This would tempt some of the more opportunistic neighbors to expropriate the agricultural produce by force, and so the industrious citizens might starve in spite of (or perhaps because of) their efficient production. Converting even a few farmers to warriors might create an effective deterrent. The notion of utility applies not only to production that is possible but also to any production. Suppose a society had a thousand times more plowshares than could be consumed and not a single warrior. It would gladly trade many surplus plowshares for just a few swords. Similarly, a society that had an enormous surplus of swords but no plowshares—to avert possible starvation in case the neighbors had the same paucity of plowshares—would be willing to trade many swords for a few plowshares. From these considerations, it follows that equal utility contours are concave *away* from the origin.

The solution for the society—once it has decided on its utility function—is to maximize utility, which it can do by moving along the production possibility frontier until it touches the highest iso-utility contour. At this optimal point, the utility contour and the production contour will be tangent to each other and have the same slope. This is classical economic theory. Modern macroeconomic theory deals with many complex additions to this model that may prevent an optimum from being achieved, but these are beyond the scope of the present discussion. Note, however, that the guns versus butter trade-off just described has numerous other economic analogies. For example, should an automobile company manufacture large or small cars? Should a scarce resource be consumed now or saved for later? Finding the optimum point along various trade-offs is at the heart of economic theory, and specialized branches of mathematics (such as linear programming) have been developed to deal with the problems of optimization. In the next section, the equivalences between trade-offs and optimization in signal detection, in concurrent tasks, and in economic theory are explored.

Optimization in Signal Detection, Attending, Attention, Economics, and Motivation

Signal Detection Theory

Consider the conditional distributions of noise and of signal plus noise in classical signal detection theory. Let the likelihood ratio (*lr*) for an observation

(x; the internal sensory representation of the stimulus on a particular trial) be the conditional probability of signal given x divided by the conditional probability of noise given x; that is,[15]

$$lr(x) = \frac{p_{s+n}(x)}{p_n(x)} .$$

In the examples considered here, the conditional density functions p_{s+n} and p_n are arranged in order of increasing likelihood ratio on the sensory continuum (x). Figure 4.11 shows p_{s+n} and p_n as normal density functions—the usual assumptions of classical signal detection theory. The criterion c represents the value of x corresponding to the *lr* below which the observer responds "noise," and above which the observer responds "signal." The DOC (mirror image ROC) at the upper right of Figure 4.11 represents the joint performance on n trials and $s + n$ trials as c is varied from $-\infty$ to ∞.

The logorithm of the likelihood ratio is illustrated in Figure 4.11, column 3. The reason for illustrating the log *lr* rather than *lr* itself is that log *lr* is symmetric around lr = 0, which reflects the actual symmetry of treatment of *lr* and lr^{-1}, and that log *lr* occurs in many statistical treatments.

The normalized distance between the mean of n and of $s + n$ density functions is the d' statistic of signal detection theory. The area under the DOC, Area(DOC), is a more general statistic, which is simply related to d' for those special cases where d' makes sense, but is itself useful in more cases. The interpretation of Area(DOC) is

$$Area(DOC) = P\ (x_{s+n} > x_n).$$

That is, Area(DOC) is the probability that a random sample drawn from the distribution p_{s+n} exceeds a sample drawn from p_n. According to a simple signal detection theory model, Area(DOC) is the probability of a correct choice in a two-alternative, forced-choice task (Green & Swets, 1966). More generally, Area(DOC) is a nonparametric measure of the amount by which the $s + n$ distribution dominates (is to the right of) the n distribution. Statistics for Area(DOC) are given in Bamber (1975).

Signal detection theory and decision theory differ only in nomenclature—not in any critical conceptual or substantive way. Both are special cases of theories

[15] The statement defining *lr* is correct for discrete probabilities but needs to be technically elaborated for the continuous probability density functions. The equation is correct for either the discrete or continuous case with appropriate interpretation of p_{s+n}, p_n. In the formal treatment of signal detection theory, the sensory variable x usually is discarded as quickly as possible and replaced by *lr* (because all decisions are based on the value of *lr*). However, when the emphasis is on the sensory continuum under investigation, it is more useful to formulate the theory directly in terms of this continuum—the approach taken here.

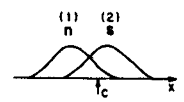
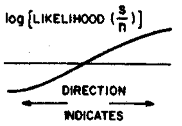
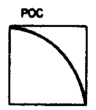
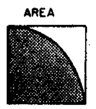
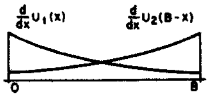
THEORY	(1) n (2) s c x 	log [LIKELIHOOD (s/n)] DIRECTION INDICATES 		c (x)	POC 	AREA 
SIGNAL DETECTION	DENSITY FUNCTION FOR CONDITIONAL PROBABILITY THAT SENSORY OBSERVATION = X GIVEN N, S	NOISE	SIGNAL	CRITERION (SENSORY CONTINUUM)	ROC RECEIVER OPERATING CHARACTERISTIC	P [$X_s > X_n$]
DECISION	SAME, GIVEN STIMULUS, 1, 2	STIMULUS 1	STIMULUS 2	CRITERION (DECISION CONTINUUM)	DOC DECISION OPERATING CHARACTERISTIC	P [$X_2 > X_1$]
ATTENDANCE	DENSITY OF CLASSROOM INFORMATION	USEFULNESS FOR CLASS 1	USEFULNESS FOR CLASS 2	CLASSROOM SWITCHING (TIME)	AOC ATTENDANCE OPERATING CHARACTERISTIC	P [INFO IN CLASS 2 IS OFFERED AFTER INFO IN CLASS 1]
ATTENTION	USEFULNESS OF A RESOURCE FOR TASKS 1, 2	USEFULNESS FOR TASK 1	USEFULNESS FOR TASK 2	ALLOCATION DECISION (RESOURCES)	AOC ATTENTION OPERATING CHARACTERISTIC	P [A TASK 2 RESOURCE IS MORE USEFUL FOR TASK 2 THAN FOR TASK 1]
ECONOMIC PRODUCTION	USEFULNESS OF A RESOURCE FOR PRODUCING PRODUCTS 1, 2	USEFULNESS FOR PRODUCING PRODUCT 1 (OBJECTIVE SUBSTITUTION RATE)	USEFULNESS FOR PRODUCING PRODUCT 2	ALLOCATION DECISION (RESOURCES)	PF PRODUCTION POSSIBILITY FRONTIER	P [A RESOURCE USED TO PRODUCE 2 IS MORE EFFICIENT FOR PRODUCING 2 THAN 1]
ECONOMIC CONSUMPTION (MOTIVATION)	$\frac{d}{dx}$[RESOURCE-UTILITY $u_i(x)$] MARGINAL GRATIFICATION OF DESIRE (APPETITE) 1 BY EXPENDITURES CUMULATED FROM LEFT; 2 FROM RIGHT $\frac{d}{dx}U_1(x)$ $\frac{d}{dx}U_2(B-x)$ 0 B 	EFFECTIVENESS FOR GRATIFYING DESIRE 1 (LEISURE; HUNGER)	EFFECTIVENESS FOR GRATIFYING DESIRE 2 (CONSUMPTION; THIRST)	MOST PREFERRED CONSUMER BUNDLE (EXPENDITURE)	BL BUDGET LINE AOC APPETITE OPERATING CHARACTERISTIC	P [AN EXPENDITURE MADE TO GRATIFY DESIRE 2 WILL GRATIFY 2 MORE THAN 1] NONSUBSTITUTABILITY

Figure 4.11 Summary of isomorphisms among six theories: signal detection, decision, attendance, attention, economic production, and economic consumption (motivation). Density functions are indicated by $n(x)$, $s(x)$; log likelihood ratio is log $[s(x)/n(x)]$; c indicates the decision criterion; and (x) is the dimensional interpretation of the decision variable, x. The performance operating characteristic (POC) is the curve indicated in the graph; the area under the operating characteristic is the shaded area in the rightmost graph. In economic consumption and in motivation theory, the shape of the n, s density functions is usually assumed to be different than in the other theories, as is illustrated.

for compound tasks. The remaining theories outlined in Figure 4.11 are resource theories for concurrent tasks.

Attendance Theory

Attendance Theory deals with two or more classes offered during overlapping time periods. The conditional distributions represent the usefulness of information offered at a time x for each of the courses, respectively. The decision axis of signal detection theory becomes a time axis in attendance theory. At early times, information is more useful for Course 1; at later times, it becomes more useful for Course 2.

The likelihood ratio (lr) of signal detection theory is now interpreted as a usefulness ratio—the higher lr is, the more useful information for Course 2 will be relative to Course 1—whereas criterion and DOC of signal detection theory become the classroom-switching time and AOC of attendance theory. The area under the DOC has an interesting interpretation in attendance theory. Let t_1 be the time when a bit of information sampled randomly from Class 1 was offered, and t_2 the corresponding time for Class 2. Then, the area under the AOC is given by *Area(AOC)* $= p[t_1 > t_2]$. The Area(AOC) is a measure of how much later Class 2 is than Class 1 in terms of the time at which classroom information actually is offered. In order for Area(AOC) to have this useful interpretation in resource theories, the coordinates of the operating graph must be normalized to maximum possible performance, which is equivalent to normalizing the density functions to unit area.

Economic Production Theory

In economic theory, the signal detection theory decision axis of "observations" (ordered in terms of their likelihood of indicating n or $s + n$) is replaced by an ordering of resources (ordered according to their usefulness for the competing production goals of the economy). For example, in the swords–plowshares example, the decision axis might represent an ordering of all the laborers in the country. Those whose usefulness as farmers (relative to warriors) was greatest would be represented at the left side of the axis; those whose usefulness as warriors was greatest would be represented on the right side.

The economic analog to the likelihood ratio of signal detection theory is sometimes called the *objective substitution rate*; in this example, it represents the rate at which swords (or warriors) can be substituted for or converted into plowshares (or farmers). The distribution $p_n(x)$ represents the net effectiveness as farmers of all the laborers whose objective substitution rate (or usefulness ratio) was x. The function $p_{s+n}(x)$ represents the effectiveness of this same group of laborers as warriors. The decision criterion c corresponds to the decision to assign all laborers with usefulness ratio less than c to farming and the remainder to fighting. The DOC of signal detection theory, generated as c varies from $-\infty$ to $+\infty$, corresponds to the production possibility frontier of economic theory, similarly generated.

The decision axis in economic theory need not represent merely labor. In the present example, it could represent some other resource—such as mining and manufacturing operations. For example, mines and other industries could be ordered according to how effective their production was in the manufacture of swords relative to the manufacture of plowshares. Or a single factory could make such an ordering of its internal production facilities. In all these cases, as the decision criterion c to allocate resources to plowshares or to swords varies over its full range, it will trace out the production possibility frontier.

These economic examples show that the concept of resource can be quite general; it may refer to labor, to production facilities, to available capital, or to some combination of all these and other resources. When many resources are involved, matters can be quite complex; and the discovery of an ordering may be nontrivial. Some of the complexities are considered later in the chapter, but they should not obscure the general principles elaborated here about how resources are allocated.

Finally, the area under the production possibility frontier has a similar interpretation to the areas under the DOC and AOC: it represents the probability that a randomly chosen sword-resource will be more useful for sword production than a randomly chosen plowshare resource would be for sword production. It is a measure of the extent to which skills or facilities (e.g., for farming, for fighting) are segregated into different people or facilities as opposed to coexisting in the same person or facility.

Motivation Theory and Economic Consumption Theory

The terms used in studies of animal motivation and consumer economics are quite parallel. Economic consumption theory is somewhat different from production theory because, in consumption, there is only one resource (money) that can be allocated to satisfy various different desires. Similarly, in the typical animal motivation experiment (see Rachlin & Burkhard's, 1978, review), the single resource is time; a hungry and thirsty rat has to allocate its limited session time to working for food rewards or for water rewards.

In the usual case, the satisfaction value (marginal utility) of each additional reward increment diminishes with the amount x already consumed, at least for large x. This nonlinear resource-utility function $u(x)$ leads to a curved POC. Let the resource-utility function $u_i(x)$ represent the utility for reward system i of an

amount x of the disposable resource. It is assumed that the $u_i(x)$ are monotonic, nondecreasing functions of x and that x is bounded ($0 \leq x \leq B$). The bound B is the budget limit; for example, the rat's total session time or the human's total disposable money. The aim is to find a critical quantity c of the resource such that when c is allocated to Reward 1 and the remainder $B - c$ is allocated to Reward 2, the total utility $u_1(c) + u_2(B - c)$ is maximized. This expenditure yields the most preferred consumer bundle. The density functions, $p_1(x), p_2(x)$, of Figure 4.11 can be interpreted as marginal utilities—the derivatives (d/dx) $u_1(x)$, $(d/dx)\, u_2(B - x)$ of the resource-utility functions. Note that, as drawn in Figure 4.11 the marginal utility function $(d/dx)\, i_1(x)$ is cumulated from left to right and $(d/dx)\, u_2(x)$ is cumulated from right to left. In consumption theory, the likelihood ratio represents the ratio of marginal utilities of an incremental expenditure for Rewards 1 and 2 at the budget allocation x. The performance operating characteristic is called a *budget line* or an *appetite operating characteristic*. It represents the joint utilities $u_1(x)$ and $u_2(B - x)$ of expenditures x, $B - x$, respectively, for the two rewards. The particular formulation given in the preceding—in terms of $(d/dx)\, u_1(x)$, $(d/dx)\, u_2(B - x)$—is only one of several ways of generating the budget line.[16] More natural ways to generate the same budget line using nonlinear resource-utility functions are considered later in this chapter (see Figures 4.21 and 4.22 and the subsequent discussion under the subheading "Single-Resource Pool" on p. 166).

The area under the budget line or under the appetite operating characteristic represents the nonsubstitutability of the competing rewards (e.g., food and water). The minimum area (area under the negative diagonal) indicates completely substitutable rewards (e.g., waters with different but equally acceptable flavors); the maximum area indicates nonsubstitutable rewards (e.g., water and dry food). That is, rewards are *nonsubstitutable* if, in a control experiment in which an animal is offered only reward A or only reward B for ½ hour, it consumes them in the same ratio as in a 1-hour concurrent session in which it is offered access to both (cf. classroom example of Figure 4.7). Rewards are *substitutable* to the extent that, in the concurrent experimental session, the animal can be induced to vary the proportion of alternative rewards consumed from the proportion consumed in the isolated control sessions.

Attention Theory

Concurrent Tasks The allocation of mental resources (attention) determines which of several concurrent cognitive tasks are performed more or less well; just

[16] See Coombs and Avrunin (1977) for various sets of conditions that generate trade offs.

as in economic theory, the allocation of economic resources—labor, capital, raw materials, and the like—determines which manufacturing goals are achieved. This analogy of attention to economic production theory was proposed by Navon and Gopher (1979). The critical aspect of the attention analogy is the interpretation of the decision axis x as an ordering of resources—in the case of attention, mental processing resources. The mental resources for which the usefulness ratio x (usefulness for Task 2 divided by usefulness for Task 1) is lowest are represented at the extreme left of the resource axis (Figure 4.11). Thus, the resource axis is directly analogous to a likelihood decision axis of signal detection theory. The conditional density function $p_1(x)$ represents the usefulness to Task 1 of resources as a function of their usefulness ratio x; $p_2(x)$ represents the usefulness of resources to Task 2. The decision criterion c represents the decision by the subject to allocate mental resources with usefulness ratio less than c to Task 1 and the remainder to Task 2. The attention operating characteristic is traced out as c is varied over its range. The area under the AOC represents the probability that a resource, chosen at random from all those useful for Task 2, really is more useful for Task 2 than a randomly chosen Task 1 resource would have been. It is a nonparametric measure of the extent to which separate—as opposed to interchangeable—resources are involved in performing the two tasks.

Compound Tasks Attentional manipulations (e.g., instructions to attend to Task 1 versus Task 2) can be interpreted as controlling resource allocation only in concurrent tasks. In compound tasks, because of the effects of stimulus uncertainty, the attentional manipulation must first be viewed as a decision manipulation (as in signal detection or decision theory). The starting hypothesis for compound tasks is that precisely the same resources are used under all conditions of attention; the quality of data input to and output by these processes does not vary with selective attention, only the decision made on the basis of the data output varies. If, after stimulus uncertainty has been accounted for, there is a residual effect of attention in a compound task; then, obviously, resource analysis would be appropriate for this residual effect.

What Are Mental Resources? There are two approaches to this question. The first is that it is not necessary to know what mental resources are. They have the status of a random variable much like the decision variable of signal detection theory. All the power and prediction of signal detection theory work whether or not the psychological (mental) dimensions of the decision variable are known precisely. All the power of optimization theory is available to predict and describe performance in concurrent tasks even when it is not known precisely where these tasks conflict. On the other hand, I would not be a cognitive psychologist if I did not have a very special interest in learning precisely what particular mental resources were involved in cognitive functions.

With respect to particular mental resources, the critical resources for which there is competition vary with the task. In the partial–whole report tasks, the critical resource was short-term memory (STM); it had a limited capacity, and that capacity was allocated to items from one stimulus row or the other according to the task demand. This memory resource seems to be quite interchangeable.

In search tasks, the critical resource probably is a processing resource involved in making comparisons. A stimulus item at one location in the visual field *can* be compared to a memory representation of a target at the same time that another item in another part of the field is being compared to a representation of another target (Sperling, *et al.*, 1971). However, the extent to which such comparisons can occur simultaneously and the extent to which they draw from a common pool of resources depends on many factors, among the most important of which is the familiarity of the target and the extent to which special resources have been developed for highly specific targets (see also Schneider *et al.*, this volume). Later in this chapter, a powerful technique to investigate whether resources from a common pool can be evenly shared by two tasks or whether they are switched in all-or-none fashion from one task to the other on different trials is examined. First, however, consequences of the unified decision theory that has been presented here are examined further.

Iso-Utility Contours

Iso-utility contours are a powerful heuristic device for studying optimization; that is, for investigating which of a number of alternative procedures or parameters produces the maximum utility or most preferred outcome. Iso-utility contours have long been commonplace in economic theory.[17] Navon and Gopher (1979) introduced iso-utility contours into the study of attention; in this chapter, they were introduced in the classroom example. Here, their use to signal detection theory experiments (DOC, ROC) and to related situations is introduced.

In a typical signal detection task, a 2 × 2 payoff matrix describes the utility of stimulus–response (S–R) outcomes. Let the payoff values be a and j respectively, for "signal" and "noise" responses on noise trials, and h and m, respectively on signal trials, representing false *a*larms, correct re*j*ections, *h*its, and *m*isses. If the fraction of signals is α (thus the fraction of noise trials is $1 - \alpha$), and probability of an "s" response given s is $p(``s"|s)$, and analogously for the other conditional probabilities; the expected utility Eu is

[17] Iso-utility contours are also referred to as "equal-utility contours" and "indifference curves" (Samuelson, 1980). According to Due (1951, p. 92), the indifference curve approach "was suggested in writings of Pareto (1909) and by the Russian economist Slutzky (1915)," and popularized by Hicks and Allen (1934).

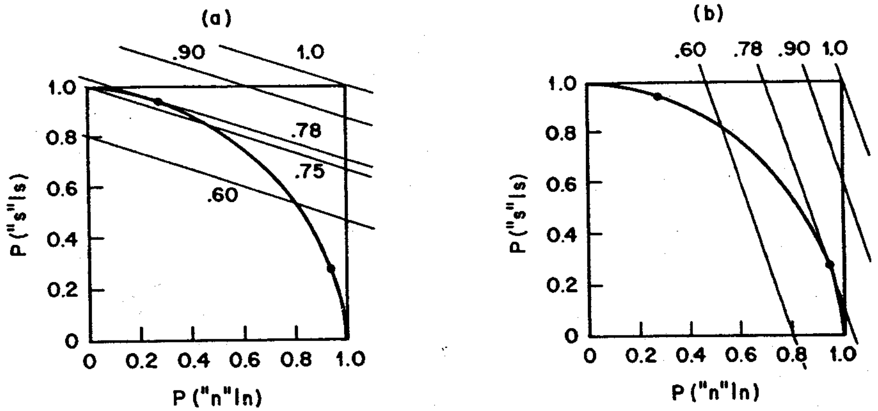


Figure 4.12 (a and b) Iso-utility contours in a signal detection task. In (a), the abscissa indicates the probability of a correct rejection response, given the stimulus was noise, and the ordinate indicates the probability of a correct detection response, given the stimulus was signal. In this example, the a priori probability α of a signal stimulus is 0.75; 1 − α, the a priori probability of a noise stimulus, is 0.25; the utility is 1.0 for correct responses and zero for wrong ones; and the expected (utility) payoff for the joint performances—$p(``S"|S)$ and $p(``N"|N)$—is represented by the labeled iso-utility contours. The DOC illustrated is for equal variance normal distributions with a d' of 1.0 (cf., Figure 4.10). Panel b is the same as Panel a, except α = 0.25.

$$Eu = \alpha\ [hp(``s"|s) + mp(``n"|s) \\ + (1 - \alpha)\ [ap(``s"|n) + jp(``n"|n)]. \qquad (4.2)$$

Given a particular payoff matrix, Equation 4.2 gives the expected utility for every possible performance level—$p(``s"|s)$ and $p(``n"|n)$. (See Equation 4.1 in the attendance example.) In fact, the limits of achievable performance levels are described by the DOC, which is a graph of $p(``s"|s)$, $p(``n"|n)$ pairs obtained as some *nonstimulus* parameter of the experiment that is varied. (The phrase "limit of performance" is used here because the subject can always do worse by making deliberate errors or by following a nonoptimal decision strategy.)

To illustrate the effect of α on performance, a particular payoff matrix is chosen; for example, wrong responses a and m earn zero and correct responses j and h earn V dollars per trial. In brief: $a = m = 0$ and $j = h = V > 0$. Figure 4.12a illustrates iso-utility contours for α = .75 and Figure 4.12b illustrates iso-utility contours for α = .25. The parameter on the contours is the expected utility per trial. Expected utility as defined by the payoff matrix and Equation 4.2 is computable for all values of $p(``s"|s)$ $p(``n"|n)$, not just achievable values. The iso-utility contours are straight lines with slope S easily calculable from Equation 4.2:

$$S = \left[\frac{h - m}{j - a}\right]\left[\frac{1 - \alpha}{\alpha}\right].$$

Suppose the outcomes of a trial are symmetrical with respect to s and n for both errors and correct responses. Then the iso-utility slope is simply the ratio of the two a priori stimulus probabilities, $-(1 - \alpha)/\alpha$. In the two examples in Figure 4.12a and b, the slopes are -3 and $-1/3$.

A typical DOC based on the assumption of equal-variance Normal distributions for n, $s + n$ is also illustrated in Figure 4.12. It is quite obvious graphically that the criterion should be adjusted quite differently to achieve the optimal performance with α = 0.25 and α = 0.75. The expected utility of each strategy (criterion value) can also be estimated from the graph so that the cost of, for example, not changing the criterion (c) when α changes can be quickly assessed. The optimal strategy has an expected utility of 0.78 per trial. This is only marginally better than 0.75, the utility that could be achieved by simply naming the a priori more probable stimulus on each trial without actually observing the stimulus. For Normally distributed signal and noise, it always pays to observe the stimulus because the likelihood ratio varies between 0 and ∞. But there are many distributions—such as the logistic distribution, which is very similar to the Normal—for which the likelihood ratio is bounded. For stimuli characterized by such distributions, when the a priori probabilities are very asymmetrical, it would be better *not* to observe the stimulus but merely to use the a priori information.

The *value* of a priori information is the expected value of a trial with this information minus the value of a trial without it. In the example in Figure 4.12, suppose that an observer has no information about the a priori stimulus probabilities and therefore sets the decision criterion symmetrically (at a likelihood ratio equal to 1.0). The expected probability of a correct response would be 0.69, which (in this example) is also the number of utility units (V) the observer would expect to earn on each trial. Note that 0.69 is the highest achievable expected probability of a correct response with equally probable stimuli or with unequally probable stimuli when the probability is unknown. The a priori information that one stimulus is three times more probable than the other enables the observer to achieve an expected probability of a correct response of 0.75 without even observing the stimulus and 0.78 if he or she chooses to actually observe it. The a priori information alone is thus worth more (0.75) than the opportunity of viewing the stimulus without a priori information (0.69). In the present example, therefore, the a priori information about stimulus probabilities is worth at least 0.25 utility units (0.75 − 0.50) per trial when the observer does not bother to look at the stimulus, and is worth 0.09 units (0.78 − 0.69) if the observer bothers to observe the stimulus. Finally, it is obvious that good detection of signal stimuli $p(``S"|s) = 1$, can be profitably traded off for good detection of noise stimuli, $p(``n"|n) = 1$, when there are more noise than signal stimuli, and vice versa.

All these properties and relations of variables in signal detection are, of course, derivable algebraically; and they are well known. The aim here is to illustrate them in a new way so that previously unobserved similarities between optimization in the various situations (detection, discrimination, attendance, attention, economics, etc.) will be made obvious.

Reaction-Time Trade-Offs

Simple Reaction Time with Alternative Stimuli

Consider the following experiment by Posner, Nissen, and Ogden (1978). A subject views a fixation point between two locations, left and right, where a light flash may appear on a given trial. Whichever flash appears, the subject is to respond as quickly as possible by pressing a key. Occasional blank trials (no flash) are introduced to reduce anticipatory responses (responses before the flash). This is a go/no-go reaction-time experiment in which the subject must respond ("go") when any stimulus is presented and must not respond ("no-go") on catch trials. The experimental manipulation of concern here is the fraction α of stimulus-containing trials on which the left stimulus appears. Posner *et al.* investigated three conditions—trials in which α was, respectively, 0.80, 0.50, and 0.20. Although trials with different α traditionally have been run in separate blocks (Audley, 1973; Falmagne, Cohen & Dwivedi, 1975; Link, 1975; Welford, 1980), Posner *et al.* ran them mixed together, using a pre-cue before each trial to inform the subject of α. The pre-cue procedure (with varying interval of pre-cue to reaction stimulus) is used to determine how quickly the pre-cue can influence the subjects' response. When the pre-cue is given well in advance, as in the present experiment, the mixed procedure is equivalent to the traditional, blocked procedure.

Posner *et al.*'s experiment is analyzed here as a two-task compound experiment in which the two component tasks are (1) press the key when the left flash appears, and (2) press the (same) key when the right flash appears. The outcome of the experiment, the reaction time for each of the two component tasks in the three conditions, is represented in Figure 4.13a. (Except for a slowing of reaction time, there was no important difference between the reaction times in this Donders Type 3 experiment and in a disjunctive reaction-time experiment in which the subject had to press a left key in response to the left flash and a right key in response to the right flash.) So far, the only measure of performance considered here has been accuracy—usually, the fraction of correct detections, correct identifications, or correct answers. In the following, the dependent measure is mean reaction time; errors that occur when the observer responds before the stimulus

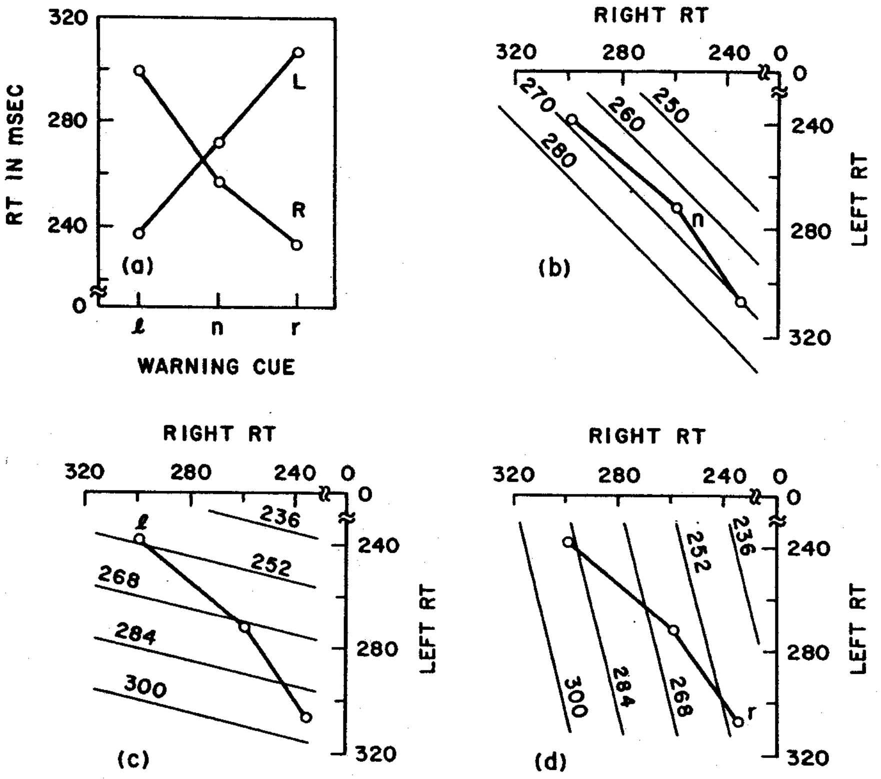


Figure 4.13 Simple reaction times for two alternative stimuli. The same data (from Posner *et al.* 1978, Figure 5) are represented in all panels. (a) Conventional representation of simple reaction times (*RT*) to a flash of a Left (L) or a Right (R) stimulus light with three different warning cues: l and r, indicating 80% probability of Left and Right lights, respectively, and n, indicating a neutral (uninformative) cue. In Panels b, c, and d, the data of (a) are regraphed as a POC. The coordinates represent simple reaction times to the Left and Right stimuli, respectively, and are oriented to show good performance up and to the right. The iso-utility contours in panels b, c, and d each represent equal mean reaction times as indicated; that is, each point along the contour represents a joint performance to Left and Right stimuli for which the overall mean reaction time is the value indicated on the contour, $u = \alpha RT_{\text{Left}} + (1 - \alpha)RT_{\text{Right}}$. In (b), the iso-utility contours represent a weighting of performance appropriate to the stimulus probabilities in effect with cue l: $u = 0.8[RT_{\text{Left}} + 0.2RT_{\text{Right}}]$. (c), the iso-utility contours represent a weighting of performance appropriate to the stimulus probabilities in effect with cue n: $u = 0.5[RT_{\text{Left}} + RT_{\text{Right}}]$. In (d), the iso-utility contours represent a weighting of performance appropriate to the stimulus probabilities in effect with cue r: $u = 0.2RT_{\text{Left}} + 0.8RT_{\text{Right}}$.

occurs, fails to respond within a reasonable time period, or responds on a catch trial are ignored for the moment. Fast reaction times represent good performance; following the convention of this chapter, they are represented up and to the right in Figure 4.13.

Iso-Utility Contours for Reaction Times What are the utilities in Posner *et al.*'s experiment? Unfortunately, the authors did not define these explicitly for the subjects, so it is necessary to guess. Suppose that utility varies in direct inverse proportion to the reaction time: The faster the reaction is, the higher the utility will be; and vice versa. With this assumption, the utility u of any performance—pair of reaction times (RT_{left}, RT_{right})—can be computed as a function of α, the proportion of left stimuli:

$$u = -[\alpha RT_{\text{left}} + (1 - \alpha)RT_{\text{right}}]. \tag{4.3}$$

The utility function (Equation 4.3) is defined in terms of minus one times the reaction times because smaller values of reaction time represent better performance, and utility, by definition, increases as performance improves. Iso-utility functions based on Equation 4.3 are illustrated in Figure 4.13b, c, and d for the three values of α for which data are available. Note that the utility functions (Figure 4.13b, c, and d) are similar to those in typical signal detection tasks, but the data are not, in that the data seem to fall on a straight rather than a curved line. Straight-line data in this experiment, as in Sperling and Melchner's (1978a) attention study, have special significance: They suggest that a mixture of just two states rather than a continuum of states is sufficient to account for the data. This point is taken up in detail later in the chapter.

Posner *et al.*'s observers seem to operate sensibly with respect to the utility function (Equation 4.3) optimizing their performance in each case (Figure 4.13b, c, and d). Note that informative precues enable the observers to shorten their mean reaction times substantially over the mean reaction time to uninformative cues. A valid cue (e.g., left warning followed by left light) "benefits" reaction time (by speeding it up) by about the same amount as in invalid cue "costs" reaction time (by slowing it down). The important point—overlooked by Posner *et al.*—is not that the costs and benefits of knowing α when α equals 0.2 or 0.8 happen to be approximately symmetrical, but that the benefits are available on 80% of the trials, whereas the costs are paid only on 20%. Thus, the mean reaction time improves with unsymmetrical stimulus probabilities in a way that is completely analogous to the improvement of *s*/*n* detection accuracy with asymmetric stimulus probabilities, as considered in the preceding section.[18]

[18] The cost and benefits in choice (as opposed to simple) reaction-time tasks have been extensively analyzed by numerous investigators. For a review of experiments, see Audley (1973); for a review of the random walk model's (RWM) predictions, see Link (1975); for examples of other models, see Green and Luce (1973).

Simple Reaction Time with Alternative Stimuli: A Compound Task Finally, the careful reader will have observed that Posner *et al.*'s task, like the signal detection tasks, is a compound task. On one trial, the observer never receives both a left and right stimulus to which independent responses must be made. A closely related task, which would be concurrent, is responding with the left hand to left stimuli and with the right hand to right stimuli when both stimuli could occur on the same trial. The concurrent tasks would make it possible to determine whether an observer can simultaneously perform two tasks concurrently as readily as one by providing the opportunity to compare concurrent performance to performance in single-task control experiments. This is the ideal task for studying attention, although it may present problems when conflicts arise in the motor system (Kantowitz, 1974). Posner *et al.*'s compound task does not enable us to come to any such conclusion about the ability to perform two tasks simultaneously. The compound reaction-time situation in which either a left or right signal occurs on each trial is analogous to signal detection in which either noise or signal plus noise occurs on each trial. Recall that signal detection theory assumes there is no loss of information by the observer; an ideal detector (matched to the observer's performance on the simple task) would show a loss similar to the observer's when presented with the compound task. In the ideal detector, only the decision criterion changes as the payoff matrix and the a priori signal probability are varied rather than the quality of the information. Is it possible, in Posner *et al.*'s task, that performance varies with instructions and payoffs and yet the quality of perceived information remains invariant? The following question arises: Is a subject slower to react when there are two locations to monitor because the subject cannot process information as efficiently from two as from one location, or does the subject's slower reaction merely reflect the same loss that an ideal detector with no information loss would show in the same situation? As with all compound tasks, a theory is necessary to decide these questions.

Before going on to a theory, it is worth noting that if any performance deficit is observed in a compound task (relative to any of the component tasks) then a full POC could be generated (i.e., a range of costs and benefits observed). To illustrate this point, consider the two-choice reaction-time experiment. This is a compound task with two component tasks, each a simple reaction-time task. (Each simple reaction-time Task *i* requires Response *i* for Stimulus i). Suppose it is known that simple reaction times are faster than choice reaction times for at least one set of a priori probabilities of the occurrence of the component tasks in the compound. That is, suppose Task 1 is slowed in the choice reaction time relative to simple reaction time. Let the a priori probability of Task 1 in the compound be increased to, for example, .9999. A Gedanken experiment is performed using this new compound task. A session consist of 1000 trials. Only once in any 10 sessions will the subject experience a component task other than Task 1. This is a very slow way to gather data about Task 2, but it is as efficient for Task 1 as the simple reaction-time Task 1. Moreover, the data from this compound Task 1 will not differ significantly from the simple Task 1 data because in more than 9 of 10 sessions the simple and compound task have exactly the same stimuli and required responses. Insofar as simple reaction times and choice reaction times are ever different, manipulating a priori probability allows a smooth transition over this range of differences. Thus once it has been observed that choice reaction time is slower than simple reaction time, the POC and costs/benefits for choice reaction times are not a discovery, but rather follow immediately from the procedure for measuring them. Probability manipulations in choice reaction time have been extensively analyzed in the literature (Audley, 1973; Link, 1975; Welford, 1980); therefore the discussion here focuses on a theory for the simple reaction time with alternative stimuli, which has not been so extensively treated.

Random Walk Model for Simple Reaction Time with Alternative Stimuli A simple theory for reaction time, closely related to signal detection theory, is the random walk model (RWM) (Link & Heath, 1975; see also Laming, 1968). Signal detection theory is a theory for the perception and decision component in detection tasks; the RWM is a theory for the perception and response-decision component in reaction-time tasks. Without going into full detail, the principle of the RWM can be summarized as follows: An ideal detector accumulates information from the start of a trial. When the information exceeds a threshold, the appropriate response is made. Each new increment of information is assumed to be somewhat unreliable so that the cumulative balance of all the information may waver between the alternatives—that is, execute a random walk. A strategy consists of a choice of response threshold (the distance from the starting point to the absorbing boundary) for each of the alternative responses.

The response threshold is adjusted so that an optimum compromise is made between several incompatible criteria. The response threshold is set high to avoid accidental incorrect responses (due to some randomness in the incoming information), but not so high that the reaction time is too long. (The higher the threshold is, the longer it will take, on the average, to reach it.) These relations are illustrated in Figure 4.14. A priori information that a stimulus is probable will cause the threshold for the corresponding response to be set lower (thereby decreasing reaction time) without a corresponding loss in accuracy. A priori information that a stimulus is unlikely forces the threshold to be raised in order to avoid errors. (The response threshold is changed by changing *A* or *C* or both together.)

In order to apply a RWM to Posner's task, it is necessary to choose one from among the many candidate configurations. For a single location being monitored in a go/no-go reaction time, I propose a RWM with two boundaries: a near one for the "go" responses and another; much further boundary for the occasional

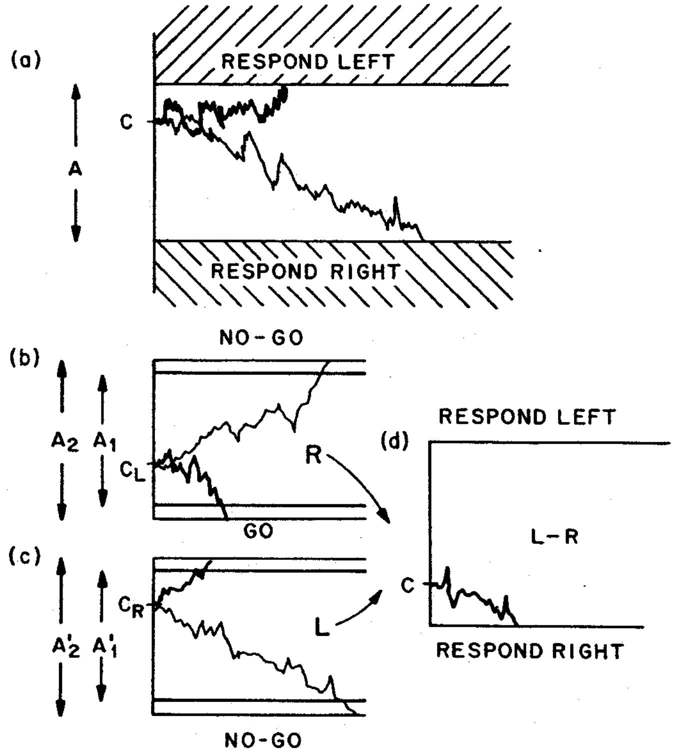


Figure 4.14 Random walk models for reaction-time paradigms. In (a), an illustration of a random walk model (RWM) for a choice reaction time with two alternative stimuli (Left, Right) is shown. A possible random walk that leads to initiation of a Left response is represented by the heavy jagged line; a possible random walk that leads to initiation of a Right response is represented by the light jagged line. When a random walk reaches a boundary, the reaction-time (RT) response is initiated. The sensitivity of the process is controlled by the parameter *A* (which is set in accordance with the penalties for errors and the rewards for speed, etc.). The bias is reflected in the starting point *C* (which is determined primarily by the expectation of Left versus Right stimuli). When *C* is in accordance with the expectation of predominantly Left stimuli (as illustrated), then RTs to Left stimuli will be faster than to Right stimuli (as illustrated). In (b), a random walk for a GO/NO GO RT experiment with a RIGHT stimulus light is shown. The inner, heavy boundaries A_1 are used when this is the only active RW. The outer boundaries A_2 are used when there is one other active random walk (such as [c]); that is, when both LEFT and RIGHT stimulus lights are being monitored. A possible RW that initiates a GO response is shown as a heavy line; and a possible RW that leads to a no response, NO-GO, is shown as a light line. In (c), a random walk for a GO/NO-GO RT experiment with a LEFT stimulus light is shown. Posner, Nissen, and Ogden's (1978) experiment involves both left and right stimulus lights; the first RW to reach the A_2 or A'_2 boundary initiates the response. In (d), a random walk for a Left–Right choice reaction time involving the same stimuli as in (b) and (c). The starting point indicates a strong Right bias; otherwise, the L minus R random walk in (d) is similar to that in (a).

"no-go" response. The no-go boundary has little influence on the simple go/no-go experiment: Catch (no-go) trials are rare, the subject is not rewarded for a quick no-go decision (in fact, this decision speed is not explicitly measured); and if the boundary is very far, only very seldom will it be crossed on go trials. Monitoring two locations is modeled by two simultaneous, independent go/no-go random walks; the response being triggered by the first completion. Because two random walks would produce more false reactions (on catch trials or premature responses) than one walk, the boundaries must be moved away (A_1 to A_2) to maintain the same accuracy in performance. Therefore, monitoring two locations produces slower decisions than monitoring just one. The explanation is exactly analogous to the previous explanation of the difficulty in searching for two targets (*1* or *2*) instead of one target. Recall that the concurrent task of searching for *1* and *2* did not have this problem. Nor would the concurrent task of presenting stimuli independently for responses with the left and right hands. The concurrent reaction-time task is composed of two simple component reaction-time tasks: (1) respond with the left hand to a left stimulus, and (2) respond with the right hand to a right stimulus. The concurrent reaction-time task—in which both left and right stimuli might occur on any given trial—is quite different from the usual choice reaction-time task, which is a compound task: Either the left or the right stimulus occurs, but never both together. The concurrent task is modeled by two independent random walks. The choice (compound) reaction-time task can be modeled in a manner equivalent to Link's (1975) by taking as the RW the difference of the two individual right and left walks considered here (see Figure 4.14).

The point of this discussion is not to provide a definitive model of trade-offs in multistimulus go/no-go reaction times, but to demonstrate a particular class of ideal detectors (RWMs) that exhibit trade-offs similar to those of the subjects in Posner *et al.*'s combined task. In the RWM, a loss of information quality is represented as a slowing of the random walk—it takes longer to accumulate the same amount of information. The attentional question is: When subjects monitor two locations instead of one, do they show more of a performance loss than the simple RWM predicts—a loss that must be described as a slowing of the walk rather than merely an adjustment of the response thresholds? Obviously, this is a very difficult question to answer—so difficult, it suggests that the concurrent task analog to Posner *et al.*'s (1978) experiment should be reconsidered—the concurrent task requires no model for its interpretation.

Conclusion To study attention (the allocation of processing resources) without the burden of a model, use concurrent tasks and abhor compound tasks. To study decision making under uncertainty (the otpimal compromise between incompatible goals when the incoming data are noisy or incomplete), compound tasks are appropriate. Optimization theory (POCs, iso-utility contours, etc.) is applicable in formally similar ways whether the problem is resource allocation (concurrent tasks) or noise (compound tasks).

Speed–Accuracy Trade-Off (SATO)

Consider the following kinds of reaction-time experiments. On each trial, a subject is presented a stimulus that he or she must classify into one of two (or more categories) as quickly as possible. For example, the subject may be shown a letter string and asked to press a reaction key with the left hand if it is a word or another key with the right hand if it is not a word (*lexical decision task;* Rubenstein, Garfield, & Millikan, 1970; Rubenstein, Lewis, & Rubenstein, 1971a, 1971b). Or the subject may be asked to classify as red or green a colored patch that has an irrelevant color name written on it (*Stroop effect;* Stroop, 1935; Kahneman & Treisman, this volume). Or the subject may be asked to classify stimuli by means of a card-sorting task, placing cards as quickly as possible into different piles according to category. Note that all these tasks are compound (not concurrent) tasks; the subject is presented only one of the possible stimuli and makes only one of the alternative responses on any one trial.

In all reaction-time tasks, subjects typically have been asked to respond as quickly as possible while making as few mistakes as possible. These are clearly incompatible goals; subjects could go faster by accepting more mistakes, or they could reduce errors by increasing their reaction times. The ambiguity of the "fast and accurate" instruction is well known: and, in contemporary experiments, the subject is rewarded according to a well-specified payoff matrix for quick correct responses and penalized for errors.

In addition to the straightforward, classical reaction-time procedure two variations should be considered: the *deadline procedure* and the Reed–Wickelgren–Dosher cued-response procedure. In the former the subject is given a time limit (the deadline) within which he or she must respond in order to avoid an explicit penalty (Fitts, 1966). In the *Reed–Wickelgren–Dosher cued-response procedure* (Dosher, 1976; 1981; Reed, 1973; Wickelgren, Corbett, & Dosher, 1980); a "respond now" cue follows the stimulus with a variable delay, with the subject having to respond within a brief interval (deadline) thereafter.

To induce the subject to respond more quickly in the three procedures (classical reaction-time, deadline, and cued-response), the rewards for fast responses and the penalties for slow responses are increased, the deadline is shortened, or the delay of the response cue is decreased. To induce the subject to be more accurate, the penalty for errors is increased, the response deadline is increased, or the delay of the "respond now" cue is increased. Thus, given precisely the same stimuli, subjects can be induced to be either fast and inaccurate or slow and accurate. The range of performance of which a subject is capable defines his or her speed–accuracy trade-off (SATO).

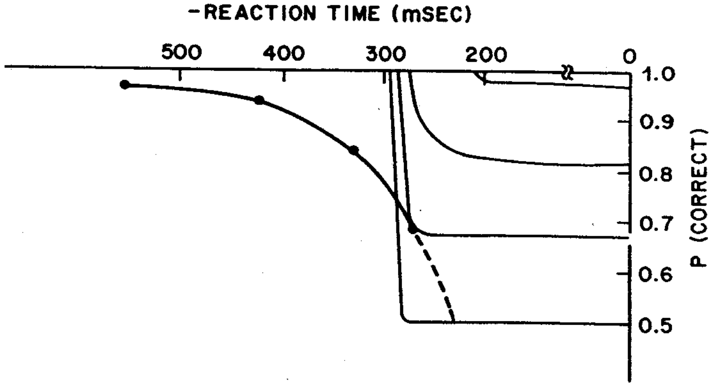


Figure 4.15 A speed-accuracy trade-off (SAT) in a two-alternative reaction-time experiment with various deadlines. The abscissa is the speed (zero minus reaction time), and the ordinate is the probability of a correct response. Iso-utility contours are illustrated for a deadline of 300 msec, corresponding to the rightmost data point. They are equally appropriate for a cued-response procedure. (Data are computed from Pachella and Fisher, 1972.)

Example of a Speed–Accuracy Trade-off and Utility Function A typical SATO is illustrated in Figure 4.15. Data are from a two-choice reaction-time experiment (Pachella & Fisher, 1972) with deadlines of 300, 400, 700 msec, and infinity. Both accuracy and speed are averaged over all responses; accuracy is represented by the proportion of correct responses, and speed is represented by zero minus the mean reaction time. Notice that good performance (fast, accurate) is again represented up and to the right. Whenever payoffs are defined in terms of individual responses rather than in terms of averages over a session, to represent the utility functions along with the SATO on a graph of accuracy versus speed requires additional information about the distribution of individual reaction times and errors. However, for the deadline or cued-response procedures, the form of the utility function is so simple that it will not be much influenced by the reaction-time distribution. A representative utility function[19] is illustrated in Figure 4.15. Utility is proportional to the number of correct responses, with a very high penalty for late responses (i.e., responses that exceed the deadline in the deadline procedure or that fall beyond the "respond now" deadline in the "respond now" procedure). A high utility is achieved by combining high accuracy with a very small fraction of late responses. Responding sooner than required yields no additional award; thus, the iso-utility contour is horizontal for short reaction times. A small increase in late responses must be compensated by a large increase in accuracy; thus, the iso-utility function is almost vertical near deadline time (in the deadline procedure) or the cue plus deadline time in the cued-response procedure.

[19] Pachella and Fisher (1972) used a tone to indicate to subjects that they had responded within the deadline and visual feedback (an arrow) to indicate the correctness of their response, but they did not explicitly assign values to the various outcomes of the trial. At least some of their subjects failed to follow instructions to be as accurate as possible; responding far in advance of the deadline, presumably because of an incorrect (but quite natural) assumption that quicker reaction times were better than slow ones or (equivalently) because of impatience. Therefore, the right halves of the actual utility functions in Figure 4.15 are not quite horizontal. The data in Figure 4.15 are derived from Pachella and Fisher's measure of information transmitted by assuming complete symmetry between alternative responses.

In all nonpathological cases, the POC is concave down, the iso-utility contour is concave up, and the two curves are tangent to each other at the optimum point. In the deadline and cued-response procedures, the corners of the iso-utility contours (where the tangent point will be) tend to be almost vertically above each other; demonstrating the overriding importance of speed (relative to accuracy) in determining the operating point on the SATO.

According to optimization theory—even with the ordinary, ambiguous "speed plus accuracy" instruction—the subject operates at the optimal point on his or her SATO; with ambiguous instructions, the optimum is determined by the subject's *implicit* utility function. Insofar as different points on a SATO can be measured, the reasoning just described can be reversed and the tangent relation between the SATO and the iso-utility contour can be used to discover the shape of the subject's implicit iso-utility contours.

Finally, it should be noted that multiresponse reaction-time experiments cannot be represented completely in a single, one-dimensional SATO. For example, the speed and accuracy of particular response alternatives can be varied inversely even as overall performance remains relatively unaffected—the problem to which RWMs are addressed. However, in symmetrical situations, where the difficulty and payoffs for the various alternative responses in the compound task are approximately equal; the SATO as defined here has interesting and useful properties—some of which are considered in the next section.

Strategy Mixture in Operating Characteristics

On all the operating characteristics considered so far, a strategy is defined as the choice by which a subject or an economy arrives at a particular point on the operating characteristic. In attendance theory, strategy is the choice of classroom-switching time; in signal detection theory, it is the criterion above which a sample will be called a "signal"; in attention theory and economics, it is the choice of how to allocate resources between competing tasks or industries; and in speed accuracy trade-offs, it is the choice of speed and accuracy level (which, according to random walk theory, is mediated by the choice of boundaries—see Figure 4.14). Consider two strategies, S_a and S_b, represented by two distinct points a, b on an operating characteristic (Figure 4.16b). Suppose on some

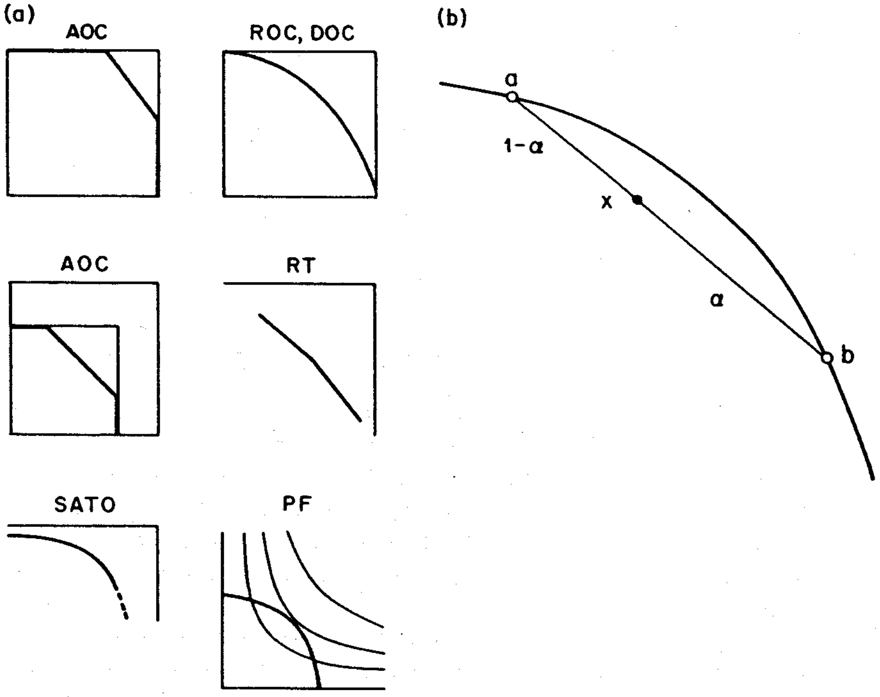


Figure 4.16 (a) Six Performance Operating Characteristics: (AOC) Attendance: (ROC) Receiver, (DOC), Discrimination or Decision; (AOC) Attention; (RT) Reaction Time Trade-off; (SATO) Speed Accuracy Trade-off; and (PF) economic Production Possibility Frontier. (b) Close-up view of a curved segment of an operating characteristic—the points a and b represent performance achieved by Strategies S_a and S_b, and the point x represents performance on a mixture of occasions on which S_a and S_b are used, with S_a being used on a fraction (α) of occasions and S_b on the remaining fraction (1 − α) of occasions.

fraction (α) of the trials, the subject uses Strategy S_a and on the remaining fraction (1 − α) of trials used Strategy S_b. The observed performance **x** would lie along the straight line connecting Performances **a** and **b**. The distance from **b** to **x** is proportioned to α; that is,

$$\frac{|\mathbf{x} - \mathbf{b}|}{|\mathbf{a} - \mathbf{x}|} = \frac{\alpha}{1 - \alpha}$$

and of course $\mathbf{x} = \mathbf{a}$ for $\alpha = 1$.

The property of strategy mixtures, that they lie along the line connecting them in a POC graph, can easily be generalized: The net result of a mixture of strategies is a point that represents the center of gravity of the mixture. That is, let $S_1, S_2, \ldots, S_N$ represent N strategies each of which produces a performance on each of M tasks $(P_{11}, \ldots, P_{ij}, \ldots, P_{MN})$. Let α_j, $\alpha \geq 0$, and $\Sigma_{j=1}^{N} \alpha_j = 1$ represent the proportion of trials on which strategy S_j is engaged. Then the mixture of strategies $S = \Sigma\alpha_j S_j$ produces the performance $P_{1\cdot}, P_{2\cdot}, \ldots, P_{M\cdot}$ where

$$\begin{aligned}(P_{1\cdot}, P_{2\cdot}, \ldots, P_{M\cdot}) &= \sum_{j=1}^{N} \alpha_j (P_{1j}, P_{2j}, \ldots) \\ &= (\Sigma\alpha_j P_{1j}, \Sigma\alpha_j P_{2j}, \ldots)\end{aligned}$$

Primarily, the interest here is in mixtures of just two strategies; so, higher dimensional generalizations can be dispensed with, and the equation can be simply summarized as: The mixture of two strategies lies on the straight line connecting them.

Contingency Analysis: Attendance Example

To explore the properties of strategy mixtures, consider an extreme attendance example. Two courses are offered at precisely overlapping time periods, for example, noon until 1:00 p.m. Suppose a student attends only Course 1 (S_1). The student's performance is perfect on examinations for Class 1, and zero for Class 2. Another student who attends only Class 2 (S_2) has perfect performance for Class 2, and zero for Class 1. To produce an equal mixture of the strategies ($S_{1\wedge 2}$), a third student flips a fair coin each day before class to determine which class to attend. The third student's performance with $S_{1/2}$ is 50% on examinations for each course. On the other hand, a fourth student attends Class 1 from noon to 12:30 and Class 2 from 12:30 to 1:00—a pure strategy. The fourth student also scores 50% on each class's examinations. How can the third student's mixture of strategies be discriminated from the fourth student's pure strategy?

When a POC is strictly concave, then a mixture of strategies lies on a straight line away from the curved POC. Insofar as an intermediate point **y** on a POC lies above the line representing the mixture of its neighbors (**a** and **b**) it cannot represent the mixture of strategies that gave rise to **a** and **b** but represents a new strategy. This procedure can be generalized. Suppose it is established that at least N straight-line segments are required to generate a curved POC. Then there are at least $N + 1$ different strategies. In the limit (for example, in the usual signal detection case with normal distributions assumed for noise and signal plus noise and with a continuously variable criterion), an infinite number of strategies is assumed.

The problem with using the shape of the POC to infer the number or existence of possible intermediate strategies is that it is a statistically weak test when the POC is not very curved, and it is useless when the POC is straight (as it is in the classroom example just described). Nevertheless, a strong differentiation of the mixed and intermediate pure strategy is possible by considering joint performance on the two tasks. Consider the examination questions asked about the material covered on a particular day in each classroom. For simplicity, assume

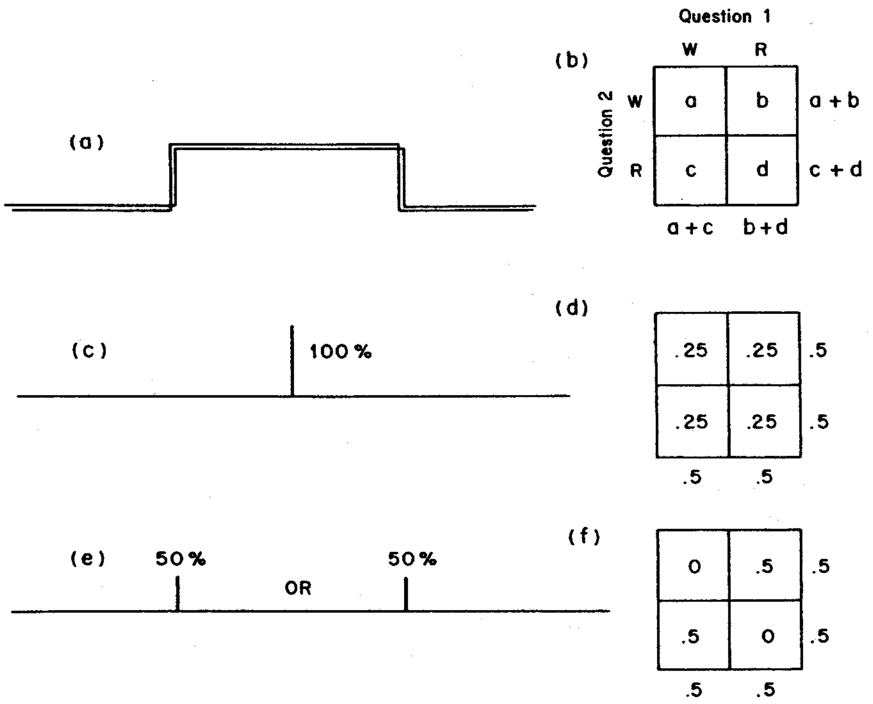


Figure 4.17 Contingency analysis of strategy mixture in a classroom attendance example. (a) Representation of information being offered in two classes scheduled at precisely the same time. (b) Contingency table representing the joint probability of answering two examination questions correctly. Question 1 is drawn at random from material offered in Class 1 and Question 2 from material offered in Class 2 on the same day. W and R represent wrong and right answers, respectively. It is assumed that if the student was in the classroom at the time the material was presented, he or she would answer the examination question correctly. (c) Representation of a (pure) strategy in which, on each day, a student switches from one classroom to the other midway through the class period. (d) Contingency table for the strategy in (c). (e) Representation of a mixture of strategies in which a student attends either one class or the other on each day (i.e., the student switches from Class 1 to Class 2 either in the first or at the last instant of each class). (f) Contingency matrix for the mixed strategy in (e).

that just one question is asked by each instructor. There are four possible outcomes of the joint response to these questions: A student can correctly answer both, neither, or one question from either one of the two classes. These outcomes are represented in the 2 × 2 contingency table shown in Figure 4.17a and b. In the mixed strategy, in which the third student attends all of one class or the other, the student always answers the question from the attended class correctly and fails the other question. Over the whole examination with questions asked about many days, the student's performance will average out to 50% in each class, but the student never answers both or misses both (Figure 4.17 e and f).

On the other hand, if it is assumed that instructors construct their examination questions independently and that they are equally likely to probe information offered in the first half as in the second half of the class period, then the fourth student (who switches classes halfway through the period) is as likely to answer any examination question as any other. This student's pure strategy results in a contingency matrix in which all cells have equal probability. Thus, the pure strategy results in a contingency matrix in which there is statistical independence and zero correlation between the two performaances. The mixed strategy results in a contingency matrix with statistic dependence and maximum negative correlation. In summary, the classroom example shows that a pure and a mixed strategy can be powerfully discriminated even when they fall on precisely the same point of a POC.

Generalized Mixtures of Statistics

The preceding argument can be generalized somewhat: The mixture of two strategies results not only in a mixture of the probabilities of a correct response in each strategy, but also in a linear combination of all the statistics that characterize the two strategies being mixed. In the classroom example, there was a matrix characterizing each component (S_1, S_2) of the mixed strategy; the mixed strategy itself was characterized by the mixture (combination) of the matrices characterizing the components (Figure 4.17). Strategy mixture is mixture indeed.

Strategy Mixture in Attention Operating Characteristics

The AOCs (Figure 4.14) reported by Sperling and Melchner (1978a) are nearly straight lines. The extreme strategies are "give 90% of your attention to the inside" and "give 90% of your attention to the outside," respectively. The equal attention strategy is near the midpoint of these extremes. One may ask, can the contingency matrix tell us whether the "equal attention" strategy is a pure strategy (attention sharing) or whether it is a mixture of switching between the extremes.

The answers differ a little for the different task combinations: In no case are the data powerful enough to reject the switching (mixture) hypothesis; the sharing hypothesis can be rejected for concurrent search for large and for small targets, and for concurrent search for numerals and for letters. For the concurrent search of noise-masked and normal numerals, performance is so close to the independence point that the mixed strategy and pure strategy predictions of the equal-attention matrix do not differ enough to make a discrimination feasible. Although mixture cannot be rejected for any individual subject or condition, all the data deviate somewhat from the pure mixture predictions in the direction of sharing. Thus, the most likely conclusion is that strategies entering into the mixture in the equal-attention conditions are not quite as extreme as the strategies

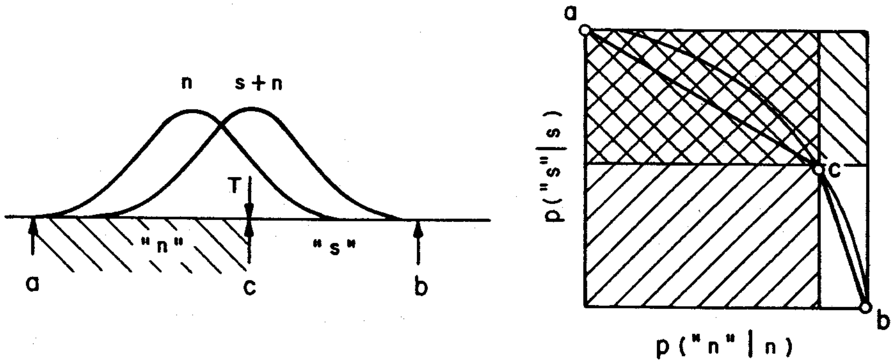


Figure 4.18 (a) Representation of a threshold in signal detection theory. The decision criterion c cannot be moved to the left of the point on the abscissa indicated by T. Shaded area indicates forbidden criteria. Of course the observer can always respond "signal" even without observing the stimulus; this is represented by a criterion at a. A response of "noise" to all stimuli is represented by a criterion at b. In (b) the decision operating characteristic (DOC) representation of the three decision criteria (a, b, and c) in (a) is shown. The shaded area indicates portions of the DOC along which threshold theory asserts no pure strategy can exist; performance within the shaded area is achievable only by strategy mixtures.

employed in the give-90%-of-your-attention conditions. In other words, there are more than two strategies: Equal attention is achieved by switching between strategies that allocate more resources to one or the other class of target, but not such an extreme allocation as is the case with the instruction, "give 90% of your attention" to one class of target.

Strategy Mixture in Signal Detection Theory

The threshold model is the most common source of the (implicit) assertion of strategy mixture in signal detection experiments. Essentially, an observer threshold means that the subject cannot adopt the full range of criteria; small values of n or $s + n$ on the subject's "internal sensory continuum" are all treated alike (Figure 4.18). Whether this is due to a decision criterion that cannot normally be induced below the high threshold or whether it is due to a sensory process that simply transmits the value zero whenever the summed value of input plus process-noise falls below the threshold is immaterial for the present analysis. The net outcome is that the DOC would have an interruption (shaded area, Figure 4.18b) unless the subject employed a mixture of strategies to bridge the gap between the strategy, "always say 'signal,'" and the strategy, "say 'signal' if the decision variable is above threshold."

In its strongest form, threshold theory also asserts that all signals above threshold are treated alike—for the purposes of detection, that is, the unshaded area of Figure 4.18b between b and c also is forbidden. The strong form of high threshold theory is an assertion that the DOC consists of three points derived from pure strategies: two degenerate extremes and one intermediate point. All other points that may be observed on a DOC are based on strategy mixtures that lie on the straight-line segments connecting pure strategies. More complicated forms of threshold models (such as two thresholds: low and high) are equivalent to assertions that the DOC consists of more than two straight-line segments, or of some straight lines plus some curved sections, etc. Still more complicated models (e.g., Krantz, 1969), which include both probabilistic thresholds (a stimulus confusion process) and probabilistic responses (a response confusion process) are not considered here and, in fact, are better considered in a more general scaling context (Shepard, 1958).

Any curve can be approximated by straight-line segments, and smooth curves—such as DOCs typically are—can be well approximated by very few line segments. From a DOC alone, it is impossible in practice to distinguish a threshold theory (N pure strategies plus mixtures) from an infinitely variable criterion theory (only pure strategies). The difficulty arises because the signal detection experiment is a compound (not concurrent) paradigm. There is only one stimulus presented on each trial; therefore, there is no possibility of a contingency analysis such as that which distinguished pure from mixed strategies in the concurrent tasks of attendance and attention theory.

The Objective Study of Strategy Mixture

Phenomenological Approach To resolve the issue of pure versus mixed strategies in signal detection (and similar compound tasks) requires going beyond the data that are usually collected. I propose three methods of resolution: phenomenological, mathematical, and experimental. The phenomenological approach is the simplest: Ask the subjects directly whether they are mixing strategies. This method does not convince skeptics who argue that subjects' answers could be uninformed, misinformed, or (worst of all) deliberately deceitful.

Mathematical Approach The proposed mathematical approach relies on confidence ratings: Instead of a simple yes or no, the subject uses one of J (typically, 5 ± 2) responses ranging from "certain n," to "probably n," to "certain s" (Egan, 1975). Consider a mixed-list detection experiment in which, on each trial, either noise (n) or a signal of intensity i (s_i) is presented, and the subject makes a confidence rating response j. The signal intensity on each trial is chosen randomly from a set of $I - 1$ different, nonzero intensity values so that, with noise, there are I different stimuli. The data consist of the $I \times J$ response matrix, $R_i(j)$ (the proportion of trials on which stimulus i elicits response j). A row of this matrix represents the *profile (distribution) of confidences elicited by stimulus i.*

Let k be the state of the subject produced by stimulus i; for example, k is the value produced by the stimulus on the subject's internal sensory continuum.

Threshold theory asserts that (1) there is a small number K of internal states, and the probability that stimulus i produces state k is $p_i(k)$; (2) each internal state k is characterized by a profile $f_k(j)$ of confidence ratings that the subject produces at random when the subject is in state k; (3) the observed confidence rating profile $R_i(j)$ elicited by a stimulis i will be the mixture of profiles f_k that represents the proportion of times the observer is in state k when i is presented.

$$R_i(j) = \sum_{k=1}^{K} p_i(k)f_k(j)$$

Thus, threshold theory is an assertion about the rank of the matrix R. For example, for $K = 3$, it asserts that even if the experimenter had used 10 different stimuli ($i = 10$), the apparently different observed confidence profiles $R_i(j)$ would be derivable from just 3 fundamental profiles $f_k(j)$ that represent the K internal states.

This formulation of threshold theory makes it equivalent to some of the most studied problems of psychology and mathematics. For example, in test theory, row i of the matrix (which represented a stimulus) becomes a subject, i. And the column j (which represented the proportion of times the confidence rating j was used) becomes a score on Test j. Internal states k become factors $f_k(j)$, and K-state threshold theory becomes the assertion that scores of the subjects on the various tests are explained by a small number K of factors, for which the loadings of tests $f_k(j)$ and the loadings of subjects $p_i(k)$ on the factors are required. From this vantage point, factor analysis of confidence ratings is an appropriate method for analyzing internal states, the number of internal states being at least as large as the number of factors needed to account for the data. Alternative approaches for deriving the minimum number of internal states in discrete-state models have been proposed by Bamber and van Santen (1983) and by van Santen and Bamber (1981), who derive statistical methods (unavailable in factor analysis) for testing the models.

With a small but important modification in the confidence rating procedure, other powerful analytic methods can be brought into play. The confidence rating method uses an arbitrary number of confidence rating levels that are unrelated, and uses no explicit correction method to reinforce the subject for using ratings optimally. Suppose the number of confidence levels is made equal to the number of stimuli. Then the confidence experiment becomes so nearly equivalent to an identification experiment that the confidence scale might as well be replaced with the stimulus names (i.e., their relative intensities). For example, with 9 signals and noise, the confidence ratings 0, 1, . . . 9 are equivalent to noise followed by the stimulus names, in order of increasing intensity. The subject does essentially the same thing whether stating a confidence or stating a belief that some particular stimulus occurred (Sperling, 1965). The resulting S–R confusion matrix can be analyzed, for example, by multidimensional scaling to yield a representation of the threshold stimuli embedded in Euclidian space (see Sperling, Figure 12, p. 160, in Levitt, 1972). The multidimensional representation does not directly answer the theoretical question of how many discrete internal states there are, but it does provide a means of answering many questions relating to discriminability of threshold stimuli.

Experimental Approach: The Contrived Control Finally, one may study the subject's ability to perform in accordance with an N-threshold theory in a case where the N-discriminable thresholds are not in doubt. That is, two experiments that should produce equivalent results if the threshold theory were correct are conducted. The first experiment is the original signal detection experiment, for example, presentation of either a weak stimulus or noise with equal probability on each trial. The subject is required to give confidence ratings on a 7-point scale. Suppose that the analysis by high-threshold theory suggests that the signal exceeds the high threshold 0.7 of the time, and noise exceeds the threshold 0.1 of the time.

A control experiment can now be contrived that is matched in all respects to the original threshold experiment except one: The control has a highly visible stimulus, say 10 *jnds* over threshold. The subject would unfailingly detect this stimulus; so it can safely be assumed that when the stimulus is presented, it exceeds the subject's high threshold. Because in the original experiment, the subject's high threshold was exceeded on 0.7 of the signal trials, in the contrived control experiment, the strong stimulus is presented on 0.7 of the signal trials, and a blank is presented on the remaining signal trials. (A signal trial means that the subject is told (after responding) that a signal was presented.) On 0.9 of the noise trials, the blank is presented; and on the remaining 0.1 of the noise trials, the strong stimulus is presented (to correspond to the subject's internal state following a false detection). For equal probabilities of signal and noise trials, this means that after 0.25 = 0.3/(0.9 + 0.3) of the blanks, the subject is told that the signal had been presented (to correspond to the misses in the original experiment). After 0.125 = 0.1/(0.7 + 0.1) of the strong stimuli, the subject is told that noise had been presented. To make this feedback plausible, the subject is told that the apparatus does not work reliably, and that the task is to base responses—a confidence level—on what the apparatus had been programmed to do, not on what it actually did (because the apparatus failures are manifest).

The contrived control experiment yields a subject whose internal states, given "signal" or "noise," are known. The subject's strategy in the use of confidence ratings, for example, can be studied directly and not merely inferentially. For that matter, a three-state or even an N-state threshold theory can be mimicked by an appropriately contrived control experiment with three or with N highly discriminable stimuli. Insofar as subjects behave similarly in contrived control experiments and in the actual signal detection experiments, it substantiates the N-

state threshold models; insofar as subjects are unable to maintain the complex strategies ascribed to them by the *N*-state theorists, it defeats such *N*-state models. In either case, these procedures provide an objective experimental approach to the study of strategy and strategy mixtures.

Strategy Mixture in Speed–Accuracy Trade-Offs

The assertion of strategy mixture in SATOs comes most commonly in the guise of the fast-guess model. This model applies, for example, to two-choice reaction-time experiments in which the subject is presented on each trial with one of two alternative stimuli and is required to make the corresponding one of two responses as quickly as possible. For example, in Ollman's (1966) and Yellott's (1967, 1971) theory, the subject is asserted to respond to the stimulus with a normal reaction time on some fraction $1 - \alpha$ of the trials, and on the remaining trials α the subject responds as quickly as possible (simple reaction time) according to a predetermined guess at what the stimulus might be. On fast-guess trials, the subject is correct with only chance accuracy ($p = 0.5$) but with very short reaction times; on the remaining trials, the subject has long reaction times and a correspondingly higher percentage of correct responses. When the experimenter demands from the subject an even lower average reaction time, the subject complies by increasing the proportion α of fast guesses.

The fast-guess model is equivalent to the assertion that the SATO is composed of a straight-line segment the end points of which represent the two strategies, the honest strategy and the fast-guess strategy. An alternative hypothesis to fast guess would be that the subject chooses a pure strategy appropriate to each payoff matrix—a process that could be modeled, for example, by boundary changes in a random walk model. This alternative strategy might generate either a curved or a straight-line SATO. As in all the previous cases, it is not efficient to discriminate pure from mixed strategies by close examination of the curvature of the operating characteristic. In the case of the SATO, associated with each point on the SATO are not only the mean reaction-time and mean accuracy (which define the point), but also two reaction-time distributions—one for correct responses and one for errors.[20] The fast-guess model not only requires the SATO to be a straight line, but also requires the reaction-time distributions associated with each point to be a mixture of the reaction-time distributions associated with the extreme points. This is a powerful test to discriminate between strategy mixtures and pure strategies[21] that is formally similar to the test for mixture of strategies in confidence ratings discussed in the preceding. Furthermore, contingency matrices also apply to the SATO. In a pure strategy, reaction speed and accuracy are uncorrelated or weakly related. In a mixed strategy, however, there is a strong negative correlation of accuracy with speed exactly analogous to the negative correlation of performance on Task 1 with Task 2 in the attendance and in the attention contingency matrices.

Mixed Strategies in Production

Consider the primitive plowshares–swords economy. Suppose it is decided to devote half of the economy to each goal. Does it make any difference whether on every odd-numbered day of the year, the whole economy is devoted to agriculture, and on every even-numbered day, the economy is devoted to defense production (mixed strategy) versus the case in which on all days, the resources are divided in half and equally devoted to each goal (pure strategy)? Certainly! The pure strategy is far more efficient in terms of the production facilities needed. But even if production facilities were not at issue and only the availability of labor was, it would still be more efficient to divide labor equally on every day than to alternate days. The reason is that if even one laborer were more efficient at making plowshares than swords, it would be efficient to assign this laborer to the task for which he or she were better suited. By similar reasoning regarding any resource, a pure strategy is preferable whenever resources are not completely equal and interchangeable with respect to the economic goals. This is the line of argument used previously to demonstrate that economic production possibility frontiers are always concave toward the origin. A mixed strategy does not take advantage of the curvature; it always lies closer to the origin and is of lower utility than the corresponding pure strategy (Figure 4.16b). There is inherent superiority in a pure strategy—optimal for the situation—over a mixture of less than optimal strategies.

Given the economic superiority of pure over mixed strategies, it is pertinent to ask, What limitation in allocation of mental processing resources prevents the utilization of pure strategies in human divided attention tasks? One possible answer is that there is a single processor or process involved in concurrent tasks, and that there is a changeover delay incurred in switching this resource from task to task. Switching the resource within a trial produces unacceptable costs. An analogous problem occurs in computer time-sharing systems in switching from one user to another. There is an overhead cost (time and memory) incurred in swapping a second user's program into the central processor unit (CPU), and the

[20] In the case of unsymmetric stimuli or responses, there are even more reaction-time distributions to be considered; but that is beyond the scope of the present treatment (see, for example, Link and Heath, 1975).

[21] Mixtures of two probability density functions $p(x) = p_1(x) + (1 - \alpha)p_2(x)$ have the interesting property (Falmagne, 1968) that there is at least one value—x_0, the fixed point of the random variable—for which the associated probability density does not change as the mixture ratio α varies. Whenever this occurs $p_1(x_0) = p_2(x_0)$. The existence of a fixed point in probability distributions is analogous to a fixed point in the absorption spectra of a putative mixture of two chemicals formed, for example, by decomposition of one into the other.

first user's program out into a buffer until it again gains access to the CPU. Trying to divide time too finely results in too many swaps per second with a corresponding, disproportionately high overhead cost. In the limit, no useful work is accomplished—only swapping is achieved. Changeover costs have some interesting consequences in other economies as well; these are considered in the next section.

Path Dependence in Performance Operating Characteristics

Path Dependence in Classroom Attendance

The simplest situation in which to discover effects of changeover costs is the classroom example. Suppose that when a student was ready to run from Class 1 to Class 2, the second class were located not in an adjacent room but in a different building, and the trip between classes would consume 5 minutes. Clearly, there would be no point in switching from Class 1 to Class 2 unless the information being offered in Class 2 were so much more valuable that it could compensate for the lost time.

The effect of a changeover cost is to maintain the status quo. The student remains in the present classroom, even when another class would be slightly more useful, because the additional utility is insufficient to compensate for the changeover cost. The student's presence in current classroom reflects not only the current utility of the competing classes, it also reflects the past history that brought the student to the class in the first place. A class that was useful in the recent past holds students even after its utility has slipped below that of its competitors. This phenomenon is called *path dependence.* It is ubiquitous in psychology, although only since the 1970s has it begun to be appreciated outside the realm of clinical psychology. It is sometimes referred to as *hysteresis*—a reference to the electromagnetic phenomenon in which a magnetic substance tends to retain its previous magnetic orientation even after an oppositely directed external magnetic field has been applied, a field that—had the previous magnetic orientation been neutral—would have been sufficient to induce a change. Of course, hysteresis can be overcome; it simply requires a stronger external magnetic field. Energy is lost in a hysteresis cycle related to the amount of path dependence, with no energy being lost when there is no hysteresis (see Figure 4.19a). The classroom dilemma is analogous. Students can be induced to switch classes, provided the required differential benefit is sufficient to overcome the cost. The classroom-switching cost—lost information during changeover—is somewhat analogous to lost energy in hysteresis. (A better analogy with magnetic hysteresis is the loss in information due to the student's being in a nonoptimal classroom (see Figure 4.19b). When classes are adjacent and there is no

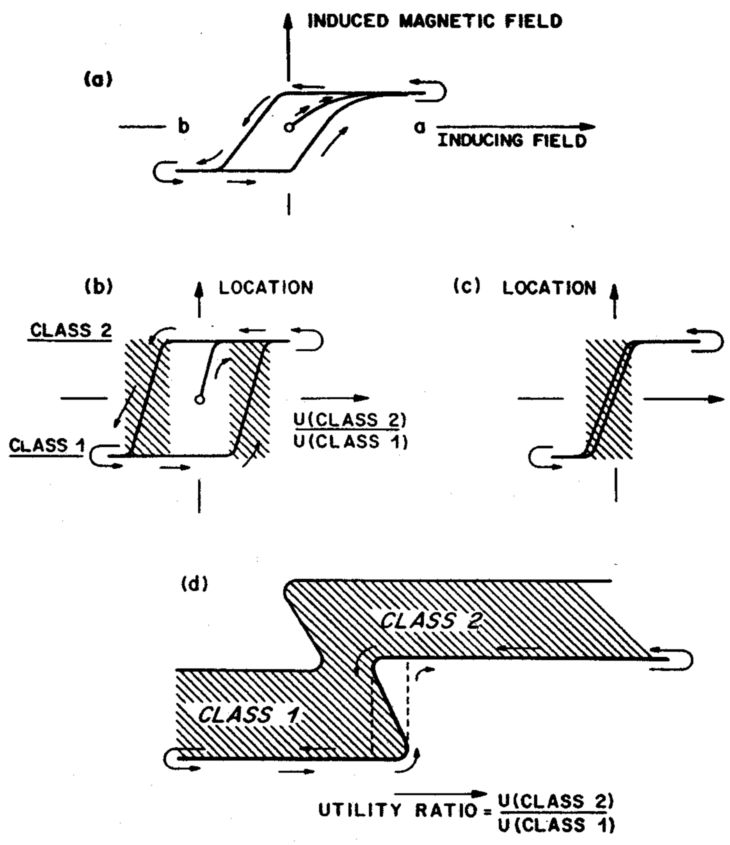


Figure 4.19 Examples of path-dependence. In (a), hysteresis in a piece of magnetized iron is shown. The electric field is initially neutral (open circle) and then varies back and forth between *a* and *b*, indicated on the abscissa. The ordinate indicates the induced magnetic orientation of the microcrystals in the iron. The curved arrows indicate the flow of time. In (b), hysteresis in the classroom is shown. Two courses are offered; the ratio of their utility to the student *u*(Class 2)/*u*(Class 1) is varied periodically during a very long class period. Initially (open circle), the inexperienced student is midway between classes. As the information being offered in Class 2 is becoming more valuable than in Class 1, the student runs to Class 2 and remains there; subsequently, Class 1 becomes more valuable so the student runs to it; when the utility of Class 1 subsides, the student returns to Class 2; etc. Information is completely lost during transit (heavily shaded area) and partially lost during the time the student lingers in the less informative class (clear center rectangle). In (c), the classroom strategy of the upperclassman is shown. Being smarter than an iron crystal or a freshman, this student anticipates the future course of events. When in Class 2, as its utility diminishes, he or she leaves while it is still more valuable than Class 1, knowing that by the time of arrival in Class 2 the relative utility will have reversed. This student loses information only as a result of transit (shaded area), never by being in the wrong classroom. In (d), the catastrophe theory representation of the events in (b) is shown. The upper surface represents Class 2, the lower surface Class 1. The abscissa represents the control parameter, the utility ratio *u*(Class 2)/*u*(Class 1). When *u* is varied and a fold in the surface is reached, the student cannot reverse direction. So, the student jumps to the other surface and continues there. The jump is the "catastrophe" of catastrophe theory.

changeover delay, there also is no path dependence, no hysteresis, and no lost information. The student's strategy at any and every instant of time can be optimal for that instant.

There is an interesting heuristic representation of path-dependent effects in *catastrophe theory* (Thom, 1975a; 1975b; Zeeman, 1976). The student's current classroom may be thought of as the dependent variable, which is under the control of an independent variable—the utility ratio of the material offered in the two competing classes. As the utility ratio changes, the state changes—as described in the preceding and as illustrated in Figure 4.19d. The "catastrophe" occurs when the student switches from one surface (classroom) to the other at a fold in the surface. A useful aspect of the catastrophe theory representation is that all possible equilibrium states and the relations between them are clearly shown. A limitation of the catastrophe theory representation is that neither the dynamic aspects of the situation nor the underlying processes are represented. By itself, a catastrophe theory representation is an insufficient description of a dynamic system (Sussman & Zahler, 1978).[22]

Path Dependence in Economics

I do not know of an analog to changeover costs in signal detection theory; but in economic theory, changeover costs (obviously) are extremely significant. For example, according to the preceding argument, it would be expected that as demand for small cars began to grow in the 1970s, the American automobile manufacturers would have continued to make too many large cars because of the retooling expense and risks involved in changing over from the manufacture of large to small cars (see Figure 4.19b). In fact, contemporary economic decisions seldom can afford to wait until they are caused by events themselves; almost always, it is the anticipation of predicted events and of trends that drives economic decisions (see Figure 4.19c). The incorporation of expectations into economic theory so complicates the theory that it can hardly serve as a simple illustrative example. Whether automobile manufacturers retooled early or late depended on their expectations of future market forces. Because it is not known what these expectations were nor how they were derived, the automobile retooling problem is hardly the simple changeover analogy that it superficially appears to be.

Path Dependence in Attention

The moral from economics for the study of human attention is that sophisticated performance requires correspondingly sophisticated theory. Sperling and Melchner's (1976, 1978b) observation that visual attention tended to be switched rather than shared suggested that there was a single, serial central processor that avoided switching between tasks within a trial in order to avoid the changeover cost; and thus, within a trial, the processor was devoted almost entirely to one task or the other. The single processor hypothesis is the most ubiquitous theory of attention. However, in this instance, it is too simple and leaves too many questions unanswered. How does one account for the imperfect but nonetheless substantial performance on the secondary task? Further, in follow-up experiments in which different sizes of targets occurred in the same or in nearby places, Sperling and Harris (Note 4) failed to find significant effects of attention instructions; performance was at the independence point. A similar result is reported by Hoffman and Nelson (1981). In case it is not already obvious, the reader is reminded that not all questions are answered in this chapter.

[22] See Sperling (1970b) for illustrative examples of the relations between path dependence and multiple stable states. See Sperling (1981) for the relation of catastrophe theory to the just-discussed phenomena and for references.

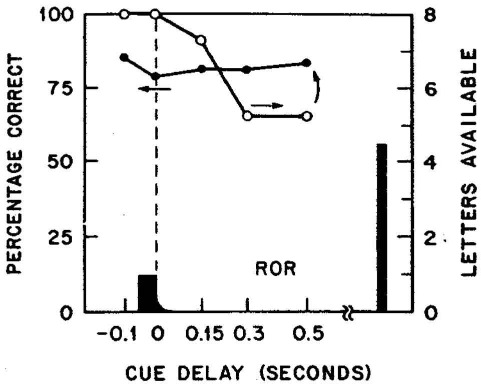


Figure 4.20 Path-dependence in a partial report experiment. Accuracy of partial reports for 10 blocks of trials conducted in one session. Arrows indicate the sequence of blocks. Bar at right indicates accuracy of whole reports. Stimulus exposure is indicated at lower left. Descending cue-delay series (open circles) corresponds to the strategy, "attend equally to both rows", and ascending series (closed circles) corresponds to the strategy, "attend primarily to the top row." The loop is not closed at the left in these data because the session ended; but from other data, it is clear that with continued exposure to prior cues (−0.10 seconds) the subject will switch to "equal attention," that is, jump to top curve. (Subject ROR from Sperling, 1960.)

Rather than close this section on a question, it seems useful to remind the reader that path dependence between trials has an honorable, but sporadic, history in the experimental psychology of attention under the pseudonyms of "set" and "Einstellung." In the realm of the examples of visual attention discussed here, for example, Sperling (1960) exhibited data from a subject with textbook hysteresis who failed to switch soon enough between "equal attention" strategies and "guessing" strategies as the conditions favoring one or the other of these strategies were gradually altered between blocks (Figure 4.20). Figure 4.1 showed this subject earlier in training using a single (pure) strategy.

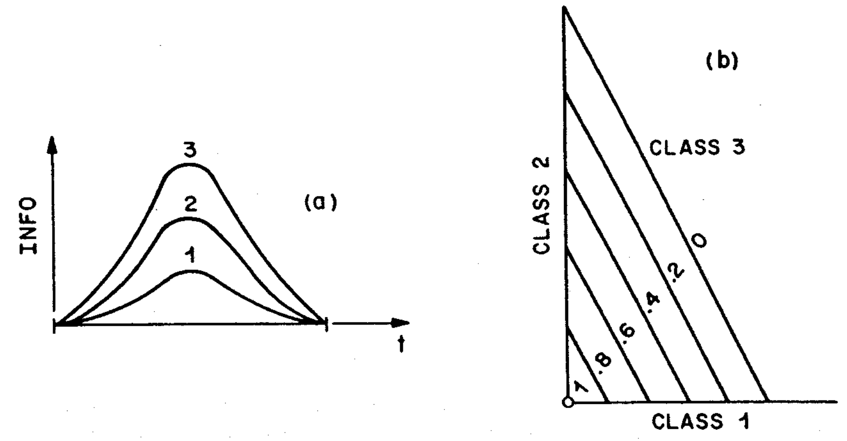


Figure 4.21 A two-dimensional strategy covers an *area* in operating space. (a) Three classes that draw on precisely the same resource pool. In an extension of previous examples, the density of information offered in each class (ordinate, INFO) is allowed to differ although the three information density functions differ only by a scale factor. (b) The amount of information a student acquires from Class 1 and Class 2 as the classroom-switching time from Class 1 to 2 is varied (the student also spends some units of time, indicated as the parameter, in Class 3 during which the student attends neither Class 1 nor 2).

Single and Multiple Resources

Single-Resource Pool

Multidimensional Strategies In the attendance analogy, a single-resource pool is exemplified by a set of competing courses offered at a particular time (e.g., noon to 1:00 p.m.) in all of which there is the same temporal distribution of information. Thus, all the competing courses use the same resource pool (the time from noon to 1:00 p.m.) with the same effectiveness, except possibly for an arbitrary scale factor that reflects the fact that performance measurements in different tasks may be incommensurate (Figure 4.21a). If the distribution of information over time $f(t)$ were not uniform, it would be possible to monotonically transform time from seconds into new units—resource units—which have the property that $\frac{1}{1000}$ of the daily lesson is covered in each resource milliunit. Insofar as a student's performance in a course is directly proportional to the number of resource milliunits for which the student has attended that course (the only assumption so far), it is trivial to compute performance when the student divides the total time between N classes. We simply add up the milliunits spent in each classroom i to determine the performance P_i in that class.

Dimensions The dimensionality of operating space equals the number of independent performance measurements. In nearly all the examples of this chapter, performance was measured on two tasks and operating space was two-dimensional. In the three-class example of Figure 4.21, operating space is three-dimensional. In two-alternative choice reaction-time experiments, operating space is four-dimensional (two speeds, two accuracies), although only the most significant two of the four dimensions have been considered in this chapter.

The trade-off between performance on two tasks—determined by a single decision criterion or a single resource-allocation parameter—is one-dimensional, a line—the POC—in two-dimensional operating space. When there are two independent strategy decisions, as in the allocation of time to three overlapping classes, the POC is a two-dimensional area in operating space—a plane in this example. Figure 4.21b illustrates various possible POCs as one strategy is varied and the other is held constant. When two strategies both vary as a consequence of an experimental manipulation, the experimentally determined POC can be almost any regular or irregular curve, being restricted only to lying on the POC surface, (e.g., the right triangle in the bottom left of Figure 4.21b).

Nonlinear Resource–Performance Function A slightly more complicated, realistic condition in which performance is monotonic with—but not directly proportional to—the number of resource units is represented by a nonlinear resource–performance function. These are the considerations introduced by Norman and Bobrow (1975) in their resource–performance function. For example, suppose the instructor tends to repeat material at random intervals. If it is assumed that listening to the repetition is wasted time, the student's performance increases slower than a linear function of milliunits because repetitions become more probable the more class units are accumulated (Figure 4.22). With repetition permitted, the marginal utility of a milliunit of class participation is a positively decreasing function of the amount of participation already accumulated. This assumption leads to a curved AOC.[23] The monotonic but nonlinear increase of performance with resource milliunits of participation is dealt with by several analytical methods, most notably conjoint measurement (Krantz, 1969) and monotonic analysis of variance (Carroll, 1972; Kruskal, 1965). These derive the nonlinear resource–performance transformation and ferret out the underlying additive structure in resource milliunits. In conclusion, when there is only one

[23] If the instructor repeats material *between* classes (on different days) as well as *within* a single class period, even the mixture of strategies would not be represented as a straight line. This would correspond to dependent trials in attention or in signal detection experiments. Although dependent trials do occur in psychophysical methods such as in the method of limits or in threshold tracking, the discussion here is confined to independent trials or, equivalently, independent topics covered in different class periods.

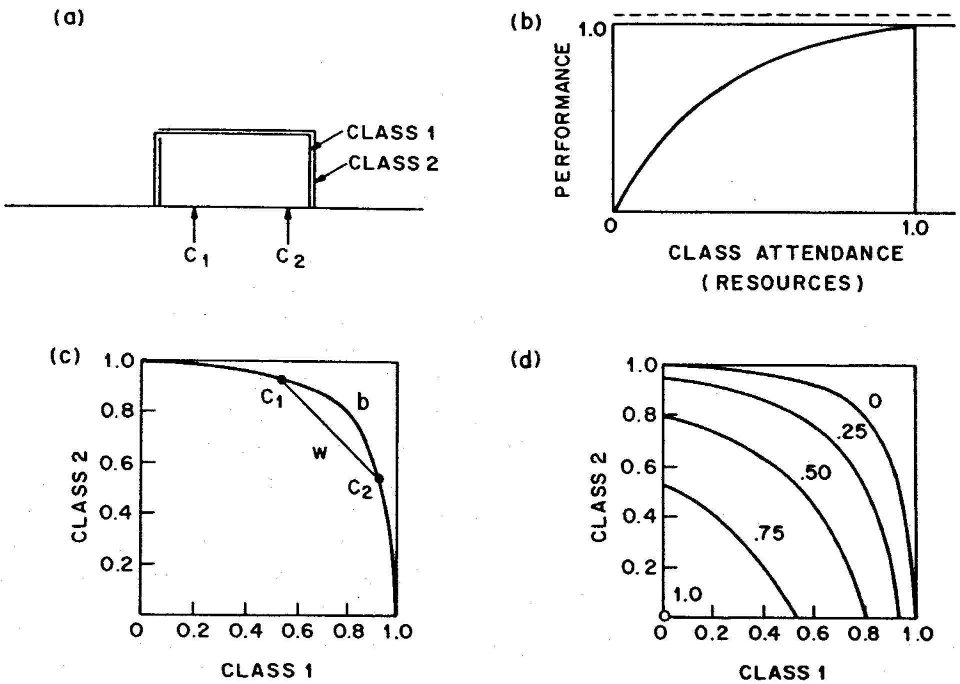


Figure 4.22 Consequences of a nonlinear resource–performance function. (a) Two classes are offered at precisely the same time. In (b), the resource–performance function in which an instructor repeats himself or herself within a class period is shown. When the student attends the entire class (attendance is 1.0), the performance also is 1.0. If the class were longer, performance would asymptote at the dashed horizontal line. In (c), the Attendance Operating Characteristic generated as classroom-switching time c varies is shown. When the student mixes two strategies, switching some days at c_1 and other days at c_2, the student's performance lies along the straight-line segment *w* (within-day repetition). If the instructor were equally likely to repeat material between days as within a day, the mixed strategy would actually lie along the segment *b* because the whole term would be, in effect, one long class period rather than a sequence of independent periods or trials. In (d), a third class is offered at precisely the same hours as the first two. If the student did not attend Class 3, the POC for Class 2 versus Class 1 will be unchanged from (c) as indicated by the POC labelled "0". If the student's participation in Class 3 increased (the curve parameter in (d), the POC would become straighter and be shifted toward the origin.

underlying resource pool shared in the same relative proportions by all the competing tasks, then, with linear resource–performance functions, description is perfectly straight forward; with nonlinear resource–performance functions, adequate methods exist for discovering and describing this situation.

Multiple-Resource Pools: Substitutability and Interference

Consider two tasks, Task 1 and Task 2. Let $u_i(r)$ be the utility of the resource *r* for Task *i*. For example, in the attendance example, a resource *r* is a particular

time interval $[t_a, t_b]$. In the single-resource pool, the utility ratio $u_2(r)/u_1(r)$ is exactly the same for all *r*. In the multiple-resource pool, $u_2(r)/u_1(r)$ varies with *r*—the situation considered in most of the previous examples of this chapter. How can the existence of multiple-resource pools be demonstrated formally.

Demonstration of Multiple Resources As has been shown in the previous sections, curvature of the POC is not a sufficient condition for concluding that there are multiple-resources; curvature could result from a nonlinear resource–performance function.

The most direct demonstration of multiple-resource pools is by way of an interaction involving the differential effect of a third task on performance in the first two tasks (see Wickens, Chapter 3, this volume). An excellent analysis and totally different kinds of examples are provided by Rachlin, Green, Kagel, and Battallo (1976) and Rachlin and Burkhardt (1978). Consider a rat given access to a dry solid food *S*1 and a liquid food *L*1. The rat spends, let us say, an equal amount of time consuming *S*1 and *L*1. Eating and drinking, respectively, can be regarded as two tasks, and consumption time as the dependent performance variable similar to a performance in attention or in classroom attendance tasks. The food and liquid dispensers are at different locations so the rat cannot simultaneously eat and drink; it has to choose to perform either one "task," or the other, or neither. When a second solid food *S*2 (a third task) is introduced, it interferes with consumption of *S*1 but not with *L*1. Similarly, a second liquid *L*2 interferes with consumption of *L*1 but not with *S*1 or *S*2.

It is hardly news that solid foods and fluids satisfy different appetites. However, casting a rat's motives into formal economic terms was an original and useful contribution by Rachlin and his co-workers (Rachlin, Batallio, Kagel, & Green, 1981; Rachlin & Burkhardt, 1978; Rachlin, Kagel, & Battalio, 1980). The subsequent treatment here differs somewhat from theirs. Figure 4.23 illustrates the three POCs appetite operating characteristics, for food *S*1 versus liquid *L*1 with the parameters: no competing task, competing *L*2, and competing *S*2. To generate points along an appetite operating characteristic, the relative amounts of deprivation (hunger or thirst) are varied. The POCs of Figure 4.23 cross each other in striking contrast to the POCs of Figures 4.21 and 4.22, which are parallel. The crossed POCs would require an interaction term in the analysis of variance, as can be seen from the more usual representation of such data in Figure 4.23b, and such data defeat any single-factor model.

The key economic concept is substitutability. Two different solid foods can substitute for each other, but eating and drinking are relatively nonsubstitutable (Rachlin *et al.*, 1976). Nonsubstitutability was treated earlier in this chapter. It occurred in the first attendance example (Figure 4.8c) in which a student had to pass each of two courses in order to graduate. This was contrasted to the case of grade point average (Figure 4.8b) in which courses are completely substitutable;

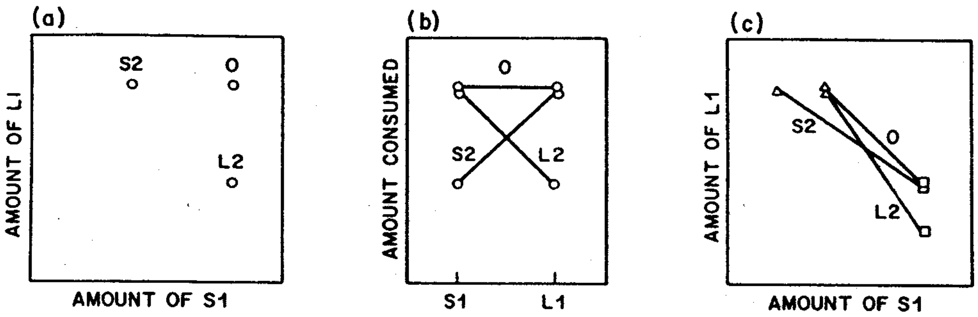


Figure 4.23 A demonstration of performance as a function of utility and interference, specifically, dry food and liquid consumption as determined by hunger and thirst and by alternative food and drink. In (a), the abscissa indicates the quantity of a dry, solid food $S1$; and the ordinate indicates the amount of liquid ($L1$) consumed by a hungry and thirsty rat. The open circle labeled zero (0) indicates that $S1$ and $L1$ were the only offerings, the circle labeled $S2$ indicates a second solid food $S2$ was offered along with $S1$ and $L1$, and the circle labeled $L2$ indicates a second liquid ($L2$) was offered with $S1$ and $L1$. In (b), a conventional representation of the data in (a) is illustrated to show interaction. The abscissa indicates what was being consumed ($S1$, $L1$), the ordinate indicates the amount consumed (of $L1$ or $S1$), and the parameter indicates the third alternative (zero, $S2$, $L2$). In (c), the appetite operating characteristics is shown. Coordinates are the same as in (a). Animals are offered $S1$, $L1$, and one of 0, $S2$, or $L2$ in different sessions. Triangles indicate hypothetical data from animals that are more thirsty (water deprived) than hungry (food deprived), and squares indicate more food than water deprivation.

a high grade in one can compensate for a low grade in the other. In the case of pass–fail, courses are completely nonsubstitutable because both have to be passed and good performance on one cannot compensate for failing performance in the other.

Interference In these examples, substitutability is a property of the utility function. Interference ultimately affects the utility that is achieved, but interference is best considered at the level of individual performance in each of the concurrent tasks. For example, in the animal experiments, eating and drinking interfere with each other in a physical sense; the animal would need two heads at opposite ends of its body to simultaneously eat and drink. Interference between two classes has been the point of all the classroom examples; attending one class totally interferes with attendance at the other. To see the interference effects of a third class, consider first two classes, one from noon until 1:00 p.m. and the other from 1:00 until 2:00 p.m. A third class scheduled from noon until 1:00 would interfere only with performance on Class 1; a class scheduled from 1:00 until 2:00 p.m. would interfere only with Class 2; a class scheduled between 12:30 and 1:30 would interfere with both Classes 1 and 2. This last class is analogous to a liquid food, such as milk, that depresses both eating of $S1$ and drinking of $L1$.[24]

Beyond variations in the content of a course itself, the classroom analogy admits two operations to manipulate the time spent in a classroom: (1) varying the relative utilities of the courses being offered and (2) varying the particular assortment of courses offered. These operations are referred to here as *varying utility* and *interference*, respectively. In the case of specific appetites, varying utility is most easily accomplished by specific deprivation; food has greater utility for a hungry animal than a sated one. Interference is accomplished by varying the assortment of foods and activities offered. The interference-generated data by themselves are somewhat cumbersome, both in collection and in utilization; so it is preferable to have access to both utility and interference data in constructing a resource model.

Interference data are probably the most widely collected data in analyzing mental processing resources. The AOC is an elementary interference method involving just two tasks. The extension to a full-fledged interference method would involve additional tasks. In detection experiments, such as Sperling and Melchner's (1978a) visual search task for numerals and/or letters, appropriate third tasks might involve memory loads (items that the subject was required to maintain in memory for subsequent recall during the search phase of the trial), irrelevant targets (targetlike items to be ignored that appear outside of the delimited search area), irrelevant auditory stimuli, etc. The various Stroop phenomena discussed by Kahneman and Treisman (Chapter 2, this volume) represent exhaustive studies of interference phenomena.

For exploiting interference data, factor analysis is an appropriate mathematical model. The appetite preference test is analogous to a battery of test items, such as the component tests of an IQ test. Each combination of competing tasks (e.g., $L1$, $S1$, $S2$) is analogous to a subject who has certain traits and abilities (e.g., $L1$, $S1$, $S2$). The factor analysis attempts to arrive at the minimum number of underlying factors needed to represent these abilities (appetites) and test items (foods and drinks). Even though factor analysis is a far more formidable technique than is demanded by any of the data currently available, it is useful to keep the interference methods and the corresponding analyses in mind when confronting these data.

[24] To complete the analogy between specific appetites and classroom attendance, additional assumptions are necessary. For example, assume classes offered between 12:00 and 1:00 are science classes, classes offered between 1:00 and 2:00 are language classes, and the student must ultimately pass both a science and a language examination. Assume it matters little to the student's performance which particular combination of classes in a category are elected. Science classes (solid food) are then substitutable for each other as are language classes (liquid foods).

Measuring the Reaction Time of a Shift of Attention

Previous sections of this chapter have been concerned with the spatial distribution of attention during stable periods when the spatial distribution is not changing. This section is concerned with the dynamics of changing attention and with the comparison of attentional dynamics to the dynamics of motor responses. The reaction time of a motor response is the time from the onset of the reaction stimulus to the onset of the required response. Both the stimulus, for example, a light flash, and the response, for example, pressing a key, are trivial to measure. In the case of attention, the stimulus is easy to measure; but measuring the attention response requires ingenuity. Attention is the allocation of mental processing resources; hence, an attention response is a shift in resource allocation. The shift is not directly observable, but it can be inferred from its consequences. Because the concern here is with an attention response that involves "grabbing" an item from a list, the attention procedure is introduced with an analogous procedure for measuring the reaction time of a motor "grabbing" response.

Measuring the "Grabbing" Response

Imagine a subject who is seated adjacent to a conveyor belt on which balls are passing by while observing a screen on which stimuli are flashed. As soon as a visual target appears on the screen, the subject reaches through a small opening that permits access to the conveyor belt and grabs the first possible ball. The balls are numbered and arranged such that, for example, the ball numbered 1 passes the opening exactly 0.1 seconds after the target, the ball numbered 2 passes the opening exactly 0.2 seconds after the target, and so on. From the number of the ball that the subject grabbed, the reaction time of the grabbing response can be inferred exactly. Of course, the subject also would know the reaction time. Imagine a long sequence of trials on which the subject has consistently grabbed Ball 5. On the next trial, the subject grabs Ball 6 or perhaps even Ball 7. When the subject is asked the number of the ball, the subject might be ashamed to tell the truth and might say 5. To keep the subject honest, the numbers must be scrambled on the balls so that the experimenter can know from the number what the grabbing time was, but the subject cannot. In fact, a random one of the numbers could even be omitted on each trial. If the subject ever reported grabbing a ball with that number, the experimenter would know immediately there was a flaw in the procedure.

The subject reports the number on the ball only seconds after it actually was grabbed. The reaction time of the response, which actually had occurred much earlier, can be inferred from the reported number. The latency of the subject's verbal report has little to do with the latency of the grabbing response; the content of the verbal report is what reveals the grabbing reaction time. Obviously, this is an indirect method of measuring a reaction time. A high-speed film of the subject's movements could have been made. From a study of the film, it could have been determined directly when the subject's hand first began to move, when it first made contact with the passing ball, when the ball first was lifted from the conveyer, and so on. These are direct measures.

In the case of motor reaction times, there is a choice of direct or indirect measures of reaction time. In the case of attention responses, there is no visible response—nothing that can be photographed; there are only indirect measures. On the other hand, little is lost in the indirect measurement. Not only the mean, but also the variance and, in fact, the whole reaction-time distribution are obtained by the indirect method. The responses are perforce quantized into discrete times—there are balls passing only every 0.1 seconds—but this is neither a serious problem nor a necessary aspect of the indirect procedure.

Attention Reaction Time Distributions

To measure the reaction of a shift of visual attention, Sperling and Reeves (1980)[25] used the following procedure. The subject maintained fixation on a fixation mark throughout a trial (see Sperling & Reeves, 1980, p. 349). To the left of fixation, a target appeared. In one series of experiments, the target was chosen at random from a letter *C*, a letter *U*, or an outline square. The target was embedded in a stream of distractors (consisting of the other letters of the alphabet) that were flashed briefly, one on top of the other, at a rate of one character/218 msec. At the right of fixation, a stream of numerals occurred (one on top of the other) at either a fast rate of from one numeral/75 msec; or, in other conditions, at rates as slow as one numeral/240 msec (see Figure 4.24).

The subject's task was to detect the target in the letter stream and then to report the first numeral he or she could from the numeral stream. The task implicitly required the subject to attend to the letter stream until the target was detected and then to shift attention to the numeral stream in order to "grab" the earliest numeral. The identity of the reported numeral is important only insofar as—like the number on the billiard ball—it indicates the numeral's temporal position. The time from the onset of the target to the onset of the named numeral defines the attention reaction time on that trial. From a block of trials, an entire attention reaction-time distribution is obtained.

[25] Many investigators have attempted to measure the speed of attentional processes. Sperling and Reeves (1976, 1978, 1980) were the first to pose the problem as one of measuring the reaction time of an attention response, to formally propose the indirect procedure, to use this indirect procedure to generate the reaction-time distribution of an attentional response, to publish an attention reaction-time distribution, and to present side-by-side comparisons of attention reaction-time and motor reaction-time distributions made in response to the same stimuli.

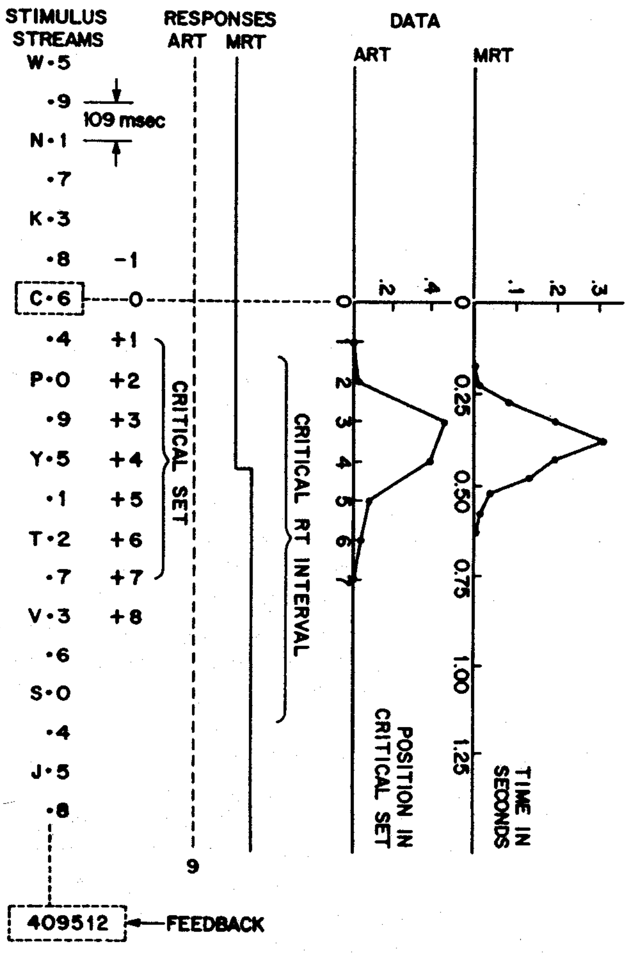


Figure 4.24 Measuring the reaction time of a shift of visual attention: procedures and typical results. Stimulus streams are shown at the left. The subject sees only the letters, the fixation dots, and the numerals. Each row represents a single display, briefly flashed and superimposed on the preceding display. The target letter is a C. The critical set of numerals and the critical reaction-time periods are indicated. The finger (motor) reaction-time key is indicated by MRT, and the attention-shift reaction (9) is indicated by ART. Feedback indicates the answer display (first six numerals of the critical set) shown to the subject after the ART response. The results show the actual observed MRT distribution and the observed ART distribution (proportion of times a numeral corresponding to each position is named) for 414 trials in this particular condition; that is, a letter stream rate of 4.6/sec, target C, and a numeral rate of 9.1/sec, in which report of one letter is tested in subject AR. (After Sperling and Reeves, 1980, Figure 17.1.)

There are certain important procedural considerations. In measuring a simple motor reaction time, for example, the interval between the warning stimulus or trial initiation and the occurrence of the target is varied randomly so that the reaction-time experiment does not degenerate into an experiment in time estimation. Moreover, the experimenter cannot simply instruct the subject to "respond as quickly as possible." There must be explicit contingencies so that responses that occur before the target stimulus are punished as premature "anticipations," and responses occurring too late are punished for being "slow." The effect of such restrictions is to define a critical interval within which the response is supposed to occur, (e.g., from 100 to 800 msec after target occurrence), and to reward the subject for responding as early as possible within the critical interval.

Similarly, in measuring an attention reaction time, one cannot simply instruct the subject to name the earliest numeral possible. Rather, a critical interval is defined; and the subject is instructed to grab the earliest possible numeral from that interval. As with the motor reaction-time procedure, the beginning of the critical interval is placed so early that it cannot be achieved by a legitimate response. An additional complication in the attention reaction-time procedure is that the numerals within the critical interval, as well as the one or two before and after it, are all arranged to be different so that the numeral's identity unambiguously indicates its position in the stream.

Using the motor reaction-time and attention reaction-time methods as outlined, Reeves (1977) simultaneously measured motor reaction times and attention reaction times for 3 subjects in 17 conditions, obtaining a total of over 50 pairs of attention reaction-time and motor reaction-time distributions. A representative pair of motor reaction-time and attention reaction-time distributions is illustrated in Figure 4.24. Although they are measured in completely different ways—motor reaction time by a direct method and attention reaction time by an indirect method—they are quite comparable in terms of their mean and variance. These data are typical of the attention reaction-time procedure when the numeral stream occurs at a high rate (7/sec or faster); with slow numeral rates, attention reaction times become shorter than motor reaction times. Attention reaction times respond similarly to motor reaction times with manipulations of target difficulty (motor reaction times and attention reaction times get slower for hard-to-detect targets) or target probability (motor reaction times and attention reaction times get faster as the likelihood of a target increases). The interested reader will find a theory for these and other interesting features of motor reaction-time and attention reaction-time distributions in Sperling and Reeves (1980).

The point of the present discussion is that the indirect method yields measures of attention reaction times—unobservable responses—that are no less reliable and no more difficult to obtain than directly observable motor reaction times. Unobservable does not mean unmeasurable. The way to measure a reallocation of mental processing resources is by its effect—by how soon the reallocated resources have an effect. This indirect method can, in principle, be extended to any of the attentional tasks described in this book.

Optimization: A Last Word

Optimization—finding the most favorable compromise between conflicting goals or demands—has been the unifying principle throughout this chapter. It is a Darwinian principle. In the classroom attendance example, optimization involved the best choice of which class to attend when there were schedule conflicts. In signal detection tasks, it was the choice of a criterion that optimized the expected rewards for correct responses minus the expected costs for errors. It is essential to recognize that in all real situations there is imperfect information, and errors are possible. The signal detection theory of optimization is applicable far beyond the case of humans detecting threshold auditory signals in noise. The same optimization principles apply, for example, to seeds that must decide whether to sprout now or to wait for a better time, to juries that must balance the risk of punishing an innocent person against the risk to society of releasing a criminal, or to a human body's immunological system that must decide whether an unknown substance is a dangerous germ to be fought off or a part of itself to be left alone. The principles of optimum decision making transpose to cases of optimum resource allocation in selective attention—such as the optimal location of mental processing resources to one of several competing tasks, or of economic behavior—such as that of animals in motivation and reinforcement experiments; or, originally and ultimately, of economic issues—such as the optimal allocation of resources by consumers, by industries, and by entire societies.

With respect to psychology, application of the calculus of optimization represents an explicit reintroduction of *purpose* to the explanation of human performance by once again explaining behavior in terms of its goals. Explicit knowledge of the utility function is necessary to understand the decision process in signal detection tasks or the allocation of mental processing resources in attentional tasks—decisions that are governed by utility. In this respect, optimization theory is a link between many branches of psychology, and it will link the psychology of the future to the neglected psychology of the past in which purpose once had been an essential ingredient of psychological theory.

Reference Notes

1. Sperling, G. *Measuring Attention.* Invited address presented at the meeting of the American Psychological Association, Montreal, Quebec, September 4, 1980.

2. Sperling, G. *A unified theory of attention and signal detection.* Invited address presented at the fifteenth annual meeting of the Society for Mathematical Psychology, Princeton University, Princeton, N.J., August 8, 1982.
3. Budlansky, Judy T., and Sperling, G., *GS Letters: A general purpose system for producing visual displays in real time and for running psychological experiments on the DOP24 computer.* Unpublished technical memorandum, 1969. Bell Laboratories, Murray Hill, N.J..
4. Sperling, G., and Harris, J. R. Unpublished experiments, 1976–1977, Bell Laboratories, Murray Hill, N.J.

References

Anstis, S. M. A chart demonstrating variations in acuity with retinal position. *Vision Research,* 1974, *14,* 589, 592.

Audley, R. J. Some observations on theories of choice reaction time: Tutorial review. In S. Kornblum (Ed.), *Attention and performance IV.* New York: Academic Press, 1973.

Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology,* 1975, *12,* 387–415.

Bamber, D., and van Santen, J. P. H. How many parameters can a model have and still be testable? *Journal of Mathematical Psychology,* 1983, in press.

Bouma, H. Visual search and reading: Eye movements and functional visual field: A tutorial review. In J. Requin (Ed.), *Attention and performance VI.* Hillsdale, N.J.: Erlbaum, 1978, 115–147.

Carroll, J. D. Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, and S. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences,* (Vol. 1, Theory). New York: Seminar Press, 1972.

Cohn, T., & Lasley, D. Detectability of a luminance increment: Effect of spatial uncertainty. *Journal of the Optical Society of America,* 1974, *64,* 1715–1719.

Coombs, C. H., & Avrunin, G. S. Single-peaked functions and the theory of preference. *Psychological Review,* 1977, *84,* 216–230.

Dosher, B. A. The retrieval of sentences from memory: A speed–accuracy study. *Cognitive Psychology,* 1976, *8,* 291–310.

Dosher, B. A. The effects of delay and interference: A speed–accuracy study. *Cognitive Psychology,* 1981, *13,* 551–582.

Due, J. F. *Intermediate economic analysis.* Chicago: Irwin, 1951.

Duncan, J. The demonstration of capacity limitation. *Cognitive Psychology,* 1980, *12,* 75–96.

Egan, J. P. *Signal detection theory and ROC analysis.* New York: Academic Press, 1975.

Falmagne, J. C. Note on a simple property of binary mixtures. *British Journal of Mathematical and Statistical Psychology,* 1968, *21*(1), 131–132.

Falmagne, J. C., Cohen, S. P., and Swivedi, A. Two-choice reactions as an ordered memory scanning process. In P. Rabbitt and S. Dornic (Eds.), *Attention and performance V.* London: Academic Press, 1975. Pp. 296–344.

Fitts, P. M. Cognitive aspects of information processing (Vol. 3): Set for speed versus accuracy. *Journal of Experimental Psychology,* 1966, *71,* 849–857.

Galambos, J. *The asymptotic theory of extreme order statistics.* New York: Wiley, 1978.

Gilliom, J. D., & Sorkin, R. D. Sequential and simultaneous two-channel signal detection: More evidence for a high level interrupt theory. *Journal of the Acoustical Socity of America,* 1974, *56,* 157–164.

Green, D. M., & Luce, R. D. Speed–accuracy tradeoff in auditory detection. In S. Kornblum (Ed.), *Attention and performance IV*. New York: Academic Press, 1973. Pp. 547–569.

Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

Gumbel, E. J. *Statistics of extremes*. New York: Columbia University Press, 1958.

Harris, J. R., Shaw, M. L., & Bates, M. Visual search in multicharacter arrays with and without gaps. *Perception & Psychophysics*, 1979, *26*(1), 69–84.

Hicks, J. R., & Allen, R. G. D. A reconsideration of the theory of value. *Economica*, 1934, *1*, 52–76, 196–219.

Hoffman, J. E., & Nelson, B. *Spatial selectivity in visual search.* Perception and Psychophysics, 1981, *30*, 283–290.

Howarth, C. I., & Lowe, G. Statistical detection theory of Piper's Law. *Nature*, 1966, *212*, 324–326.

Kantowitz, Barry H. Double stimulation. In B. H. Kantowitz (Ed.), *Human information processing: Tutorial in performance and cognition*. Hillsdale, N.J.: Erlbaum, 1974. Pp. 83–131.

Kinchla, R. A. The role of structural redundancy in the detection of visual targets. *Perception & Psychophysics*, 1977, *22*, 19–30.

Kinchla, R. A. The measurement of attention. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, N.J.: Erlbaum, 1980.

Kinchla, R. A., & Collyer, C. E. Detecting a target letter in briefly presented arrays: A confidence rating analysis in terms of a weighted additive effects model. *Perception & Psychophysics*, 1974, *16*, 117–122.

Kowler, E., & Steinman, R. M. The effect of expectations on slow oculomotor control—II: Single target displacements. *Vision Research*, 1979, *19*, 633–646.

Kowler, E., & Steinman, R. M. The effect of expectations on slow oculomotor control (Vol. 3): Guessing unpredictable target displacements. *Vision Research*, 1981, *21*, 191–203.

Krantz, David H. Threshold theories of signal detection. *Psychological Review*, 1969, *76*, 308–324.

Kruskal, J. B. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society*, series B, 1965, *27*, 251–263.

Laming, D. R. *Information theory of choice-reaction times*. New York: Academic Press, 1968.

Levitt, H. Decision theory, signal detection theory, and psychophysics. In E. E. David and P. B. Denes (Eds.), *Human communication: A unified view*. New York: McGraw-Hill, 1972, 114–174.

Link, S. W. The relative judgment theory of two-choice response time. *Journal of Mathematical Psychology*, 1975, *12*, 114–135.

Link, S. W., & Heath, R. A. A sequential theory of psychological discrimination. *Psychometrika*, 1975, *40*, 77–105.

Mertens, J. J. Influence of knowledge of target location upon the probability of observation of peripherally observable test flashes. *Journal of the Optical Society of America*, 1956, *46*, 1069–1070.

Metz, C. E., Starr, S. J., Lusted, L. B., & Rossmann, K. Progress in evaluation of human observer visual detection performance using the ROC curve approach. In C. Raynaud and A. Todd-Pokropek, Eds., *Information processing in scintigraphy*. Orsay, France: Commissariat a l'Energie atomique, Departement de Biologie, Service Hospitalier Frederic Joliot, 1975. Pp. 420–439.

Murphy, B. J. Pattern thresholds for moving and stationary gratings during smooth eye movements. *Vision Research*, 1978, *18*, 521–530.

Murphy, B. J., Kowler, E., & Steinman, R. M. Slow oculomotor control in the presence of moving backgrounds. *Vision Research*, 1975, *15*, 1263–1268.

Navon, D., & Gopher, D. On the economy of the human-processing system. *Psychological Review*, 1979, *86*,(3), 214–255.

Neisser, U. Decision time without reaction time: Experiments in visual scanningl *American Journal of Psychology*, 1963, *76*, 376–385.

Neisser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1966.

Neisser, U., Novick, R., & Lazar, R. Searching for ten targets simultaneously. *Perceptual and Motor Skills*, 1963, *17*, 955–961.

Norman, D. A., & Bobrow, D. G. On data-limited and resource-limited processes. *Cognitive Psychology*, 1975, *7*, 44–64.

Ollman, Robert. Fast guesses in choice reaction time. *Psychonomic Science*, 1966, *6*, 155–156.

Pachella, R. G., & Fisher, D. Hick's Law and the speed–accuracy trade-off in absolute judgment. *Journal of Experimental Psychology*, 1972, *92*, 378–384.

Pareto, V. *Manuel d'economie politique*. 1909.

Pohlman, L. D. & Sorkin, R. D. Simultaneous three-channel signal detection: Performance and criterion as a function of order of report. *Perception and Psychophysics*, 1976, *20*, 179–186.

Posner, M. I., Nissen, M. J., & Ogden, W. C. Attended and unattended processing modes: The role of set for spatial location. In H. I. Pick, Jr., & E. Saltzman (Eds.), *Modes of perceiving and processing information*. Hillsdale, N.J.: Erlbaum, 1978.

Rachlin, H., Battalio, R., Kagel, J., & Green, L. Maximization theory in behavioral psychology. *The Behavioral and Brain Sciences*, 1981, *4*, 371–417.

Rachlin, H., & Burkhardt, B. The temporal triangle: Response substitution in instrumental conditioning. *Psychological Review*, 1978, *85*, 22–47.

Rachlin, H., Green, L., Kagel, J. H., & Battalio, R. C. Economic demand theory and psychological studies of choice. In G. Bower (Ed.), *The psychology of learning and motivation*, (Vol. 10). New York: Academic Press, 1976.

Rachlin, H., Kagel, J. H., & Battalio, R. C. Substitutability in time allocation. *Psychological Review*, 1980, *87*, 355–374.

Reed, A. V. Speed–accuracy trade-off in recognition memory. *Science*, 1973, *181*, 574–576.

Reeves, A. *The detection and recall of rapidly displayed letters and digits*. Unpublished doctoral dissertation, City University of New York, 1977.

Rubenstein, H., Garfield, L., & Millikan, J. A. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 487–494.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior*, 1971a, *10*, 645–657.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. Homographic entries in the internal lexicon: Effects of systematicity and relative frequency of meanings. *Journal of Verbal Learning and Verbal Behavior*, 1971b, *10*, 57–62.

Samuelson, P. A. *Economics* (11th ed.). New York: McGraw-Hill, 1980.

Schneider, W., and Shiffrin, R. M. Controlled and automatic human information processing: I Detection, search, and attention. *Psychological Review*, 1977, *1*, 1–66.

Schuckman, H. Attention and visual threshold. *American Journal of Optometry and Archives of the American Academy of Optometry*, 1963, *40*, 284–291.

Shaw, P. Processing of tachistoscopic displays with controlled order of characters and spaces. *Perception & Psychophysics*, 1969, *6*, 257–266.

Shaw, M. L., and Shaw, P. Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 201–211.

Shepard, R. N. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 1958, *55*, 509–523.

Sorkin, R. D., Pastore, R. E., & Pohlmann, L. D. Simultaneous two-channel signal detection: II. Correlated and uncorrelated signals. *Journal of the Acoustical Society of America*, 1972, *51*, 1960–1965.

Sorkin, R. D., Pohlmann, L. D., & Gilliom, J. D. Simultaneous two-channel signal detection: III. 630- and 1400-Hz signals. *Journal of the Acoustical Society of America*, 1973, *53*, 1045–1050.

Sorkin, R. D., Pohlmann, L. D., & Woods, D. D. Decision interaction between auditory channels. *Perception and Psychophysics*, 1976, *19*, 290–295.

Sperling, G. *Information available in a brief visual presentation*. Unpublished doctoral dissertation, Department of Psychology, Harvard University, 1959.

Sperling, G. The information available in brief visual presentations. *Psychological Monographs*, 1960, *74*(11, whole No. 498).

Sperling, G. A model for visual memory tasks. *Human Factors*, 1963, *5*, 19–31.

Sperling, G. Temporal and spatial visual masking. I. Masking by impulse flashes. *Journal of the Optical Society of America*, 1965, *55*, 541–559.

Sperling, G. Extremely rapid visual scanning. *Bulletin of the British Psychological Society*, 1970a, *23*, 58.

Sperling, G. Binocular vision: A physical and a neural theory. *American Journal of Psychology*, 1970b, *83*, 461–534.

Sperling, G. The search for the highest rate of search, *Symposium on Attention and Performance*, August 1973, *5* Saltsjobaden, Stockholm, Sweden.

Sperling, G. Multiple detections in a brief visual stimulus: The sharing and switching of attention. *Bulletin of the Psychonomic Society*, 1975, *9*, 427. (Abstract)

Sperling, G. Mathematical models of binocular vision. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology*. American Mathematical Association (SIAM–AMS) Proceedings, 1981, *13*, 281–300.

Sperling, G. Unified theory of attention and signal detection. *Mathematical Studies in Perception and Cognition*. New York University, Department of Psychology, 1983, *83*(3), 1–64.

Sperling, G., Budiansky, J., Spivak, J. G., and Johnson, M. C. Extremely rapid visual search: The maximum rate of scanning letters for the presence of a numeral. *Science*, 1971, *174*, 307–311.

Sperling, G., & Melchner, M. J. Visual search and visual attention. In V. D. Giezer (Ed.), Information processing in visual system. *Proceedings of the Fourth Symposium of Sensory System Physiology*. Leningrad, U.S.S.R.: Academy of Sciences, Palov Institute of Physiology, 1976a, 224–230.

Sperling, G., & Melchner, M. J. Estimating item and order information. *Journal of Mathematical Psychology*, 1976b, *13*, 192–213.

Sperling, G., & Melchner, M. J. The attention operating characteristic: Some examples from visual search. *Science*, 1978a, *202*, 315–318.

Sperling, G., & Melchner, M. J. Visual search, visual attention, and the attention operating characteristic. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, N.J.: Erlbaum, 1978b, 675–686.

Sperling, G., & Reeves, A. Reaction time of an unobservable response. *Bulletin of the Psychonomic Society*, 1976, *10*, 247. (Abstract)

Sperling, G., & Reeves, A. Measuring the reaction time of a shift of visual attention. *Investigative Ophthalmology and Visual Science*, (ARVO Supplement), 1978, *17*, 289. (Abstract)

Sperling, G., & Reeves, A. Measuring the reaction time of a an unobservable response: A shift of visual attention. In R. Nickerson (Ed.), *Attention and Performance VIII*. Hillsdale, N.J.: Erlbaum, 1980, 347–360.

Stroop, J. R. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 1935, *18*, 643–662.

Sussman, H. J., & Zahler, R. S. Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese*, 1978, *38*, 117–216.

Swets, J. A. (Ed.), *Signal detection and recognition by human observers: Contemporary readings*. New York: Wiley, 1964.

Swets, J. A. The relative operating characteristic in psychology. *Science*, 1973, *182*, 990–1000.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. Decision processes in perception. *Psychological Review*, 1961, *68*, 301–340.

Thom, R., & Zeeman, E. Catastrophe Theory: Its Present State and Future Perspectives. In A. Manning (Ed.), *Dynamical Systems—Warwick 1974: Proceedings of a Symposium Held at the University of Warwick 1973/74*. Berlin: Springer-Verlag, 1975b.

Wald, A. *Statistical decision functions*. New York: Wiley, 1950.

van Santen, J. P. H., and Bamber, D. Finite and infinite state confusion models. *Journal of Mathematical Psychology*, 1981, *24*, 101–111.

Welford, A. T. (Ed.), *Reaction time*. London: Academic Press, 1980.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. Priming and retrieval from short-term memory: A speed accuracy analysis. *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 387–404.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology* (rev. ed.). New York: Holt, 1954, Ch. 17, 492–527.

Yellot, John I. Correction for guessing in choice reaction time. *Psychonomic Science*, 1967, *8*, 321–322.

Yellott, John I. Correction for fast guessing and the speed–accuracy tradeoff. *Journal of Mathematical Psychology*, 1971, *8*, 159–199.

Zeeman, E. C. Catastrophe theory. *Scientific American, April* 1976, *234*, 65–83.