

CHAPTER 1

BENNET B. MURDOCK, JR.

Recognition Memory

The term *recognition memory* can have at least two quite different meanings. In one sense, it refers to the act of recognizing someone or something: a familiar face, a familiar piece of music, a familiar scene. (See Mandler, 1980, for a discussion of this meaning.) In the second sense, it refers to a method of testing memory that presents one or more alternatives and asks for a judgment of familiarity. I shall concentrate on the second meaning, though at times I must consider the first as well. Recognition is one of the oldest and best established techniques of testing memory. Much is known about recognition memory. By now we have standard techniques and, as a result of their use, much data and theory. The techniques are popular today, and will likely continue to be widely used.

Designing experiments in recognition memory can sometimes get tricky. As an example, see the rather elaborate precautions necessary to guard against confoundings in some tests of a scanning model (Murdoch, Hockley, & Muter, 1977). In this chapter, I will not go into such detail; rather, I shall indicate only general principles.

It has long been known that retention is a function of the method of measurement. In one of the early classic studies Luh (1922; see also Postman & Rau, 1957) compared recognition, relearning, reconstruction, written reproduction, and anticipation. The recognition scores were the

highest. This finding agrees with our intuitions; we have all had situations in which we could not recall something but could recognize it. It is appropriate that a book on research methods in memory and cognition start with recognition.

It is generally thought that the distinction between recall and recognition involves whether or not the alternatives are presented to the subject. If I ask, *What is the capital of X?* I am testing recall, but if I ask, *What is the capital of X: A, B, C, D?* then I am testing recognition. This distinction may be too simple. Another distinction (Murdock, 1970; Tulving, 1976) might be in terms of the availability of the alternatives. If the subject can readily generate all possible alternatives or if they are present, then the method would be recognition; otherwise, it would be recall. Thus, the question *What is the fifteenth letter of the alphabet?* would be testing recognition, not recall, even though the form of the question implies recall. Actually, even such a basic distinction as this is probably a theoretical issue, and cannot be decided arbitrarily.

DEFINITIONS

Basic Processes

The three basic processes of memory are *encoding*, *storage*, and *retrieval*. Encoding refers to the process by which information presented to a person is transformed from some physical form (light waves or sound waves impinging on the receptors) into a memorial representation (memory trace or engram). Storage refers to the persistence of this information over time. Forgetting is generally considered to be a storage phenomenon, since the information does not persist with adequate fidelity to support remembering. Retrieval is the utilization of this stored information at the time of test, whatever the test: recall, recognition, or some other. Failure to retrieve information does not necessarily imply forgetting; the difficulty could be in accessing the information or in the initial encoding. Encoding, storage, and retrieval are sequential processes—encoding first and retrieval last. Successful performance on a memory test means that all three processes are above some minimum required level. However, failure on a memory test is not sufficient, by itself, to localize which process may be at fault.

Stages

Utilization of stored information, or retrieval, involves two stages: *memory* and *decision*. This point is illustrated in the simple flow chart

1. RECOGNITION MEMORY

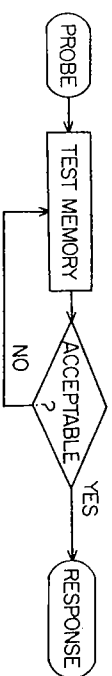


Figure 1.1. A flow chart showing the role of memory and decision.

of Figure 1.1. The output from the memory system is input into the decision system, and, speaking metaphorically, the decision system decides whether to output a response or query the memory system again. Suppose a short list of items is presented to a person and is followed by a single item about which a "yes" or "no" decision is required: ("yes," the item had been in the list, or "no," the item had not been in the list). Somehow the person makes a comparison between the encoded version of the probe item and the memory trace(s) comprising the list. If a clear match or nonmatch occurs, the person has enough information to give a "yes" or "no" response. In the event of uncertainty, the person might quiz the memory system again. Decision factors certainly affect performance on many tests of recognition memory. Failure to take these two stages of information processing into account may lead to erroneous conclusions on the part of the experimenter. The section on accuracy will deal with this problem in detail.

Types of Information

By now it seems quite clear that there are at least three different types of information in human memory. These are *item information*, *associative information*, and *serial-order information*. Item information is information that some item, object, or event has occurred in the past. Item information may be tied to some context—have you seen the word *decision* in this chapter? Associative information relates two items, objects, or events. Common examples are names and faces, words and their meanings, and artists and their works. The discussion of the association of ideas and the role of contiguity and similarity has been part of Western culture since the time of the early Greek philosophers. Serial-order information is sequence information, information about the order in which items or events occur. We all know the days of the week, the letters of the alphabet, and, in spelling, that "i" comes before "e" except after "c." In a more naturalistic vein, we could put many events of our past life in chronological order if we were told the events and asked to arrange them in the order in which they occurred. How item, associative, and serial-order information is encoded, stored, and retrieved is far from

a settled matter, but there is abundant experimental evidence to support these distinctions (Murdock, 1974).

The distinction between recall and recognition is orthogonal to the distinction between item, associative, and serial-order information. To name the first five presidents of the United States requires recall of serial-order information, whereas a subject asked to arrange pictures of items in the order in which they were just presented requires recognition. Thus, one can use recall or recognition with serial-order information. Likewise with associative information: If I present the pairs A-B and C-D, I can test for recall with A-? or C-?, or I can test for recognition by asking whether A-B or A-D was a correct pairing. Whether one can have recall of item information is a moot point; one could argue that the experimental paradigm of free recall would be an example. On the other hand, it could be that free recall involves associative or serial-order information. In any case, it is important to appreciate that recognition is not just for item information. One can test for recognition of item, associative, or serial-order information.

Types of Tests

There are three main types of tests that are used in the study of recognition memory: *yes-no tests*, *forced-choice tests*, and *batch testing*. Yes-no (or true-false) tests simply require a binary response to a question. *Was item X in the list, or Is Brasilia the capital of Brazil?* In forced-choice tests, *m* alternatives are presented, of which *m* - 1 are incorrect and one is correct. The multiple-choice test, so popular in large undergraduate classes, is a well-known example. The guessing, or chance, level is lower in forced-choice than in yes-no, though by how much depends on the value of *m*. In signal-detection theory there is a standard way of relating performance on yes-no tests to performance on *m*-alternative forced-choice tests, and forced-choice tests are often considered to bypass the criterion (or guessing) problem. Both these matters will be discussed in the section on accuracy. Finally, in batch testing all alternatives (correct and incorrect) are simultaneously presented to the subject, who must then work through the items, indicating in some fashion the old and new items. This procedure is less popular than the first two, and with good reason. For one thing, the experimenter loses control of the pacing of the test session. Also lost is the opportunity to measure the latency of each response. Finally, continual adjustment of the subject's own criterion (generally considered to be undesirable) is quite easy because the subject can always review the answer sheet.

1. RECOGNITION MEMORY

Types of Procedures

There are two main types of procedures used in the study of recognition memory: the *study-test procedure* and the *continuous-task procedure*. The study-test procedure was (to my knowledge) first used by Strong (1912, 1913), and it has not changed greatly since then. A list of items (words, pictures, sentences, advertisements) is shown to the subject, followed by a test list. Whereas forced-choice or yes-no can be used, assume the test is yes-no. Old and new items are shown, generally in randomized order, and the subject must respond to each as it occurs. One trial, then, consists of a single study-test sequence. Generally a number of trials are given in a single experimental session. The continuous-task procedure was introduced by Shepard and Teghtsoonian (1961). Here, study and test items are randomly intermingled so there is no clear separation of the study and test phase. The subject responds to each item as it appears. On the first presentation of an item, the correct response is "no" (*I have not seen this item before in this list*) and on subsequent presentations the correct response is "yes" (*I have seen this item before in this list*). Thus, the subject responds to each item each time it appears from the beginning of the list to the end, so there is no clear separation of the study and the test phase of the experiment.

The single-item probe technique is a popular variant of the study-test procedure. A short list of items is presented, followed by a single-item probe. The one item constitutes the test phase. This procedure has become popular through the work of Sternberg (1966, 1969, 1975), who has used this method to obtain latency data under conditions of relatively error-free performance. Whereas one could consider this method a special type of procedure, it seems more logical to view it as a special case of the study-test procedure. (Parenthetically, it might be noted that there is an "inertia" effect in switching over from the study phase to the test phase in a study-test procedure. The first few response latencies are abnormally high, so this might pose problems in comparing study-test data with single-item probe data. Evidence for this inertia effect may be found in Murdock and Anderson [1975].)

Experimental Measures

The three main measures in the study of recognition memory are *accuracy*, *latency*, and *confidence*. Accuracy is generally determined by the proportion of correct responses, either summed over old and new

items in a yes-no procedure or given separately for each. (The d' of signal detection theory is a derived measure and will be described in the section on accuracy.) Latency is the length of time elapsing between presentation of the test item and initiation of the response. (Visual rather than auditory presentation obviates the problem of determining precisely the onset of a spoken word, and key presses rather than spoken responses do likewise on the response side.) Confidence judgments are the subject's assessment of his own accuracy, and they can be integrated with yes-no responses or given separately. If given separately, the subject first makes the yes-no response and then evaluates it on a separate confidence-judgment scale. If integrated, a 6-point scale—"sure no" (---), "probably no" (---), "maybe no" (-), "maybe yes" (+), "probably yes" (+++), and "sure yes" (+++)—is popular. The data seem quite comparable for these two procedures (Mandler & Boeck, 1974). Finally, latency and confidence can be combined by simply recording the length of time to make each confidence judgment.

ACCURACY

One cannot discuss accuracy of recognition memory in any depth without recourse to signal-detection theory. Consequently, this section on accuracy will be devoted to an account of signal-detection theory and its application to recognition memory.

Signal-detection theory has been widely used in the study of recognition memory. This application stems from a technical report by Egan (Note 1), who not only spelled out in considerable detail the nature of the application, but also provided some relevant data. Other early references are Murdock (1965) and Norman and Wickelgren (1965). There are probably two main advantages in using a signal-detection analysis of recognition memory. First, it provides an economical summary statistic (d') to characterize overall accuracy. Otherwise, there are two separate measures, one for performance on old items, and one for performance on new items. The d' measure combines them both. One measure is much easier to handle than two. Second, signal-detection theory provides a clean way of separating memory and decision (see Figure 1.1). Thus, even though both memorial and decision factors affect performance, the theory allows separation of the two effects in the data.

Standard Signal-Detection Analysis

The basic idea of signal-detection analysis is that there are two underlying distributions: a noise (or new-item) distribution and a signal

1. RECOGNITION MEMORY

(or old-item) distribution. As far as recognition memory is concerned, it does no great harm to think of these as "strength" distributions. The concept of memory-trace strength is old and venerable and, though no one quite knows what it means, many feel comfortable in using it. The general idea is that new items (items not presented in the study list) have some variability. This variability is the starting point for the theoretical development. Old items (items presented in the study list) have comparable, and in some cases, equal, variability, but their mean strength is higher. It is conventional to represent this state of affairs with overlapping normal distributions—see Figure 1.2.

Not only are there two distributions, but there is also a criterion, or cutoff, as well. The recognition process is assumed to go as follows. The subject is presented with the probe item and must interrogate memory to determine whether the probe item is an old item or a new item. The probe is compared to the contents of memory and the result of this comparison process must be assessed. The assessment is the decision process and is logically equivalent to the statistical problem of inferring the origin of a single observation drawn at random from one of two possible distributions. The criterion is the cutoff for the decision. If the observation falls below the criterion, the subject says "no," but if the observation falls above, the subject says "yes." The criterion thus constitutes the line dividing "yes" and "no."

A few comments are now in order. First, on a single trial (i.e., a single probe test) the assessment of the probe has a fixed value. The variability is the variability of many such samples pooled over trials. If the distributions did not overlap, there would be no problem. Observations falling in the range of the lower distribution would always be called new, whereas observations falling in the range of the upper distribution would always be called old, and the subject would be 100% accurate. Since in practice this does not occur (conditions where it might occur are not very informative, so the experimenter makes sure it does

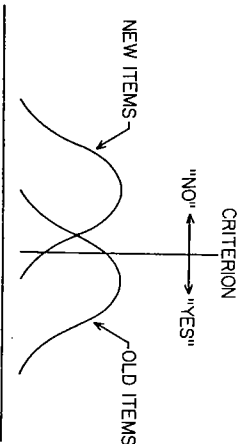


Figure 1.2. The underlying old- and new-items distributions as suggested by an application of signal-detection theory to recognition memory.

not occur), the criterion is necessary to partition the responses in the case of uncertainty. Second, there may be some variability in the placement of the criterion. Criterion variability has the effect of lowering d' (making the distributions seem closer together; see, for example, McNicol, 1972), but trying to separate criterion variability from other sources of variability is not an easy undertaking. Third, an extension to latency data has been made by Norman and Wickelgren (1969), who suggested that latency was an inverse function of distance from the criterion. Thus, extreme observations, both "yes" and "no," are made quickly, but intermediate observations take longer. Whereas this suggestion has considerable intuitive appeal, some experimental problems have been pointed out by Murdock and Duffy (1972).

Definitions

A number of standard definitions from signal-detection theory will be given next. There are four types of responses: *hits*, *false alarms*, *correct rejections*, and *misses*. Hits are "yes" responses to old items. False alarms are "yes" responses to new items. Correct rejections are "no" responses to new items. Misses are "no" responses to old items. Hits and correct rejections are correct responses, misses and false alarms are errors. These four types of responses can be reported as frequencies, but they are more commonly reported as conditional probabilities. The hit rate would then be the proportion of old items given a "yes" response, and the false alarm rate would be the proportion of new items given a "yes" response. These are not the only two conditional probabilities that could describe the data, but are commonly used. This terminology is illustrated in Figure 1.3.

ITEM TYPE	RESPONSE	
	"YES"	"NO"
OLD	HIT	MISS
NEW	FALSE ALARM	CORRECT REJECTION

Figure 1.3. The terminology for "yes" and "no" responses to old and new items.

1. RECOGNITION MEMORY

The two basic dependent variables are d' and β . A simple formula for d' is

$$d' = z(\text{FA}) - z(\text{H}) \quad (1)$$

where $z(\text{FA})$ is the standard (or z) score corresponding to the given false alarm rate and $z(\text{H})$ is the standard score corresponding to the given hit rate (see, e.g., McNicol, 1972). As an example, suppose, out of 120 old items, the subject said "yes" 102 times; the hit rate would be .850. Suppose, out of 120 new items, the subject said "yes" nine times; the false alarm rate would be .075. From tables of the normal distribution, $z(\text{H}) = -1.04$ and $z(\text{FA}) = 1.64$. Consequently, $d' = 1.64 - (-1.04) = 2.68$.

One can think of d' itself as a simple standard score; it is the mean of the old-item distribution expressed in units of the standard deviation of the new-item distribution. (The assumption is always that the mean of the new-item distribution is zero.) As a dependent variable, d' is generally regarded as a measure of the average strength of the old items.

The second measure, β , is a measure of the location or placement of the criterion on the strength axis. Technically, it is the ratio of the ordinates of the two distributions (old to new) at the criterion. It can easily be computed from any table of the unit normal curve given, for example, the hit rate and the false alarm rate. If $\beta = 1.0$, then the criterion is located at the point of intersection of the old- and new-item distributions. Other measures of the criterion are possible and sometimes preferable (see, e.g., Donaldson & Glathie, 1970). One such measure is the location of the criterion as a standard score relative to the mean of the new-item distribution.

The a priori probability is the probability that a probe item will be an old item. This probability is fixed by the experimenter, and generally is told to the subjects so they can set their criterions appropriately. The a posteriori probability is the probability that a probe item was old, given a "yes" response. This probability is not fixed by the experimenter; rather, it depends upon the subject's performance. It is quite useful in confidence-judgment experiments, since it gives an indication of how accurate the subject is at different points along the strength dimension. Costs and payoffs refer to the values associated with errors and correct responses. Costs are negative values and payoffs are positive values. A full payoff matrix specifies the cost for both types of error and the payoff for both types of correct responses. From this and the a priori probability, one can derive the optimum criterion placement.

The optimum criterion placement is that placement of the criterion which maximizes the "net profit" (in a monetary sense, what is won

from being correct minus what is lost from being incorrect). In fact, this is why β is used in signal-detection theory. According to a maximum-likelihood principle, β should equal $p/(1 - p)$ where p is the a priori probability. If $p = .5$, and the payoff matrix is symmetric, then, by this principle, β should equal 1.0. That is, the criterion should be located at the point of intersection of the two distributions. This analysis is based on a rational analysis of the subject's performance in a recognition-memory task. Whether the subjects actually behave in this fashion (and at times they clearly do not) is an open question.

Experimental Manipulations

One can always obtain a value of d' from a single 2×2 contingency table. This table can be either pooled over subjects or applied on a subject-by-subject basis. Unfortunately, such an analysis does not give any indication of whether the theory is appropriate for the given application. Given the many applications of the theory, this question should not be ignored. There are two standard manipulations that can be used to shed light on the matter. One is to vary the a priori probability, so that there are as many 2×2 tables as levels of the probability manipulation. Then one can construct a receiver operating characteristic, as described in the next section. Generally, the probability manipulation is a between-sessions (or between-lists) manipulation, and the subject is informed as to what the probability setting is under each condition.

This procedure necessitates collecting a lot of data. An alternative is to use a confidence-judgment procedure. Only a single a priori probability value need be used, but on each test the subject gives a confidence judgment (either incorporated into a yes-no judgment or kept separate). To do so, the subject is assumed to set up multiple criteria (one less than the number of scale values), and each judgment locates that observation within a region of the strength (decision) axis. Suppose a 6-point scale is used. The confidence-judgment matrix would then be 2

Table 1.1
Frequencies of Six Confidence Judgments to Old and New Items (Illustrative Data)

		Confidence judgments					
		---	--	-	+	++	+++
Old items	25	35	40	40	28	32	
New items	90	50	28	18	10	4	

1. RECOGNITION MEMORY

$\times 6$; the rows would indicate probe type (old or new); and the columns, the six confidence judgments. Cell entries would be frequencies. An example is given in Table 1.1.

Extracting Signal-Detection Measures from Data

Given that one has several 2×2 tables from varying the probability, or given that one has a confidence-judgment matrix, how can d' and β be obtained? One method is to construct a *receiver operating characteristic*, or ROC curve. An ROC curve is a plot of hit rate as a function of false alarm rate.

An example is shown in Figure 1.4 of an ROC curve for the data shown in Table 1.1. To obtain the hits and false alarms, compute cumulative proportions for old and new items separately. That is, start with the strictest criterion (here, ++++); the hit rate is $32/200 = .16$ and the false alarm rate is $4/200 = .02$. For the next strictest criterion (+++), the hit rate is $(32 + 28)/200 = .30$, and the false alarm rate is $(4 + 10)/200 = .07$. The five pairs of values for the six confidence judgments are shown in Table 1.2.

Given an ROC curve, how is the value of d' obtained? One method is to take the intersection of the ROC curve with the negative diagonal, as illustrated in Figure 1.4. [Determine the hit rate and the false alarm

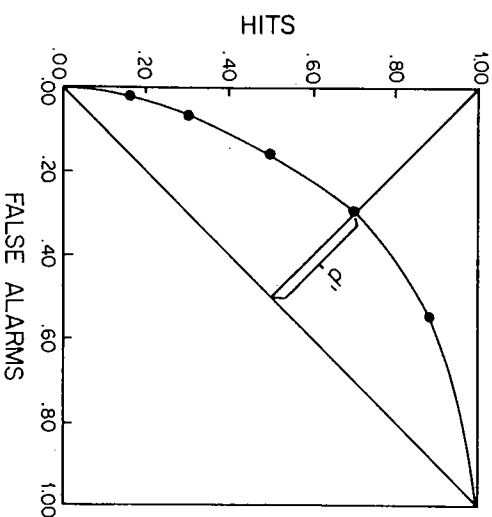


Figure 1.4. An ROC curve for the data of Table 1.1.

