VOL. 74, NO. 2 AUGUST 1970

Psychological Bulletin

SIGNAL DETECTION THEORY AND HUMAN MEMORY [1]

WILLIAM P. BANKS [2]

Pomona College

Some applications of signal detection theory (SDT) in the study of memorial processes are critically reviewed in four categories: (*a*) uses of SDT to scale memory strength, (*b*) use of SDT in criterial interpretations of data that seem to indicate forgetting, (*c*) attempts using SDT to determine the form of trace storage and to settle the question of all-or-none learning, and (*d*) extensions of SDT to scale memory-based discriminability in finer analyses of retention. The techniques that SDT offers the student of memory are explained, their limitations and past misapplications are discussed, their advantages in various situations are enumerated, and future applications are suggested.

Signal detection theory (SDT) has had a considerable influence on psychological experimentation and theory since it began to find a place in the psychological literature in the late 1950s and early 1960s. Although the relevance of SDT is most obvious in areas concerned with sensory and perceptual processes, the techniques of SDT have potentially a much wider application, and the impact of this theory is only beginning to be felt in many areas where it has much to offer. One such area of moderate impact but great potential for SDT is human learning and retention. Since Egan (1958) applied SDT successfully in measuring verbal retention, a few investigators have used techniques of SDT in a variety of analyses of human memory. These studies have presented improved, unbiased measures of retention as well as new approaches to classical problems. This paper attempts to organize the objectives and contributions of these applications of SDT, and to make clear the potential of the SDT measurement techniques where they provide an advantage over others in the study of memory. An attempt is also made to define the limits of validity of the SDT measures and to indicate where these limits have been overstepped.

Signal detection theory is a more general theory than is commonly realized. Its approach to analyzing performance allows assumptions that range from very flexible nonparametric ones to the relatively strict assumptions of the familiar *d'* statistic. The following section is devoted to an exposition of SDT and a summary of some measures of performance possible under various sets of assumptions. For a more thorough coverage of the basic theory of signal detection, the reader should consult Swets, Tanner, and Birdsall (1961), Swets (1961), or Green and Swets (1966). Specific references pertaining to each measure are cited when that measure is discussed so that the interested investigator may make use of it.

[1] This paper, a general review of issues in the application of signal detection theory to the study of memory, was received as an ordinary paper. The paper which follows it was commissioned by the editor. This paper is published out of its normal order of publication because it provides an introduction to assist the reader who is unfamiliar with the issues discussed in the commissioned paper, which follows.

[2] This paper grew out of a series of luncheon talks given by the present author at the Institute of Human Learning, University of California, Berkeley. Much of the writing was done at the Institute while the author was supported by a Public Health Service postdoctoral traineeship (5T01-GM01207-05). The comments and encouragement of Leo Postman, Geoffrey Keppel, Donald A. Riley, and others at the Institute are gratefully acknowledged.

Requests for reprints should be sent to William P. Banks, Department of Psychology, Pomona College, Claremont, California 91711.

Signal Detection Theory as Applied to Memory

In detection of a weak signal an observer might be insensitive because of a limitation in the sensory organ in question; on the other hand, he might appear insensitive because he is overcautious and reports only signals that he is sure were presented. Either factor could serve equally well to raise an observer's measured sensory threshold. There are, of course, many other ways an observer's response biases can be superimposed over his basic sensory processes, and SDT was evolved in an effort to separate the truly sensory aspects of detection from the decision aspects. The application of SDT to memory depends on conceiving of a memory trace as a signal that the subject must detect in order to perform in a retention task. Given this conception of memory performance, it is reasonable to assume that percentage correct scores may be biased indicators of retention—just as thresholds may be biased indicators of sensory performance—and, in addition, that SDT techniques should be used where possible to separate the truly retention-based aspects of memory performance from the decision aspects. It is, furthermore, reasonable to take advantage of the analytic techniques of SDT to gain access to additional questions about the process of remembering that are analogous to questions about the detection of signals.

Because the concepts of SDT fit quite naturally into the study of memory, the following discussion proceeds as though SDT were developed precisely for that purpose. Students of SDT in its sensory guise will find that "noise" and "signal plus noise" distributions have been replaced by "new item" and "old item" distributions and that "signals" have been replaced by "memory traces." It is hoped that they will understand that these and other substitutions have been made in the interest of simplicity.

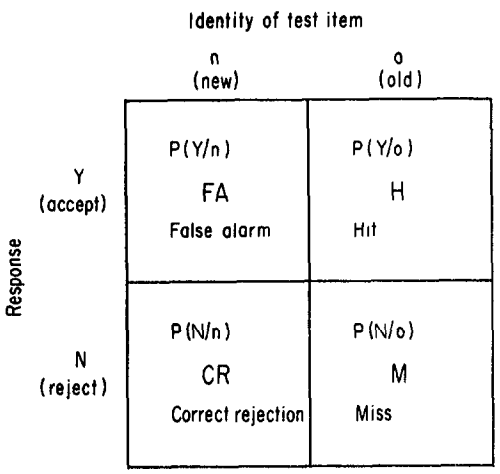


Fig. 1. The decision matrix.

The Decision Matrix

In the study of memory the experimenter must know which items in a recognition list were on the learning list (old items) and which were not (new items); or, in a recall protocol, which responses are correct and which are intrusions. The subject, in turn, must make one of at least two possible responses on each occasion: "yes" (this is an old word) or "no" (this is new). Given the subject's yes and no responses to old and new items, a decision matrix like that shown in Figure 1 can be constructed. Such a matrix is generally the basic datum for any experimental condition considered by SDT. A hit (H) occurs when the subject correctly identifies an old item, and a false alarm (FA) occurs when the subject incorrectly accepts a new item as old. When the subject incorrectly rejects an old item he generates a miss (M), and rejection of a new item is termed a correct rejection (CR). The four cells of the decision matrix can be summarized with two values. The conventional technique is to express the H rate as the proportion of old (o) items correctly identified, and the FA rate as the proportion of new items (n) incorrectly accepted as old. Thus, H rate $= P(Y/o) = H/(H + M)$ and FA rate $= P(Y/n) = FA/(FA + CR)$. These two values adequately represent the entire matrix, for $P(N/o) = 1 - P(Y/o)$ and $P(N/n) = 1 - P(Y/n)$.

It is a fact of fundamental importance to SDT that the two values, $P(Y/o)$ and $P(Y/n)$, covary when, all else being equal, motivation varies. If, for example, a subject has marked on a recognition list only those items he is sure are old, he will have a low H rate and a very low FA rate. If he is then asked to find some more old items, he must accept some items of which he is less con-

fident, and will surely increase his FA rate as well as his H rate. In SDT the form of the relationship between H and FA rates is usually studied with use of a receiver operating characteristic (ROC), which is simply a graph of all obtained (or theoretical) H and FA rates possible in a given situation. The present paper, being concerned with retention, adopts Norman and Wickelgren's (1965) suggestion that the H, FA graph be called the memory operating characteristic (MOC).

The MOC function

Figure 2 contains the graph on which MOC functions are plotted. Any decision matrix is a single point in MOC space, the H rate being plotted along the ordinate, and the FA rate along the abscissa. Let Point A represent the performance of a subject judging which words in a recognition set were on the learning list. The placement of Point A indicates that he is behaving very cautiously. Acting a little less cautiously he will increase his H rate, but will necessarily increase his FA rate as well. The decision matrix resulting from this slightly less cautious behavior is represented by Point B. Further willingness by this subject to accept doubtful words will result in performance at Points C, D, E, and F. Although the form of the MOC function shown in Figure 2 is one commonly observed in psychological data, it is by no means the only possible one, and the shape of the function is itself often an important empirical question.

The MOC is also referred to as the isomnemonic function. This term is derived from the fact that it is the locus of all points possible with a single memory strength. Any two points on a single MOC, however much they differ in percentage correct (H rate), represent equivalent retention. They differ only in the degree of caution exercised by the subject. Possession of a complete MOC therefore allows one to determine whether two decision matrices represent equivalent retention, but with a complete MOC even more can be known. The MOC divides the space in which it is plotted into two regions: Any point falling above it represents better performance, and any point falling below, worse. For illustration, consider Point G

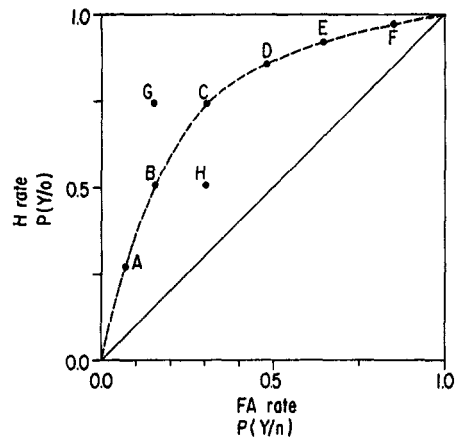


FIG. 2. The memory operating characteristic (MOC); graph of H rate as a function of FA rate. (Function shown is that consistent with the d' model. Points A–H are discussed in text and in Figures 3 and 4.)

in Figure 2. It has the same H rate as Point C, but one obtained at the expense of fewer FAs and therefore represents, by any reasonable standard, better performance. Since all the points on the MOC A-F are equivalent, Point G (and any other point in the region it occupies) is better than any point on the MOC. By a similar argument, any performance falling in the area occupied by H represents poorer retention than the A-F MOC. Therefore, as long as a sufficient number of complete, nonintersecting MOCs have been obtained, little in the way of theory is necessary to determine the relative levels of retention in comparable conditions. The statistic A_g, discussed later in this review, provides a valid basis for nonparametric tests of the differences between such MOCs.

While it might be prohibitively laborious to attempt to construct complete MOC functions by varying payoff in the manner outlined in the discussion of Figure 2, there exists, fortunately, a confidence-rating technique that can be used in many situations to generate MOC functions from subjects in a single retention test without variations in payoff. The technique requires a more finely graded response of a subject than "yes" or "no." The subject might have, for example, five categories of response, ranging from five = "Certain it was on the list," through three = "Don't know," to one = "Certain it was

(a)

	o	n
5	.26	.05
4	.24	.10
3	.20	.15
2	.15	.20
1	.15	.50

(b)

	o	n
C_A	.26	.05
C_B	.50	.15
C_C	.70	.30
C_D	.85	.50
	1.00	1.00

(c)

	C_A		C_B		C_C		C_D	
	o	n	o	n	o	n	o	n
Y	.26	.05	.50	.15	.70	.30	.85	.50
N	.74	.95	.50	.85	.30	.70	.15	.50

FIG. 3. Construction of Type II MOC from confidence ratings of a hypothetical recognition task. (In Figure 3a the cells show the proportion of old and new items rated in each category of confidence, 5 being the highest. In Figure 3b proportions are cumulated from highest confidence to lowest. Cell entries give proportion of items rated at or above a given category boundary. Figure 3c shows the four decision matrices giving H and FA rates at each criterial level. These matrices represent performance at levels C_A–C_D of Figure 4a and are plotted as Points A–D of Figure 2.)

not on the list." The n categories of response are used to set up a $n \times 2$ decision matrix, and, from this $n \times 2$ matrix, $n - 1$ points can be plotted in the MOC graph. The technique for constructing a MOC function from confidence ratings is explained and illustrated in Pollack and Decker (1958) and in Green and Swets (1966, pp. 40–43). Figure 3 shows the mechanics of this technique.

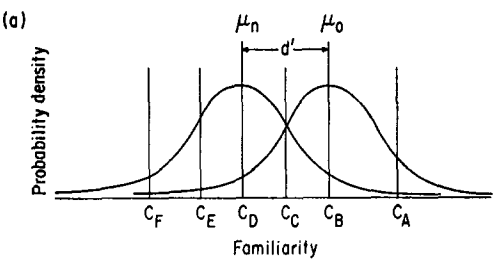
The ROC constructed from confidence ratings is termed the Type II ROC, and there is some question whether it should be treated in the same way as the Type I ROC obtained by payoff variations (cf. Clarke, Birdsall, & Tanner, 1959). In this paper it is assumed that Type I and Type II MOCs can be treated in the same way. Generally, as is seen, it is the shape of the MOC, rather than the way it was obtained, that determines how it will be treated.

THE SDT MEASURES OF STRENGTH AND CRITERION

It is necessary, for the majority of SDT analyses of memory, to go beyond the MOC representation of performance and obtain a single measure which denotes a given MOC. Such a measure is necessary if statistical tests of differences between MOCs are to be made, or if MOC data are to be summarized in a convenient form. A further, and quite important use of the single measures of MOCs is extrapolation to the entire MOC function from a single point in MOC space. Thus, for example, it would be possible to determine whether Points D and G of Figure 2 represented equivalent performance when no other data were available, as long as it was known what assumptions could be made about the MOCs passing through them.

Gaussian Measures: d' and d_s'

These measures are based on the assumption that the effects of old and new items are distributed normally along an abstract dimension known as the likelihood axis. Figure 4a illustrates the model assumed in the calculation of d', in which the standard deviations (σs) of the positive and negative distributions are equal. Points along the likelihood axis represent different criteria, or degrees of the subject's leniency or strictness when he judges a series of items. In analyses of memory performance the likelihood axis has been labeled the familiarity axis (Parks, 1966). Thus, if a subject selects Point C_A in Figure 4a as his



(b)

	n	o
Y	0.05	0.26
N	0.95	0.74

FIG. 4. The normal distributions of new and old items along the familiarity axis assumed with the d' model are shown in Figure 4a. Figure 4b is the decision matrix resulting from operation at criterion C_A in the model of Figure 4a.

criterial cutoff value, all items with an apparent familiarity greater than that value will be accepted as old, and all items with less will be rejected. Operation at Point C_A will produce the Decision Matrix *A* shown in Figure 4b and represented as Point A in Figure 2. Successively more lenient criteria, C_B–C_F, will result in greater H and FA rates, represented by Points B–F in the MOC of Figure 2. Such a MOC, plotted in probability coordinates rather than linear ones, will be a straight line parallel to the positive diagonal. The larger d' is, the further the MOC will be displaced from the diagonal.

In the model of Figure 4a, d' is the distance between the means of the two distributions, scaled in z units with the common variance as the metric. With varying degrees of accuracy, d' can be estimated from a decision matrix obtained under any criterion: It is simply the distance between the means that is consistent with the obtained matrix. Tables are available for determining d' from H and FA rates (Elliott, 1964; Freeman, 1964). Green and Swets (1966, p. 405) discussed a graphical method for obtaining d', and Ogilvie and Creelman (1968) have written a computer program that performs a least-squares solution for d' from confidence-rating data. Gourevitch and Galanter (1967) have derived a test for the difference between two d's and Marasculio (in press) has extended this test to k d's.

The model used in the calculation of d_s is the same as the d' model except that the standard deviations of the old and new distributions are not assumed equal. Successive criteria taken along the likelihood axis of this model will generate a MOC which, unlike Figure 2, is assymetrical about the negative diagonal. The d_s MOC, while straight in probability coordinates, is not parallel to the positive diagonal, and its slope is equal to σ_n/σ_o, the ratio of the two standard deviations. The index d_s is the z distance between the two distributions, and it can be measured with the standard deviation of either distribution, or an average of the two, as the unit. Egan, Greenberg, and Schulman (1961) presented a graphical method of measuring d_s in terms of σ_n: The point at which the MOC crosses the negative diagonal is determined, and twice the z score of the FA value at this point is equal to d_s. Green and Swets (1966, pp. 96–98) discussed other detectability measures that can be used with the d_s model. Significance tests for measures based on this model are not available, but the tests of Gourevitch and Galanter and of Marasculio are sufficiently robust to be used within reasonable deviations of the ratio σ_n/σ_o from unity.

Area under the MOC

A valuable nonparametric measure of performance is the area under the MOC, A_g. Green's (1964) demonstration that area under the ROC curve is equal to the percentage of correct responses in a two-alternative–forced-choice task established A_g as a meaningful measure of performance. Since this demonstration did not rely on any particular model of the distributions generating the ROC, A_g can be used in a wide variety of situations and by experimenters who do not wish to accept any of the stronger assumptions of SDT. A further advantage of A_g is that it can be used to compare performance across conditions for which different distributional assumptions may be necessary. The relevance of A_g to studies of memory was established—for at least one situation—by Green and Moses (1966) who demonstrated empirically that forced-choice recognition performance is predicted by the area under the MOC derived from confidence ratings.

Pollack and Hsieh (1969) made an exhaustive study of A_g, using Monte Carlo methods to determine its sampling variability under a wide variety of assumptions. They also discussed the relationship between A_g and a number of other indexes of sensitivity, and illustrated the algorithm used in their computer program to derive A_g from confidence ratings.

The MOC obtained with the confidence-rating method can be used to achieve a great sensitivity of measurement. With "yes-no" recognition testing the subject is often unable to convey all the information he can about his memory. Some recognition items are definitely identifiable as old or new, but many occupy a middle ground and the subject is forced to collapse whatever distinctions he can make into two categories. Although it has

been recognized that confidence rating allows the subject to express gradations of memory strength, there has been no general agreement on single-valued indexes of rating data that express the available information validity. The A_g, and for that matter, d' and d_s, being based on the MOC rather than on the confidence ratings themselves, can do full justice to the information contained in the rating data. With any of these measures there is freedom from the bias effects that can exist in confidence ratings, and with A_g there is no need to make restrictive parametric assumptions.

Measures of Criterion: β and C_j

The classical measure of criterion (Swets et al., 1961) is β. It is the value on the likelihood axis at which the criterion resulting in a given decision matrix is set and is equal to the ratio of the densities of the two distributions at that point. Green and Swets (1966, pp. 36–40) have shown that the slope of the MOC at any point (in linear coordinates) is numerically equal to the β that resulted in operation at that point, thus an approximate graphical determination of βs is possible given smooth MOCs. Freeman's (1964) tables give βs for a wide range of H and FA rates, but they are based on the equal variance model of the d' statistic. Other models would result in different values of β at the same H and FA rates.

Although β is a mathematically elegant criterial measure, and one of considerable utility in the calculation of optimal decision strategies (Swets et al., 1961, p. 308), it has a number of disadvantages as a psychological index of criterion setting. The range of values β can take is a function of the detectability index. For example, if $d' = 0.0$, β can have only one value, unity; but when $d' = 1.0$ its measurable range extends from near 0.0 to approximately 10.0. And with a d' of 3.0 the upper measurable limit might be near 100. Another difficulty with β is that in the d_s model, it is not monotonic with the likelihood (or familiarity) axis (Green & Swets, 1966, pp. 62–64; Lee, 1969). Thus paradoxical situations can arise in which rather different familiarity values can have the same β value. The measure suggested here as an alternative to β is referred to as C_j. It is the distance along the likelihood axis from the mean of the new distribution to the point of the criterion setting, scaled in z units with σ_n as the unit. The subscript j indicates which criterion of the $N - 1$ possible with N rating categories is being referred to. The range of C_j is not a function of detectability, C_j is always monotonic with the likelihood axis, and it is susceptible of statistical test by a straightforward extension of the Marascuilo or the Gourevitch and Galanter techniques. The C_j for any point on the MOC can be determined graphically from its z score on the FA axis. The computer program of Ogilvie and Creelman (1968) yields $N - 1$ C_js for every N categories of confidence in the rating data put into it. It should be mentioned that C_j is not without disadvantages. It can be used with full validity only to compare points lying on the same MOC, and then only when the MOC conforms to a Gaussian model. While most researchers will find C_j a useful measure in spite of its limitations, the need for a perfectly general psychological index of criterion still exists.

Applications of SDT in Studies of Memory

The SDT studies of memory covered in this paper are discussed under four headings. The first of these, Strength Measures, includes studies where SDT has contributed highly sensitive measures or measures uncontaminated by response bias as an improvement over previous techniques. In this category papers that treat SDT measures as parameters of a learning theory are treated separately from those which use them simply to improve on traditional measures. Studies under the second heading, Criterion versus Strength, use SDT to determine whether experimentally induced differences in measured retention are due to changes in trace strength or response bias. The third topic, Underlying Distributions, covers some studies that analyze empirical MOC functions to determine what models of memory processes are possible. This heading includes some recent assaults on the question of incremental versus discrete learning. In the fourth category, Scaling Ap-

plications, fall some studies that make explicit use of d' or d_s as scale values rather than as measures of a single dimension of memory strength.

Strength Measures: Strength as a Theoretical Construct

Wickelgren and Norman (1966) have developed a formal model for recognition-memory experiments in which d', rather than conventional measures of recognition, is the fundamental statistic for experimental tests. Their model assumes the underlying distribution of Figure 4a together with the statistical decision theory of SDT. No specific predictions are made about the effect of criterion setting on conventional measures in any particular situation. It is assumed, rather, that performance is always a joint function of memory strength and a decision process, and that measures of strength alone are the only valid measures for a theory of memory.

This model has been developed and tested in a number of experimental studies. In their 1966 article Wickelgren and Norman presented some data that allow a decision among a number of assumptions possible within their model. They found that the strength of an item decays exponentially in short-term memory (STM) as a function of the number of intervening items, or:

$$d(i) = a\phi^i,\ a \geqslant 0,\ 0 \leqslant \phi \leqslant 1, \qquad [1]$$

where $d(i)$ is the strength (d') of the item at test, a is its acquisition (initial strength) parameter, ϕ is the decay parameter, and i is the number of items intervening between presentation and test. The exponential decay law is quite pervasive in their data, and appears as a fundamental equation in terms of which other findings are expressed. In the same study, Wickelgren and Norman found that the primacy effect—earlier serial positions (SPs) better retained than middle SPs—is best explained with the acquisition parameter:

$$d(i) = a(i)\phi^i, \qquad [2]$$

where $a(i)$, the initial strength parameter, is no longer a constant as in Equation 1 but is a function of i. Thus, earlier SPs are retained better because they begin their exponential decay from a higher initial strength.

Norman (1966) investigated the effects of input modality (auditory versus visual), rate of presentation, type of item, and type of test (recall versus recognition) on the parameters of memory strength in STM. These manipulations were found to affect the acquisition parameter (a) almost exclusively; rate of decay was quite constant over all conditions. One unexpected finding, which casts some doubt on the generality of Norman's strength measures, was that the initial strength of items (a) was much greater when estimated from recognition than when estimated from recall. Since learning conditions were equivalent, the difference found between recognition and recall cannot be due to differences in the absolute strengths of the items being tested. The d' measure is not, however, the absolute memory strength of the item. Its value depends both on memory strength and on the similarity between the new and old items in the test list. As the similarity between targets and distractors increases, confusion between them will increase; and d' for detection of targets, given a constant memory trace, will decrease. Distractors must, therefore, be equivalent in all conditions to be compared. In Norman's (1966) experiment it is likely that the distractors in recall (provided as intrusions by the subject) and in recognition (provided by the experimenter) were not equivalent. And, in fact, Norman (1966, p. 380) cited some evidence for a lack of equivalence. The fact that a and not ϕ differed for recognition and recall suggests that there was only a zero-point difference in the two measurement procedures, but Bernbach's (1967) treatment of the effects of test-list construction (as discussed later in this review) would demand a multiplicative relationship. Comparisons of recall and recognition should wait, it seems, for the problem of the distractor set to be worked out.

Wickelgren (1967) showed that Equation 1, when applied to the strength of association in STM between item n and $n+1$ in a serial list, still holds when n is followed by a different item elsewhere on the list. Thus the strength of interitem (forward) associations is independent of the strength of competing as-

sociations. This finding provides further support for the generality of the exponential decay law and suggests that at least in this situation, the law is not simply a result of retroactive or proactive inhibition (RI or PI, respectively), as such inhibition would have been increased by specific competing associations. The result also suggests an important divergence between strength measures and conventional probability measures of retention. Wickelgren pointed out that cued recall for $n + 1$, given n as cue, would have been depressed when n was also followed by a different item on the list. Such a result should not be taken as evidence for a weakening of the $n, n + 1$ association by the conflicting one, if the SDT results are accepted, but rather as evidence for response competition in recall. Wickelgren did not present any data on criterial levels or percentage correct for the different conditions of his experiment, however, and it is therefore not possible to determine whether other measures of retention would have been biased.

Norman and Waugh (1968) used the Wickelgren and Norman strength model to assess the relative effects of subsequent items presented for learning, and for recognition on the strengths of items previously learned. They found that items to be stored and items presented for test both produce an exponential decline in the strength of items already in STM. In one of their two experiments the two effects were equivalent, but, in the other, test items caused a less rapid decline than did items to be learned. Their results depended critically on the use of memory-strength measures because the regularities they found in the strength data are not apparent in the raw probabilities of correct response, although the general direction of effects is the same. Furthermore, their analysis required a separation between performance mediated by STM and LTM (long-term memory), and the strength model provided a technique for this separation that seems more valid and more suited for parameter estimation than previous techniques used by these authors (Waugh & Norman, 1965).

Norman and Wickelgren (1965) tested the strength in STM of digits and pairs of digits with the intention of predicting the strengths of pairs from the strengths of the two single digits composing the pair. Such prediction was not possible. They found that the distribution on the familiarity axis of old and new items was somewhat different for pairs of digits than it was for single digits, and that a different model of acquisition of strength was required for the two kinds of item. These models will be covered in a subsequent section (see Underlying Distributions).

Parks (1966) has presented a model of recognition-memory performance which is based on SDT and is similar to Wickelgren and Norman's (1966). This model is intended to deal with the problem that after a given amount of learning, performance on a recognition task declines as the number of distractors increases, even when correction-for-guessing is applied. The model assumes the distribution of old and new items of Figure 4a, and memory strength is the distance between the means of the two distributions. The model also employs a decision process which Parks, unlike Wickelgren and Norman, formalized so that H and FA rates can be predicted. The decision formula simply assumes that the proportion of items in a recognition set accepted as old will equal some proportion (K) of the test items which are actually old. In assessing a number of experiments, including his own, which vary the composition of test lists, Parks estimated K and d' from the results with one test list, and then used these parameters to predict the results of the others. His success in the majority of these predictions is unquestionable and surprising as well, considering the wide variety of experiments and materials he considered. Besides being a verification of the SDT approach, Parks' analysis indicates that the effect of recognition-list composition on performance is consistent with a single strength of memory trace, and no complex hypotheses of learning or interference during testing are necessary to account for the effect.

Strength Measures: Descriptive Indexes

While Wickelgren and Norman and others, discussed earlier in this review, are concerned as much with testing and developing their strength models of memory as they are with

solving the problems to which they apply them, there are a number of studies that have employed strength indexes simply to overcome some measurement problem. These studies are less concerned with the conceptual apparatus of SDT than with its serviceability: SDT provides some very sophisticated correction-for-guessing techniques.

Hopkins and Schulz (1969) studied the effects of stimulus (S_1) and response (S_2) meaningfulness on paired-associate (PA) recognition learning. They wished to determine whether variation in S_1 meaningfulness still has a greater effect on learning than S_2 variations when the need for response learning is eliminated. To obtain a measure free of response biases they tested with a confidence-rating technique and calculated A_g for each combination of S_1 and S_2 meaningfulness. The A_g data agreed with the rest of their results in showing that S_1 meaningfulness variations had more effect on retention than S_2 variations. Some uncontaminated strength measure was necessary in this experiment because a response bias to say yes to recognition items was observed, and this bias varied as a function of meaningfulness. The A_g measure was chosen for its virtual lack of parametric assumptions.

Other uses of SDT techniques strictly for the sake of measurement include Allen and Garton's (1968) and Schulman's (1967) studies of the word-frequency effect in recognition memory, as well as Slamecka's (1969) replication (with a recognition test and MOC analysis substituted for recall) of Marks and Miller's (1964) study of semantic and grammatical effects on memory. Kintsch and Carlson's (1967) examination of changes in the MOC during PA recognition learning also falls into this category of atheoretical use of SDT, although they were working in the context of a two-stage theory of learning which is heavily indebted to SDT (Kintsch, 1967). Wickelgren (1966), in a study unrelated to his and Norman's strength theory, used a MOC analysis of recognition performance to demonstrate the effects of phonemic similarity on RI. Murdock (e.g., 1968) has been a consistent and frequent user of d' as a measure of retention but has remained among those who use SDT indexes more as descriptive statistics than as parameters of a learning theory.

Criterion versus Strength

Murdock's (1966) study of the SP effect asked whether this effect results from a variation in the strengths of items as a function of SP, or whether it is a result of variation in criterion settings. He cited the difference between a priori and a posteriori recall curves as evidence for a criterial explanation of the effect. What these curves show is that items from later SPs are recalled well, but are also likely to appear as intrusions at many other positions, while items from the earlier SPs are not recalled as well, but neither are they intruded so loosely. It would appear, then, that some of the difference between H rates of early and late SPs might be accounted for if the FA (intrusion) rates were considered.

To test the criterial hypothesis of the SP effect, Murdock presented a 5-item list and then asked for immediate recall, requiring subjects to respond with something for each SP. They rated their confidence in the correctness of each response and from these ratings d' and β were computed. For all SPs, d' was virtually constant, but β (for the highest confidence category) varied in excellent accord with the SP effect. It is, however, difficult to see how criterial differences could produce the SP effect (or any effect) in this situation. Recall was forced, and criterial differences could not, therefore, have had a chance to operate: all subjects at all SPs must operate at a very low criterion. A possible solution to this problem is that extraneous items have different intrusion probabilities at different SPs, thus placing a different lower limit on β at each SP, but this is an explanation in terms of response competition, not criterion shifts.

It is more likely that Murdock's results are simply an artifact of his measuring procedure. The decision matrices for each SP can be reconstructed by using a priori percentage recall (Murdock, 1966, Table 1) to determine H rates. With Elliott's (1964) tables FA rates can be determined from H rate and d'. These constructed matrices, expressed in terms of frequencies rather than proportions, show roughly constant FA and M frequencies over

all SPs, H frequencies that reflect percentage recall, and CR frequencies that increase by whatever amount H may decrease. It is as though subjects recall what they can and then, because of the forced-responding requirement, guess and place these guesses in the lower categories of confidence. The virtual complementarity between CRs and Hs means that as recall declines, H rates and FA rates will both decline together. (The H rates will decline with lower recall because the numerator of the H proportion will decrease. The FA rates will also decline with lower recall because the denominator of the FA proportion will *increase*.) The variation in recall over SP is thus translated into a covariation between H and FA rates, and such covariation will indicate greater changes in β than d' as a function of SP. In fact, with certain values of m and n such that CR + H = m, and M + FA = n, artificial decision matrices will show only an increase in the strictness of β and no change in d' as H frequencies are decreased and CRs increased. Murdock seems to have had the misfortune of obtaining such values of m and n.

The strong possibility that Murdock's results are artifactual vindicates Bower's reservations about them. Bower (1967, p. 286), in a mathematical derivation from his multicomponent theory of memory, showed that d' should vary directly with amount recalled and found Murdock's results puzzling. Bernbach's (1967) argument for a two-state theory of memory, on the other hand, suffers if Murdock's results cannot be believed, for the argument was based, in part, on the invariance of d' with level of recall.

The possibility of an artifact in Murdock's results should serve as a caution to those who would base SDT measures on forced recall. Nonforced recall can, however, have artifacts just as serious, for without explicit instructions about the number of responses to be given at recall, the CR frequency is a matter of the subject's discretion. At the very least, conclusions about strength or criterion drawn from recall should be supported by other experimental evidence.

The present author (Banks, 1969) studied interference in the A-B, A-C PA paradigm to test the extinction theory of retroactive inhibition (RI). According to the extinction theory, forgetting of responses from a first list (A-B) after learning a second list (A-C) is caused by a weakening of the A-B associations. This theory was initially compared to a hypothesis of reduced confidence, according to which the first list is not forgotten but is simply more difficult to guess about after second-list learning. In the experiment designed to allow a decision between these two possibilities, d' and β for the first list were measured after 2 and 20 trials of second-list learning. The use of a modified modified free recall (MMFR; Barnes & Underwood, 1959) augmented by confidence ratings allowed this measurement. A decrease in d' as interference increased would support the extinction hypothesis, while increase in the value of β would support the restricted guessing notion. The results showed no change whatever in d' and some increase in β as a function of interference. For a number of reasons, however, the restricted guessing hypothesis was not accepted. First, a payoff variable in the experiment had no effect on recall, and it should have if criterion setting was depressing recall. Furthermore, in a second experiment in the study, first-list recall was forced, but it increased only 30% when so forced. This is not enough of an increase to indicate a very large criterial effect. It was concluded that RI was best explained in terms of generalized response-set competition between the two lists and that the criterial effects observed were a byproduct of this competition.

Donaldson and Murdock (1968) used SDT techniques in an analysis of performance in the continuous-recognition task of Shepard and Teghtsoonian (1961). The SDT analysis was used to determine whether criterial shifts or a buildup in proactive inhibition (PI) is responsible for the increase in false-positive rate that is characteristic of the subject's progress in later trials of the task. The results showed that response criteria (the measure here termed C_j was used) became more lax over successive blocks of trials while d' stayed relatively constant. The approximate constancy of d' was accepted as evidence that PI did not build up beyond an initial plateau and that a "steady state" of memory ca-

pacity was obtained early in the task. Criterial shifts were held responsible for the progressive increase in FA rates.

An unexpected offshoot of the SDT analysis was the finding that MOC functions for successive blocks in the experiment had progressively steeper slopes, indicating a steady increase in the ratio σ_n/σ_o. Donaldson and Murdock speculated that "diffusion" of item traces from earlier trials produces a progressive increase in the variability of new distribution. It is also possible that learning-to-learn processes or the evolution of coding schemes could reduce the variability of the old distributions of later trials. Whatever the reason for the change in MOC slope, the very fact that it occurred raises serious doubts about the comparability of C_j and d' across the experiment. One would, at least, like to see the constancy of d' supported by a nonparametric measure such as A_g or—a better approach—to have the constancy verified in an experiment using a criterion-independent response task such as forced-choice responding.

Underlying Distributions

This section covers some recent attempts to resolve the question of incremental versus discrete learning, as well as some more quantitative attempts to establish a model for the threshold characteristics of memory traces. The studies covered here generally begin with an empirical MOC function and work backwards to the underlying distribution of old and new items that must have given rise to it. These attempts are parallel to those of psychophysical theorists in answering questions about sensory thresholds on the basis of ROC data, and many fruits of the controversy over this approach in psychophysics can be applied to similar attempts in the study of memory.

The logic of reasoning from an MOC to the underlying distribution is straightforward, although the mathematics is not always so, and some of the pitfalls in the approach are only now becoming apparent. It is the case that with a given underlying distribution, H and FA rates obtained at various criterial levels will trace out a MOC of specifiable shape. The model of Figure 4a will, for example, always result in a MOC of the form shown in Figure 2. Figure 5 illustrates MOCs expected under two threshold models. To determine the underlying distribution requires a converse operation: The MOC is derived from an experiment, and the investigator must discover a distributional model that will account for it. Thus, when an MOC like that of Figure 2 is found, it has been concluded that the model of Figure 4a was the appropriate one.

It is conceded that in most cases the MOC will not have a unique solution in terms of a single underlying model. This lack of unique solutions is not a great handicap. Theorists

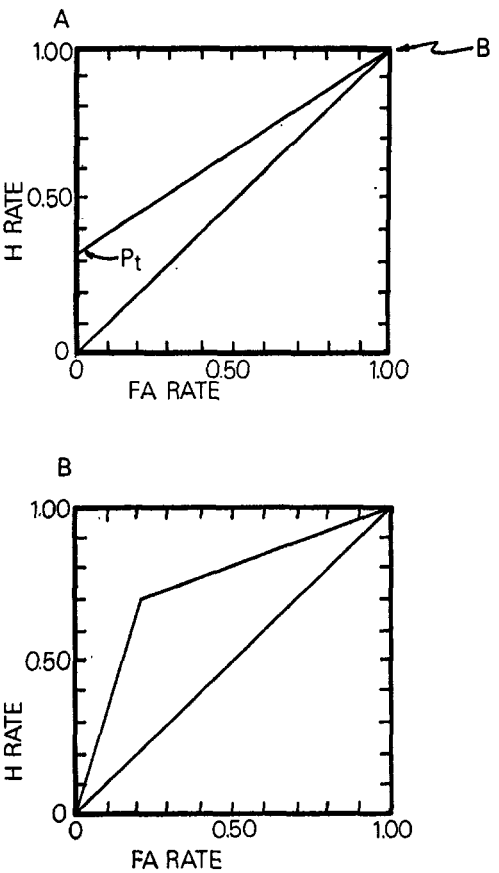


FIG. 5. Figure 5a is the theoretical MOC based on a high-threshold theory (after Swets, 1961, Fig. 4). (This is the MOC assumed by the traditional correction-for-guessing formula, where any performance on the line P_TB is corrected to the "true" probability of being correct, P_T.) Figure 4b is the theoretical MOC based on a low-threshold theory (after Swets, 1961, Fig. 5).

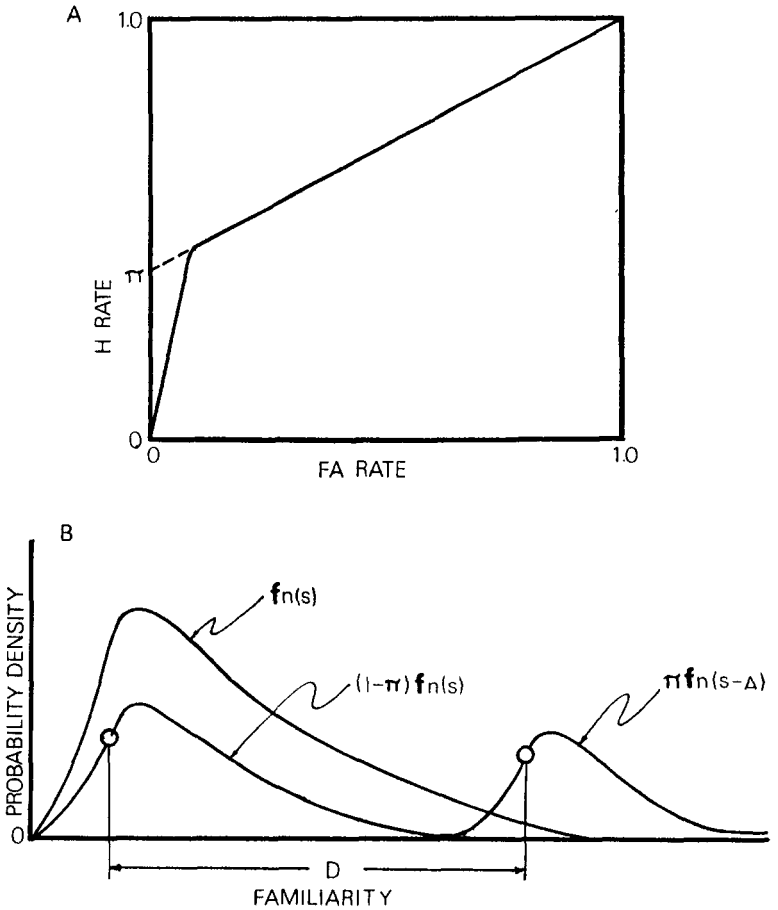


FIG. 6. Figure 6a is the idealized MOC function describing Norman and Wickelgren's (1965) pair-recognition data. Figure 6b represents the theoretical distributions of trace strength in Norman and Wickelgren's (1965) model. (The new distribution is $fn(s)$. Old items receive on learning trials an increment of D in strength with probability π, and join the π $fn(s\text{-}D)$ distribution. Otherwise, with probability $(1\text{-}\pi)$ they are not incremented and form the $(1\text{-}\pi)$ $fn(s)$ distribution.)

are generally willing to accept the simplest or most plausible model, and the all-or-none learning question has, in the past, hinged simply on the smooth curvilinearity of the MOC. Other more serious theoretical problems do, however, plague attempts to infer backwards from the MOC. When the MOC is derived from confidence ratings a smooth, nonlinear function can result from the way the subjects use the rating categories. Thus, with only two memorial states the subject may spread his judgments out on the basis of irrelevant factors or because of criterial fluctuations, and give the appearance of operating with traces distributed over a continuum of strengths. Krantz (1969) has shown that a two-state–low-threshold model of sensory detection will, except at the limits of certain response parameters, result in a curvilinear ROC function and appear to conform to a continuous model of detection. While Krantz's demonstration (anticipated in part by Larkin, 1965 and Wickelgren, 1966) vitiated much of the analysis of MOC functions, he suggested a possible method for rejecting threshold interpretations of detection data which might be applied with profit to the all-or-none learning question. Bernbach's (1967) success in

showing curvilinear MOCs consistent with a finite state model of memorial representation also renders the direct argument from curvilinear MOCs to a continuous model of memory invalid. Bower's (1967) multicomponent theory can also derive a smooth MOC from a discrete memory model. With these cautions in mind, some studies of MOC shapes that have taken a stand for a continuum of memorial strengths are considered before turning to a discussion of Bernbach's theory.

J. P. Egan (1958), in his pioneering application of SDT to memory, was concerned with showing that the conceptual apparatus of SDT was appropriate to measuring verbal retention. After administering different numbers of learning trials on a list to different groups, he obtained MOC functions by the confidence-rating method. These MOCs were of the form shown in Figure 2 and were displaced further away from the positive diagonal as practice increased. This result indicated to Egan that a statistic such as d' should be used as an index of recognition performance: Not only did d' increase with practice, but d' or a similar measure is necessary when MOCs are of that shape. If the MOCs had been rectilinear like those of Figure 5, a simple percentage score, corrected for guessing, would have been a sufficient measure of retention. It should be said that the main force of Egan's conclusion need not bear on the continuous versus discrete controversy. He simply showed that SDT techniques are vastly superior to correction-for-guessing for measuring recognition-memory performance. Conclusions about the nature of storage are not necessary.

Murdock's first (1965) use of SDT in short-term memory (STM) was a test of a high-threshold concept of memory performance according to which associative strength, while continuous, must reach a certain level before correct responding can occur. Murdock presented lists of A-B pairs, one pair at a time, in brief, single exposures and immediately tested either correct A-B pairs or incorrect A-X pairs with new second members. The subjects confidence-rated their yes-no judgments about the correctness of the pairs and MOC functions were constructed. The MOCs were smoothly curvilinear, causing

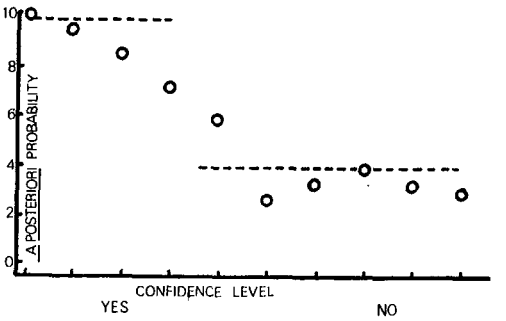


FIG. 7. A posteriori probability function for ordered pairs from Norman and Wickelgren (1965, Fig. 2), with prediction of two-valued strength model shown as broken lines.

Murdock to reject the high threshold hypothesis.

Although Norman and Wickelgren (1965) were unable to predict digit-pair performance from single-pair performance on the same lists (see p. 88), they did make a number of hypotheses about the threshold characteristics of the items. The MOC for single items presented little problem: It was the symmetric curve (Figure 2) consistent with the d' model. The MOCs for pairs were more difficult to explain. Their form, idealized, is shown in Figure 6a. While these functions are closely approximated by the intersecting straight lines of the two-valued strength model (Figure 5b), the a posteriori probabilities of correct responses in each category of confidence mitigate against this model. Figure 7 shows the empirical probabilities for ordered pairs along with the predictions (dotted lines) of the two-valued model. This graph indicates that a threshold exists, below which there are no gradations of response strength, but that there is not a single state of strength above the threshold. (It should be noted that a two-valued a posteriori function is no more dictated by a two-state theory than is a rectilinear MOC. Bower [1967, p. 279] showed how Norman and Wickelgren's a posteriori function could have resulted from the use of too many rating categories.) A model of the underlying distribution more complicated than the two-valued model seemed necessary, and Figure 6b illustrates the one suggested by Norman and Wickelgren. According to this model, presentation of an item causes it, with

probability π, to be incremented in strength by a constant amount D. When π is less than unity, the bimodal distribution of Figure 6b results. An advantage of this model is that it will account for the single-digit MOC when $\pi = 1.0$.

Bernbach's (1967) theory of memory performance couples a two-state representation of the memory trace with the decision process of SDT. He used the fact that a recognition-memory test is a kind of matching task in developing the likelihood functions necessary for the decision process. According to Bernbach, a subject in a recognition test must compare items presented with the items stored in memory, and the decision to respond "old" or "new" will be based on the minimum difference in apparent oldness between the test item and any stored item. The similarity of test-list distractors to old items and imperfection or noise in the matching function generate likelihood distributions for new and old items on the apparent oldness scale. It is to these distributions that the hypothetical decision process is applied.

In Bernbach's model it is assumed that an attended-to-item on the learning list enters the learned state (R) at least temporarily. His memory parameter is therefore δ, which is the probability of the item's returning from state R to the nonlearned state N, before the time of test. Since recognition performance depends both on memory and on the construction of the test list, a parameter besides δ is necessary to relate memory to performance. Bernbach employed d^*, which is the ideal distance between the means of the old and new distributions on the apparent oldness scale. This parameter represents the asymptotic value of d' as forgetting approaches zero. It is subject to variation only in list construction, not in memory. Bernbach then derived the following relationship between the d' index and his two underlying parameters:

$$d' = (1 - \delta)d^*. \qquad [3]$$

According to Equation 3, d' is only an indirect measure of retention. It can, nevertheless, be reliable as long as d^* is held constant by the use of the same lists in all conditions. Certain questions that involve changes in list construction across conditions, such as the effect of meaningfulness on retention, cannot be validly studied with d' measures alone.

Bernbach showed his theory consistent with a wide variety of results in recognition memory but felt that the crucial test was in recall since other theories can account for recognition data. It is unfortunate that the recall finding Bernbach chose to explain was the invariance of d' with *SP* (Murdock, 1966). This finding is very likely artifactual (see section entitled Criterion versus Strength), and one might be suspicious of a theory that would predict it. The foray into recall does not, however, detract from the important contributions of Bernbach's paper: (*a*) the demonstration that a finite state model is adequate for recognition memory, and (*b*) the explicit mathematical treatment of the effect of list composition on recognition.

Scaling Applications

In these studies d' is used to measure the retention of something other than simply the target items of a list. Reviewed here are three scaling applications: one measuring retention of list tags, one measuring retention of serial positions, and one measuring the confusing effect of synonym fillers in recognition of categorized material. Traditional measures of retention are, in most cases, too insensitive or too subject to criterial effects (or, simply, impossible) to apply to these and similar problems. The present applications are cited as examples of what can be done with SDT techniques to open up some questions to inquiry.

Winograd (1968), in his thorough examination of list differentiation (LD), performed a number of experiments using two lists of 25 common words. He varied both the relative and the absolute number of trials on the lists, then presented subjects with all 50 words and required them to indicate list membership with a rating scale. He subjected the ratings to analysis by conventional measures (mean confidence rating and a posteriori probability) as well as by the d_s index. The conventional and the SDT analyses were in substantial agreement. The main advantage of d_s in this study was that it provided a bias-free check

on the other statistics. It also provided a convenient single measure of effects otherwise stated less elegantly and suggested a conceptual model of the LD findings. According to this model the experimental parameters are conceived of as moving the two lists about on a hypothetical decision axis.

In his analysis of serial effects in STM, Murdock (1968) made ingenious use of SDT to overcome some difficult measurement problems. With the data of his Experiments I, II, and III (pp. 5–8) he reexamined the criterial explanation of the SP effect (cf. Murdock, 1966). After presenting a 5-, 8-, or 11-word list, he gave positional cues for recall. For the SDT analysis the responses from each list were cast into an $N \times N$ intrusion matrix where rows (i) are input (list position of word) and columns (j) are output (position of word in response protocol). The number in cell ij represents the total number of times the word presented in the ith position was given as a response for position j. Thus, the diagonal cells of the matrix represent H frequencies, and all off-diagonal cells in a given column are FAs. In the SDT analysis, H rate for the jth position equals the proportion of the ith row total (where $i = j$) falling into the diagonal cell, and FA rate is the proportion of the remaining rows falling in the jth column. Figure 8 illustrates Murdock's method for obtaining H and FA rates from the $N \times N$ matrix. These two proportions for each SP of the lists were entered into Freeman's (1964) tables and translated into d' and β values. The results showed that criterial variations definitely contribute to the SP effect: β is very high at the beginning of the lists and drops off rapidly toward the end. Contrary to Murdock's previous (1966) conclusion, however, d' also varies with SP.

Murdock performed a second novel analysis in Experiments IV and V of his 1968 study. In these experiments, subjects saw lists of common words, and after a single presentation of each list they were shown one word from it. Their task was to respond with the serial position of that word. For each length of list, a $N \times N$ matrix was set up as before, with rows and columns designating input and output positions, respectively. The SDT analysis was designed to measure the confusability of adjacent positions, and from each $N \times N$ matrix, $N - 1$ different 2×2 confusion matrices were taken, one matrix for each two SPs analyzed. From these matrices the distance in discriminability units (a Thurstonian distance—cf. Lee, 1969) between the two SPs was determined. For example, the 2×2 matrix summarizing the data from the first and second positions of a list consisted of the following cells of the $N \times N$ matrix: a_{11}, a_{12}, a_{21}, a_{22}. In these 2×2 matrices the H rate was defined as $a_{11}/(a_{11} + a_{12})$ and the FA rate as $a_{21}/(a_{21} + a_{22})$. The H and FA rates were defined similarly for each of the remaining comparisons of adjacent positions. From these H and FA rates, d' was determined for each pairwise comparison. The results for each length of list showed d' for the pairs to obey a U-shaped SP function. The first and last pairs were discriminated better than the intermediate ones.

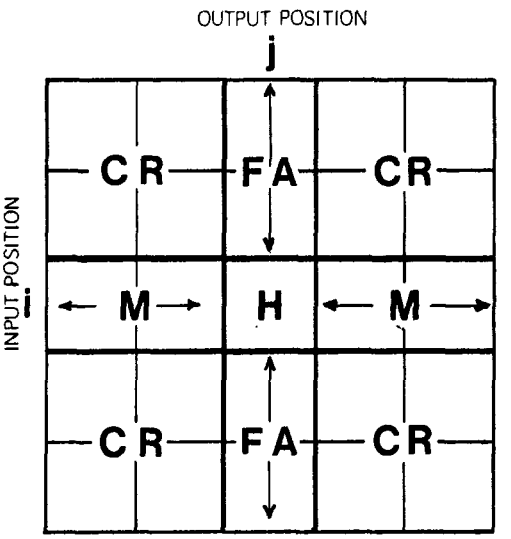


FIG. 8. Illustration of Murdock's method (1968, pp. 5–6) for determining d' of a SP from the $N \times N$ input-output matrix (see text). All items in column j were given as a response for SP j, thus those in cell ij are Hs (here $i = j$). The Ms are items presented in SP i but given as a response for a different SP. The FAs are items presented at positions other than i but responded for that position, that is, appearing in output position j. The CRs are all responses neither presented in, nor responded for, position i. The total H, FA, M, and CR frequencies are cast into a decision matrix (Fig. 1), from which FA and H rates are computed.

Both of the techniques that Murdock employed for determining d' are of special interest because they illustrate methods for analyzing recall without confidence rating. All other SDT recall-analysis techniques are based on the Type II confidence-rating MOC, and extensions of Murdock's procedures may provide fruitful Type I analyses of recall and target discrimination in other situations. It should be determined, however, whether d' is the proper discriminability index for these analyses. Such a determination could be made, for example, by varying payoff in experiments parallel to one of Murdock's conditions, and taking H and FA rates by his methods. The resulting Type I MOC functions would indicate whether d' or some other measure was the appropriate one.

A number of flaws in Murdock's analyses should be pointed out. First, while he measured overall strength or discriminability in his analysis of Experiments I–III, he determined only the adjacent-pair discriminability for Experiments IV and V. The data from the two sets of experiments are not, therefore, comparable, although Murdock does compare them, and should not be lumped together as supporting the same conclusions. Generous as it was for Murdock to present two, rather than just one innovative SDT analysis, he would have done better to have based his comparisons across experiments validly on the same analysis. Another dubious point is Murdock's discovery of the constancy, or near constancy, of the sum of d' across SPs for the longer lists in the five experiments. Even if the d's had the same meaning in all experiments, such a constancy could be meaningful only if it were known that d' measured degree of retention with the exactitude of a ratio scale. This constancy of summed d's must be viewed as accidental until d' is shown to be at least a ratio quantity, and it is hoped that future investigators will await that time before they inquire into such constancies.

Mandler, Pearlstone, and Koopmans (1969) studied the effects of categorization on recognition performance with a number of different techniques. Of interest here is their use of SDT to measure the effect of synonym distractors on recognition. When distractors were similar to or synonymous with the targets, H rates declined but FA rates were relatively unaffected. They interpreted the constancy in FA rates to mean that the confusing effects of semantic similarity are small. Actually, the d' differences showed a sizable effect. The lack of variation in the FA rates indicates a change toward strictness in the criterion as targets became more difficult to tell from the distractors. The utility of d' is precisely that it is not affected by such criterial variations and can be used in situations where criteria may vary.

One can imagine a great number of questions about memory approached with SDT in the manner of Mandler et al. (1969)—that is, by using d' to assess variations in d^* (Bernbach, 1967) rather than retention, per se. For example, the relative use of phonetic, orthographic, or semantic features in the retention of words can be assessed by requiring their recognition in lists containing distractors similar to the targets in one of these characteristics.

Conclusions and Discussion

In this paper a number of points about uses of SDT in the analysis of retention have been made in the context of particular studies discussed. Those points of general interest or of significant practical consequence for experimental use are now summarized, with additional remarks in some cases.

1. At least two sorts of benefit can be realized by using a SDT index of recognition performance: (*a*) SDT recognition measures of retention are unbiased by response criteria, implicit or uncontrolled payoffs, size of distractor set, and any other factor unrelated to retention itself that may influence percentage correct measures. Investigators who do not wish to assume the theoretical models of SDT would still profit by using some SDT index in place of more primitive correction-for-guessing methods. Use of a SDT technique does not commit one to any strong parametric versions of the theory. For the most theory shy, the statistic A_g is recommended, or—even more atheoretical—percentage correct in a fixed-alternative–forced-choice situation. It should be remembered that traditional cor-

rection-for-guessing techniques are based on a threshold theory, but one that is more restrictive in its assumptions and less in accord with psychological data than SDT. (*b*) With the Type II MOC, summary statistics of rating data are possible which extract the maximum of information from the ratings with a minimum of the bias effects that may result from the way subjects interpret the rating categories.

2. Conclusions drawn from detection analyses of recall data are on much more tenuous ground than those drawn from recognition. First, there is the dilemma of whether to force or not to force recall. Forced recall (i.e., forced list-length matching) may produce a complementary relationship between Hs and CRs and generate artifactual results. Nonforced recall, on the other hand, leaves intrusion probabilities up to uncontrolled variables—perhaps ones confounded with experimental variables. Second, whether or not recall is forced, the identity of the intrusions emitted presents an additional problem for any analysis, such as the Type II MOC, which is based on confidence ratings of recalled material. Whatever intrusions the subject chooses to emit will function as the new or "noise" distribution in SDT analyses, and analyses based on such a distribution cannot be compared to one based on an experimenter-provided distractor set. Furthermore, the similarity of the subject's intrusions to his recalled targets may vary as a function of experimental conditions and thereby contaminate comparisons across conditions.

3. Murdock's (1968) discriminability analyses of recall offer some promise of overcoming the problems of previous techniques. If Murdock's methods can be shown to be relatively insensitive to recall instructions, and if they can be extended to data not tied to SP cues, they will be of great value. Until these or other methods are developed, investigators should validate any conclusions drawn from detectability analyses of recall with other evidence.

4. Further uses of SDT for explicit scaling of memory-based discrimination promise a detailed and exact analysis of a number of memory processes.

5. Fuller use of SDT's ability to isolate criterial factors in memory would be realized if C_j were used instead of β as a measure of response criterion. The C_j is the z score of the criterion, measured in terms of the new distribution, and is a more suitable psychological index of criterial threshold than β. The limitations of C_j should, of course, be observed. This measure is probably invalid when used to compare criterial levels across MOCs that are not practically coincident or are non-Gaussian. A measure of criterion with the advantages of C_j, but with greater flexibility, would be a welcome addition to SDT.

6. The period of overenthusiastic use of MOC functions to test models of the threshold characteristics of memory traces is undoubtedly over. The SDT techniques will continue to provide data for controversies about all-or-none learning and related topics, but future arguments on the nature of the memory trace will be about theories (e.g., Bower, 1967; Norman & Rumelhart, in press) more refined than those dominating past arguments, and requiring data more sophisticated than the MOC for verification.

7. The SDT techniques extract from appropriate data a scale value representing the degree to which a given set of learned items has been retained. Percentage correct is another such scale value, but it has been shown to be inferior to those of SDT in situations where it can be compared with them. There are, however, approaches to measuring memory strength which are based on scaling theories other than SDT, and which may have favorable characteristics comparable to those of SDT. These other theoretical approaches have not, it is true, been pursued as far as SDT in retention studies, but this may be a defect in popularity, not potential. For examples of nonSDT scaling in memory, the reader is referred to papers by McConkie (1969) and Morton (1969).

REFERENCES

ALLEN, L. R., & GARTON, R. F. The influence of word-knowledge on the word-frequency effect in recognition memory. *Psychonomic Science,* 1968, 10, 401–402.

BANKS, W. P. Criterion change and response competition in unlearning. *Journal of Experimental Psychology,* 1969, 82, 216–223.

BARNES, J. M., & UNDERWOOD, B. J. "Fate" of first-list associates in transfer theory. *Journal of Experimental Psychology,* 1959, **58**, 97–105.

BERNBACH, H. A. Decision processes in memory. *Psychological Review,* 1967, **74**, 462–480.

BOWER, G. H. A multicomponent theory of the memory trace. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation.* N. Y.: Academic Press, 1967.

CLARKE, F. R., BIRDSALL, T. G., & TANNER, W. P. Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America,* 1959, **31**, 629–630.

DONALDSON, W., & MURDOCK, B. B., JR. Criterion change in continuous recognition memory. *Journal of Experimental Psychology,* 1968, **76**, 325–330.

EGAN, J. P. *Recognition memory and the operating characteristic.* (Tech. Rep. No. AFCRC TN-58-51, AD 152650) Hearing and Communication Laboratory, Indiana University, 1958.

EGAN, J. P., GREENBERG, G. Z., & SCHULMAN, A. I. Operating characteristics, signal detectability, and the method of free response. *Journal of the Acoustical Society of America,* 1961, **33**, 771–778.

ELLIOTT, P. B. Tables of d'. In J. A. Swets (Ed.), *Signal detection and recognition by human observers.* N. Y.: Wiley, 1964.

FREEMAN, P. R. *Table of d' and β.* (Tech. Rep. No. APU 529/64) Applied Psychology Research Unit, Cambridge, England, 1964.

GOUREVITCH, V., & GALANTER, E. A significance test for one parameter isosensitivity function. *Psychometrika,* 1967, **32**, 25–33.

GREEN, D. M. General predictions relating yes-no and forced-choice results. *Journal of the Acoustical Society of America,* 1964, **35**, 1042.

GREEN, D. M., & MOSES, F. L. On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin,* 1966, **66**, 228–234.

GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics.* N. Y.: Wiley, 1966.

HOPKINS, R. H., & SCHULZ, R. W. Meaningfulness in paired-associate recognition learning. *Journal of Experimental Psychology,* 1969, **79**, 533–539.

KINTSCH, W. Memory and decision aspects of recognition learning. *Psychological Review,* 1967, **74**, 496–504.

KINTSCH, W., & CARLSON, W. J. Changes in the memory operating characteristic during recognition learning. *Journal of Verbal Learning and Verbal Behavior,* 1967, **6**, 891–896.

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

LARKIN, W. D. Rating scales in detection experiments. *Journal of the Acoustical Society of America,* 1965, **37**, 748–749.

LEE, W. Relationship between Thurstone category scaling and signal-detection theory. *Psychological Bulletin,* 1969, **71**, 101–107.

MANDLER, G., PEARLSTONE, Z., & KOOPMANS, H. S. Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 410–423.

MARASCULIO, L. A. Extensions of the significance test for one-parameter signal detection hypotheses. *Psychometrika,* in press.

MARKS, L. E., & MILLER, G. A. The role of semantic and syntactic constraints in the memorization of English sentences. *Journal of Verbal Learning and Verbal Behavior,* 1964, **3**, 1–5.

MCCONKIE, G. W. Response hierarchy changes resulting from learning additional associations. *Journal of Experimental Psychology,* 1969, **79**, 495–503.

MORTON, J. Interaction of information in word recognition. *Psychological Review,* 1969, **76**, 165–178.

MURDOCK, B. B., JR. Signal detection theory and short-term memory. *Journal of Experimental Psychology,* 1965, **70**, 443–447.

MURDOCK, B. B., JR. The criterion problem in short-term memory. *Journal of Experimental Psychology,* 1966, **72**, 317–324.

MURDOCK, B. B., JR. Serial order effects in short-term memory. *Journal of Experimental Psychology,* 1968, **76**, (4, Pt. 2).

NORMAN, D. A. Acquisition and retention in short-term memory. *Journal of Experimental Psychology,* 1966, **72**, 369–381.

NORMAN, D. A., & RUMELHART, D. E. A system for perception and memory. In D. A. Norman (Ed.), *Models of memory,* N. Y.: Academic Press, in press.

NORMAN, D. A., & WAUGH, N. C. Stimulus and response interference in recognition-memory experiments. *Journal of Experimental Psychology,* 1968, **78**, 551–559.

NORMAN, D. A., & WICKELGREN, W. A. Short-term recognition memory for single digits and pairs of digits. *Journal of Experimental Psychology,* 1965, **70**, 479–489.

OGILVIE, J. C., & CREELMAN, C. D. Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology,* 1968, **5**, 377–391.

PARKS, T. E. Signal-detectability theory of recognition-memory performance. *Psychological Review,* 1966, **73**, 44–58.

POLLACK, L., & DECKER, L. R. Confidence ratings, message reception, and the receiver operating characteristic. *Journal of the Acoustical Society of America,* 1958, **30**, 286–292.

POLLACK, I., & HSIEH, R. Sampling variability of the area under the ROC curve and of d'_e. *Psychological Bulletin,* 1969, **71**, 161–173.

SCHULMAN, A. L. Word length and rarity in recognition memory. *Psychonomic Science,* 1967, **9**, 211–212.

SHEPARD, R. N., & TEGHTSOONIAN, M. Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology,* 1961, **62**, 302–309.

SLAMECKA, N. J. Recognition of word strings as a

function of linguistic violations. *Journal of Experimental Psychology,* 1969, **79**, 377–378.

SWETS, J. A. Is there a sensory threshold? *Science,* 1961, **134**, 168–177.

SWETS, J. A., TANNER, W. P., & BIRDSALL, T. G. Decision processes in perception. *Psychological Review,* 1961, **68**, 301–340.

WAUGH, N. C., & NORMAN, D. A. Primary memory. *Psychological Review,* 1965, **72**, 89–104.

WICKELGREN, W. A. Short-term recognition memory for single letters and phonemic similarity of retroactive inhibition. *Quarterly Journal of Experimental Psychology,* 1966, **18**, 55–62.

WICKELGREN, W. A. Exponential decay and independence from irrelevant associations in short-term recognition memory for serial order. *Journal of Experimental Psychology,* 1967, **73**, 165–171.

WICKELGREN, W. A., & NORMAN, D. A. Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology,* 1966, **3**, 316–347.

WINOGRAD, E. List differentiation as a function of frequency and retention interval. *Journal of Experimental Psychology,* 1968, **76**, (2, Pt. 2).

(Received September 29, 1969)