Deferred decision in human signal detection: A preliminary experiment[1]

JOHN A. SWETS, *BOLT BERANEK AND NEWMAN INC.*
THEODORE G. BIRDSALL, *UNIVERSITY OF MICHIGAN*

In the detection task characterized by deferred decision, the observer is allowed to determine how many observations he will make before deciding whether or not a signal is present, and he is called upon to balance the goals of maximizing accuracy and conserving time. The human observer appears capable of using the optimal process of cumulating sensory information over successive observations, but certain common training procedures lead him to use a less efficient process. Though he displays a consistent decision bias, his performance is also in good agreement with the optimal model of the process of terminating a sequence of observations.

The observation procedure most frequently considered in studies of signal detection is the *fixed-observation procedure.* On each trial a decision about signal existence is made after an observation interval of fixed length, or after a fixed number of observation intervals. The sole concern is for the quality of this decision. To maximize its quality, the observer must observe for as long as possible, that is, throughout the fixed interval or intervals.

Two other observation procedures employ an observation interval of variable length, or, in practice, a variable number of observation intervals. On each trial, in addition to making a decision about signal existence, the observer decides how many intervals there will be. These procedures recognize the goal of conserving time as well as the goal of maximizing the quality of the decision about signal existence. Because the quality of the decision about signal existence increases with increasing observation time, the two goals are in opposition, and the observer must establish an appropriate balance between them.

Under the *predetermined-observation procedure* the observer is called upon to decide in advance of each trial how many observations he will take on that trial. The optimal number depends upon the signal-to-noise ratio, the a priori probability of signal existence, the values of the outcomes of decisions about signal existence, and upon the cost of observing. Under the *sequential-observation, or deferred-decision, procedure,* the optimal number of observations depends upon the same factors and also upon the observations themselves. The observer in this procedure is permitted to decide as the trial proceeds how many observations to take; he can balance the two goals throughout a trial, and the decision to terminate a trial can therefore depend upon what has already been observed on that trial.

Signal detection theory has drawn heavily in recent years from statistical decision theory, particularly from Abraham Wald's classic works on fixed and sequential sampling (1947, 1950). Theory for fixed and sequential observation procedures in signal detection was presented by Peterson, Birdsall, and Fox (1954). Birdsall and Roberts have developed theory for the predetermined-observation procedure (1965a) and have refined the theory of sequential observation (1965b). In this paper we consider the performance of human observers, primarily in the sequential-observation, or deferred-decision, procedure. Several earlier studies of human observers, which emphasized the fixed-observation procedure, have been collected in one volume (Swets, 1964) and have been summarized in a comprehensive review by Green and Swets (1966).

The experiment reported here was designed, in part, to answer several questions concerning *the quality of human decisions about signal existence—or the detectability of the signal—as a function of observation time.*

(1) What process does the observer use to cumulate the sensory information in successive observations? According to modern detection theory the optimal process is to integrate the evidence from the multiple observations into a single basis for decision, specifically, to base the decision on the product of the likelihood ratios of the individual observations. An alternative process, often considered in the context of classical threshold theory, is to make a yes-no decision for each observation and to combine these binary decisions according to some rule.

(2) What is the rate of gain in decision quality over successive observations, or, stated otherwise, what is the rate of gain in signal detectability over successive observations? In a procedure with a fixed number of observations on each trial, if the observer multiplies likelihood ratios, the index of quality or detectability that is denoted d' should increase as the square root of the number of observations. The growth of detectability if the observer combines binary decisions depends upon the rule of combination he uses. This growth also depends upon the analysis employed. When only correct detections (hits) are considered in the analysis, detectability apparently increases more rapidly than the prediction based on the optimal process; the predicted gain in detectability falls short of the optimum if false-alarm responses are also taken into account.

(3) Does prior training influence the process the observer uses to cumulate information over successive

observations, and does this training affect the amount of gain in detectability over successive observations? We consider the possibility that the kind of response required of a naive observer in a procedure with a single observation interval on each trial will influence later performance under a multiple-observation procedure. Specifically, we ask if experience with a *rating* response—a relatively fine-grained representation of the observation according to the probability of signal existence—will predispose the observer to use a process of cumulation similar to the integration of likelihood ratios, and if experience with a *yes-no* response—a binary representation of the observation —will predispose the observer to adopt some process of cumulating individual, binary decisions.

(4) What role is played by memory in the cumulation of information over successive decisions? In particular, in this experiment, we ask whether or not memory is aided, and whether or not the rate of gain in detectability is thereby greater, if the observer is required to record a decision after each observation in a trial—either a rating or yes-no decision—that summarizes the information he has gained up to that point.

The experiment reported here was also designed to answer some questions concerning *the nature of the human observer's decision to terminate a trial under the deferred-decision procedure.*

(1) Does the observer make a sequential analysis, so that the decision to stop observing depends upon the information in preceding observations on that trial, or does he predetermine the number of observations on each trial? The observer possesses a large amount of information before a trial begins—about signal strength, signal-presentation probability, and values and costs—perhaps enough to tempt him to fix in advance the number of observations per trial even though such predetermination is not required of him. He might also adopt a compromise between the sequential-analysis and predetermination processes, by generally fixing the number of observations but shading this number a little one way or the other if some unusually compelling evidence is gained during the trial.

(2) Is there a significant advantage in economy of the deferred-decision procedure over the fixed-observation procedure? The indication from statistical theory is that a sequential analysis will produce a given level of detectability in about one-half as many observations as required under the fixed-observation procedure.

(3) Do the observer's criteria for "yes" and "no" responses vary systematically with the number of observations? Given that the allowable number of observations per trial is bounded, the optimal process is to hold the response criteria essentially constant until the bound is closely approached, and then to make both criteria rapidly more lenient, so that upon the last observation allowed even the slightest deviation of the likelihood ratio from the neutral value will determine the appropriate response. An alternative process is to hold the criteria fixed throughout a trial; another is to make the criteria gradually more lenient over the course of a trial.

(4) What are the effects of changes in the a priori probability of signal existence? If the signal-presentation probability is high, the observer should be willing to respond "yes" after relatively few observations and "no" only after several observations, and conversely. Whereas predetermination of the number of observations may be almost as effective as performing a sequential analysis in the case of symmetric probabilities, predetermination becomes definitely less effective as the presentation probability is made more extreme. The information inherent in asymmetric probabilities should result in fewer observations, on the average, than the number taken with symmetric probabilities.

PROCEDURE

The signal was a tone of 1000 cps pulsed for 0.1 sec. It was presented through earphones in a continuous background of noise. The masking noise had a spectrum level of approximately 50 dB re 0.0002 d/cm^2.

In most of the experimental conditions the signal was presented on a trial with a probability $P(s) = 0.50$. On trials with multiple observation intervals, the signal was presented in all, or none, of the intervals in a trial. Following each trial, the observers were informed whether or not a signal had been presented on that trial. The observers were male high-school seniors. Daily sessions of 2 hr. contained five or six blocks of trials, with brief rest period between blocks.

Two groups of three observers served for six weeks, throughout ten experimental conditions. The first three conditions, termed "initial conditions," established the signal strength appropriate for the remainder of the experiment, and gave differential training to the two groups of observers. The last two conditions, termed "final conditions," were replicas of two early conditions, and thus provided a check on the stability of observers and equipment through the experiment. Of principal interest are the five intervening conditions which employed trials with multiple observation intervals, both fixed and variable numbers of observation intervals.

Condition 1, for both groups, used five levels of signal strength to obtain from each observer a psychometric function, that is, a graph showing signal detectability as a function of signal strength. The two-interval forced-choice procedure was employed: one of the two observation intervals on each trial contained a signal, with equal probability, and the observer had to choose the more likely interval. The proportion of correct responses, P(C), was converted to the detectability index d' by means of Elliott's tables (see Swets, 1964)—and d' was plotted against the quantity E/N_0, where E is the signal energy and N_0 is the noise-power density. For each group, a single value of E/N_0 was selected to best represent an average d' = 1.0, which corresponds in two-interval forced-choice to P(C) = 0.76, and only this value was used in the remainder of the experiment. Condition 2 simply determined P(C) and d' in the two-interval forced-choice procedure for each observer at the value of E/N_0 selected for his group.

Condition 3 gave training with a yes-no response to Group I and training with a rating response to Group II. In both cases the single observation interval on each trial contained a signal with probability P(s) = 0.50. The rating response communicates more finely than the yes-no response the observer's estimate of the likelihood that a signal existed during the interval. In this experiment the rating response consisted in placing each observation in one of four categories of signal likelihood. The results obtained with both kinds of response were converted to the index d' as described in the next section.

Condition 9 was a replication of Condition 3 (yes-no response for Group I, and rating response for Group II, based on a single observation interval) and Condition 10 was a replication of Condition 2 (two-interval forced-choice for both groups).

Trials with multiple observation intervals were presented in Conditions 4 through 8. The fixed-observation procedure was used in Conditions 4 and 5, with six observation intervals on each trial. In Condition 4 the observer recorded his decision about signal existence only after the sixth interval; in Condition 5 decisions about signal existence were made after each interval. As before, the observers in Group I made a yes-no response, and the observers in Group II made a rating response, in both conditions. In both conditions P(s) = 0.50.

The three conditions remaining to be described (6-8) yielded data on the deferred-decision procedure. They were alike except that the signal-presentation probability varied from one to another; these probabilities were 0.25, 0.50, and 0.75. Both groups of observers made a "yes" or "no" response to terminate a trial and a "continue" response prior to termination. The cost of taking an observation was set at one point; the value of a hit and of a correct rejection was +30 points, and the value of a false alarm and of a miss was -30 points. The observers were told that each of the points they accumulated was worth some fraction of a cent, and that this fraction would be determined later such that a cash bonus paid at the end of the experiment would amount to about $1.00 per session for "good" performance.

A deferred-decision trial could be terminated at any time after the first observation interval; however, in order to fix both the number of trials and the length of a session, the same number of observation intervals (10) were presented on every trial. (If only the length of the session were fixed, the observers could be expected to terminate all trials quickly and thereby have many trials, or to not terminate even the first trial, depending upon whether the expected value of a decision were positive or negative. If only the number of trials were fixed, the value to the observer of concluding the session would distort the cost per observation and the values of the decision outcomes as announced by the experimenter.) On trials containing a signal the observation intervals presented after termination also contained a signal, so that exposure to the signal was equated among the observers.

A summary of the experimental conditions, together with the number of sessions and trials devoted to each, is given in Table 1. The entries in parentheses under the heading "Sessions" indicate that the first session of Conditions 1 and 4 through 8 was considered as practice; only the data obtained in the remaining sessions are reported here. There were 22 sessions in addition to the six practice sessions. In these 22 sessions approximately 9000 trials and 23,000 observation intervals were presented to each observer.

Table 1. Summary of the Ten Experimental Conditions

No.	Condition	Sessions	Trials
1	Two-interval forced-choice (5 signal levels)	(1)3	3000
2	Two-interval forced-choice	1	1000
3	Fixed-Observation: 1 interval per trial. Group I: Yes-No; Group II: Rating	1	1000
4	Fixed-Observation: 6 intervals per trial. Response only after last interval. Group I: Yes-No; Group II: Rating	(1)3	510
5	Fixed-Observation: 6 intervals per trial. Response after each interval. Group I: Yes-No; Group II: Rating	(1)2	340
6	Deferred Decision P(s) = 0.50	(1)4	400
7	Deferred Decision Group I: P(s) = 0.25; Group II: P(s) = 0.75	(1)3	360
8	Deferred Decision Group I: P(s) = 0.75; Group II: P(s) = 0.25	(1)3	360
9	Fixed Observation: 1 interval per trial. Group I: Yes-No; Group II: Rating	1	1000
10	Two-interval forced-choice	1	1000

Table 2. Results of Initial and Final Two-Interval Forced-Choice Procedure

	Group I				Group II			
Observer	1	2	3	Average	1	2	3	Average
d', Condition 2	0.78	1.32	0.43	0.84	1.16	0.96	0.73	0.95
d', Condition 10	1.10	1.74	0.89	1.24	1.32	1.24	--	1.28

Table 3. Results of Initial and Final Yes-No and Rating Procedures

	Group I: Yes-No				Group II: Rating			
Observer	1	2	3	Average	1	2	3	Average
d', Condition 3	0.72	1.47	0.55	0.91	1.12	0.98	0.56	0.89
d', Condition 9	1.22	1.77	0.97	1.32	1.40	1.20	--	1.30

RESULTS OF INITIAL AND FINAL CONDITIONS

Psychometric Functions

The psychometric functions obtained from the six observers in Condition 1 are shown in Fig. 1. Approximately 600 trials were presented at each signal level. The plots of d' versus E/N_O can be seen to be reasonably linear, at least in the middle range, in accordance with the usual finding (see Green & Swets, 1966, Chapter 7).

The difference in sensitivity among the observers in Group I is somewhat greater than we would like, largely because Observer 2 is atypically sensitive. The top scale of Fig. 1 shows that the range in signal energy at d' = 1.0 is about 2 dB—from 10 to 12 in 10 log E/N_O. The range, and the absolute values, of the observers in Group II are in good agreement with the other two observers of Group I, and with previous experiments. The range is about 0.5 dB—from 11.5 to 12 in 10 log E/N_O. The value of E/N_O desired for use in the remainder of the experiment was the value corresponding to d' = 1.0. As indicated in the figure, $E/N_O = 13.7$ was selected for Group I, and $E/N_O = 14.9$ was selected for Group II.

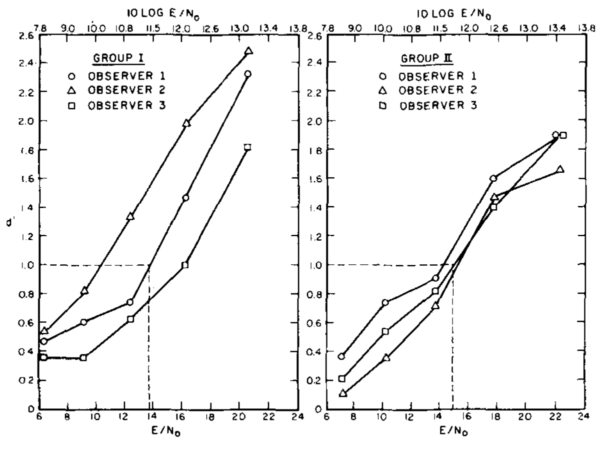

Fig. 1. Psychometric functions obtained in Condition 1.

Two-Interval Forced-Choice Procedure

The two-interval forced-choice procedure in Condition 2, at the single values of E/N_o used for each group, yielded the values of d' shown in Table 2. The average d' is 0.84 for Group I and 0.95 for Group II.

The results of the same procedure in Condition 10, also shown in Table 2, indicate a general increase in detectability during the experiment. In Condition 10, the average d' is 1.24 for Group I and 1.28 for Group II. If we use the individual psychometric functions of Fig. 1 to convert values of d' to values of 10 log E/N_O, we find that the average increase in detectability is the equivalent of about 0.75 dB in signal energy. (Observer 3 of Group II is not represented in Condition 10; he withdrew from the experiment after dition 6.)

Yes-No and Rating Procedures: Single Observation Interval

Values of d' obtained from the yes-no and rating procedures in Conditions 3 and 9 are given in Table 3. The values from Condition 3 agree rather well with the values from Condition 2 shown in Table 2. The same general increase in detectability from initial to final conditions that was observed in Table 2 is seen here: on the average the increases in d' are again the equivalent of about 0.75 dB in signal energy.[2]

As in the case of two-interval forced-choice responses, values of d' were obtained from yes-no responses by means of Elliott's tables (see Swets, 1964). These tables are based on the assumption that the probability density functions of noise and of signal plus noise are Gaussian and equal in variance. Values of d' were obtained from rating responses by a graphical procedure that is free of any assumption about the relative size of the variances of the density functions.

The graph employed is the receiver-operating-characteristic (ROC) graph, which is a plot of the proportion of hits versus the proportion of false alarms. In a yes-no procedure, these two proportions vary directly as the observer changes his criterion for a "yes" response—and trace a curve that corresponds to a given signal strength and a given value of d'. When probability scales are used on both axes, that is, when the normal deviates are spaced linearly, a linear ROC curve results from Gaussian density functions. This curve has a slope of unity if the density functions are equal in variance, and a slope less than unity if the variance of the signal distribution is greater than the variance of the noise distribution; in particular, the slope is equal to the ratio of standard deviations, σ_n/σ_s (Green & Swets, 1966, Chapter 3).

Rating responses distributed among four categories of signal likelihood produce three points along an ROC curve. Essentially, the three boundaries separating the four response categories are treated in analysis as response criteria, at three different levels, for a "yes" response, and thus three pairs of hit and false-alarm proportions are obtained (see Green & Swets, 1966, Chapter 3).

The ROC curves obtained from Group II in Condition 3 are shown in Fig. 2. Typical of rating data (Green & Swets, 1966, Chapter 4), the slopes are less than unity. A value of d' can be determined from the point where the ROC curve crosses the negative diagonal; this index is usually denoted d_e', but we dispense here with the subscript. At the negative diagonal, d' is equal to twice the value of the normal deviate, which is scaled at the top (and right) of the figure. In fitting a straight line to the three points, we have drawn the line through the middle point with a slope determined by the end points. This arbitrary procedure is based on the assumption that the middle point is the most reliable; the middle point is usually higher than the end points (note Observers 1 and 3), and, since this point is presumably the one that would be obtained if only a yes-no response were required, it may be less subject to depression caused by criterion variation. Of course, binomial variance also has a smaller range in the middle of the graph.

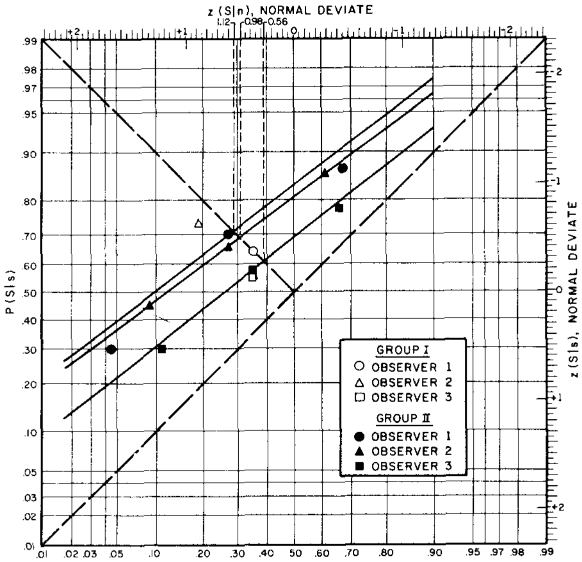

Fig. 2. Rating ROC curves obtained from Group II in Condition 3. The ordinate is the probability of a "yes" response, S, given a signal, s, or the probability of a hit. The abscissa is the probability of a "yes" or S response given noise alone, n, or the probability of a false alarm. These probabilities are estimated from response proportions obtained with the rating procedure as indicated in the text. Twice the value of the normal deviate (see the top scale), read from the point where the ROC curve intersects the negative diagonal, is taken as the detectability index d'. Also shown are the yes-no ROC points obtained from Group I in Condition 3, the open points.

Also shown in Fig. 2, as open points, are the yes-no ROC points obtained from Group I in Condition 3. Values of d' can be obtained from these points by taking the absolute value of the difference between the corresponding normal deviates, as scaled along the top and right side of the figure. These values, it can be seen, agree with those shown in Table 3. The location of these points, near the negative diagonal, indicates that each of the observers had adopted an approximately symmetrical decision criterion. This is as it should be, for both the signal-presentation probability and the payoff matrix were symmetrical; that is P(s) = 0.50, and the observers were told that the two kinds of correct response were equally to be sought, and that the two kinds of errors were equally to be avoided.

DETECTABILITY AS A FUNCTION OF OBSERVATION TIME

We turn now to results of the multiple-observation procedures to examine some factors and processes involved in the growth of detectability with increasing observation time. We concentrate on the fixed-observation procedure and fully use its results in this examination. We consider those partial results of the deferred-decision procedure that are relevant to the topic at hand. The larger part of the deferred-decision results is presented in the next major section of the paper where we consider the nature of the decision to terminate observation.

Recall that the fixed-observation procedure, with six intervals on each trial, was used in Conditions 4 and 5. In the former an overt decision about signal existence was made only at the end of the trial; in the latter cumulative decisions were recorded after each interval in the trial. Condition 5 thus yielded more detailed results, and they shall receive the bulk of our attention. Recall also that three experimental conditions (6, 7, and 8) employed the deferred-decision procedure, with three different signal-presentation probabilities.

We question first our assertion that there is a regular and substantial improvement in detectability over successive observations. We then ask whether or not the difference in the prior training of the two groups led to different rates of improvement, and we

Table 4. Results of Fixed-Multiple-Observation Procedure with Interim Responses. Entries are Values of d'

	Group I: Yes-No				Group II: Rating			
Observer	1	2	3	Average	1	2	3	Average
Stage 1	1.32	1.44	0.80	1.19	1.34	1.04	0.84	1.07
2	1.70	1.84	1.04	1.53	2.00	1.64	1.72	1.79
3	2.11	2.00	1.01	1.71	2.14	1.90	2.02	2.02
4	2.17	2.83	1.12	2.04	2.64	2.26	2.64	2.51
5	2.19	3.10	1.32	2.20	2.70	2.42	2.96	2.69
6	2.36	3.22	1.44	2.34	3.12	2.60	3.04	2.92
Prediction for stage 6 based on stage 1 and $\sqrt{m}$	3.23	3.53	1.96	2.91	3.28	2.56	2.06	2.62

make a preliminary attempt to infer from the observed rates of improvement the kind of process of cumulation that is used by the observers in each group. Lastly, we ask if detectability at the end of a trial is greater when responses have been made after each interval than when a response is made only after the last interval, that is, whether interim responses can be presumed to serve as an aid to memory in the process of cumulating sensory information.

Growth of Detectability

Table 4 shows d' at each stage of observation for Group I (yes-no response) and Group II (rating response) in Condition 5. There is clearly a steady and appreciable growth in detectability over the six observations of a trial for all six observers.

We can note in passing that much of the previously-noted increment in detectability from initial to final conditions has taken place by this point in the experiment: five of the six observers show an increase in d' from Condition 3 to stage 1 of Condition 5 (compare Table 3), and for four of them the increase is substantial.

Effect of Differential Training

Fixed-Observation Procedure. Perhaps the most striking result seen in Table 4 is that the observers who made a rating response improved more over successive observations than the observers who made a yes-no response. The average values of d' for the two groups are approximately the same at stage 1, with Group I having a slightly higher average—however, the average values of d' at stage 6 are 2.34 for Group I and 2.92 for Group II. The difference is seen quite clearly by comparing certain individual observers in the two groups: Observer 1 in Group I and Observer 1 in Group II have essentially the same d' at stage 1, but the latter shows a definitely greater rate of increase over the six stages of observation; similarly, Observer 3 of Group I and Observer 3 of Group II have very nearly the same d' at stage 1, and the latter shows a decidedly greater rate of improvement over the six stages.

We have come to assume a near invariance of the index d' over different psychophysical procedures, including the yes-no and rating procedures, but, of course, it is prudent to check the assumption with the present data. Our check took the form of analyzing the rating data as yes-no data. We considered in analysis the rating categories 1 and 2 as a "yes" response and the categories 3 and 4 as a "no" response. We obtained in this way proportions of hits and false alarms for the approximately symmetrical criterion, and, just as with the yes-no data of Group I, we used these quantities to determine values of d' from Elliott's published tables. The assumption of invariance was justified by this test. At stage 1 the rating and yes-no indices for the three observers of Group II differ by no more than 0.06. At stage 6, the yes-no d' for Observer 1 is 3.28, as compared with the rating d' of 3.12; the corresponding values for Observers 2 and 3 are 2.47 and 2.60, and 3.02 and 3.04, respectively. The averages at stage 6 exactly coincide at 2.92. Thus, the interpretation that rating training leads to greater improvement over successive observations than yes-no training is supported; this result is evidently not an artifact caused by the use of different response modes in the multiple-observation procedure.

The difference between the two groups is not large —however, it is real. The same differential effect was found in a repetition of this experiment.[3] Moreover, as we shall see next, in this experiment the advantage of rating training persisted throughout the three deferred-decision conditions.

Deferred-Decision Procedure. Table 5 shows the average number of observations, $\bar{m}$, taken by each observer in each of the three deferred-decision conditions, together with the corresponding values of d'. Remarkably, and conveniently for our present purpose, the two groups yielded very similar average values of d' at each signal-presentation probability. Thus, at P(s) = 0.50, the average values of d' were 2.79 and 2.78. However, Group I required, on the average, about one more observation than Group II to reach this level of detectability (compare 4.1 and 3.2). At P(s) = 0.25 the difference is again approximately one observation (3.9 versus 3.0). At P(s) = 0.75 the absolute difference is slightly smaller, but the percentage difference is of the same magnitude (3.0 versus 2.3). The difference between the two groups in the average number of observations required to reach a given detectability is on the order of 30%.

The Process of Cumulating Sensory Information

We would expect the greater rate of increase in detectability in Group II than in Group I if, as mentioned earlier, training with the relatively fine-grained rating response led the observers in Group II to use a process of cumulation similar to the multiplication of likelihood ratios, while training with the binary, yes-no response led the observers in Group I to combine binary decisions. The likelihood ratio preserves all of the information in an observation relevant to a decision about signal existence. If the observation is a continuous variable, or if it is discrete with more than two values, then a binary classification of the observation discards relevant information.

If likelihood ratios are used in the optimal manner, that is, if their product is taken, then d' will increase as the square root of the number of observations. (The derivation of this prediction is given by Green and Swets, 1966, Chapter 9.) Specifying the prediction for the combination of binary decisions is more difficult, for many rules of combinations are possible. The observer may say "yes" after multiple observations if any one of his individual decisions is "yes," or only if all of his individual decisions are "yes," or if a majority of the individual decisions are "yes," and so forth. Indeed, he might form the likelihood ratio of the individual, binary decisions. We expect to consider these alternatives in a later paper, in connection with the larger numbers of data obtained in a second experiment. Here we shall briefly consider one rule for combining binary decisions, the one appearing most often in the literature, namely, that the final decision is positive if any one of the individual decisions is positive. First, however, let us examine the results of Groups I and II in relation to the prediction that d' increases as the square root of the number of observations.

The bottom row of Table 4 shows the values of d' predicted for each observer at stage 6 based upon the d' obtained at stage 1. It can be seen that none of the observers in Group I reach the predicted value of d'; on the average they fall short of the prediction by about 0.6 in d'. Two of the observers in Group II reach or exceed the prediction; on the average they exceed the prediction by 0.3 in d'.

Figure 3 compares obtained and predicted values of d' at all six stages of observation. The top of the figure shows results for individual observers: Group I on the left, and Group II on the right. On the logarithmic coordinates the prediction of $\sqrt{m}$ improvement in d' is a straight line with a slope of 0.5. The solid lines are lines with this slope; they originate at the value of d' for each observer obtained at stage 1. It can be seen that the observers in Group I generally fall beneath the prediction, and that the observers in Group II generally reach or exceed the prediction, throughout the six stages. Average results are given at the bottom of the figure. Here the dashed lines are straight lines fitted by eye to the six points. The line fitted to the average result of Group II has a slope of 0.5—and thus indicates that these observers are cumulating sensory information as effectively as if they were multiplying likelihood ratios. The line fitted to the average result of Group I has a shallower slope, of about 0.4—and thereby indicates that these observers are using a less efficient process, perhaps a combination of binary representations of the several observations.

As mentioned, we shall not attempt here to deter-

Table 5. Values of d' and Average Numbers of Observations per Trial, $\bar{m}$, in the Three Deferred-Decision Conditions

		Group I: Yes-No Training				Group II: Rating Training			
Observer		1	2	3	Average	1	2	3	Average
P(s) = 0.50	d'	2.83	3.93	1.62	2.79	2.57	2.74	3.02	2.78
	$\bar{m}$	4.6	3.8	4.1	4.1	2.5	3.2	3.8	3.2
P(s) = 0.25	d'	2.86	3.79	1.96	2.87	2.66	2.98	--	2.82
	$\bar{m}$	4.2	4.2	3.3	3.9	2.7	3.3	--	3.0
P(s) = 0.75	d'	2.57	3.96	1.50	2.68	2.40	2.81	--	2.60
	$\bar{m}$	3.3	3.2	2.5	3.0	2.2	2.4	--	2.3

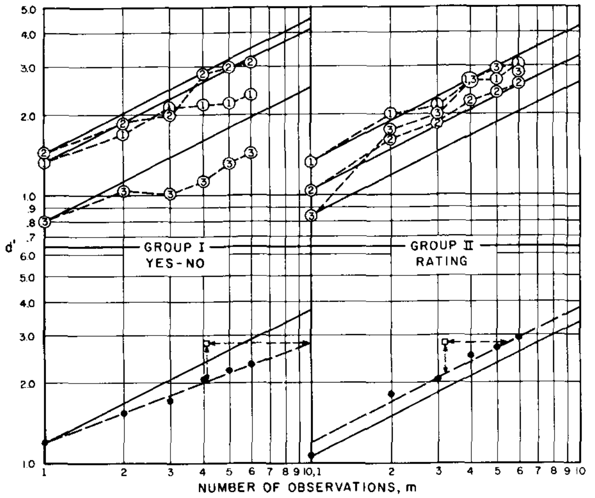

Fig. 3. Value of d' as a function of the number of observations for Groups I and II in Condition 5. The top of the figure shows individual results; the bottom of the figure shows average results. The circles represent data. The solid lines represent the prediction of $\sqrt{m}$ improvement in d'. The points plotted as squares were obtained with the deferred-decision procedure, and are discussed later in the text.

mine exhaustively which of the several possible rules of combination of binary decisions are consistent with the performance of the observers in Group I. The present data and space, however, are adequate to reject the simple rule that is most familiar. According to this rule, the observer says "yes" if a single observation indicates a "yes" response. The probability of a correct detection based on m observations is then $P_m = 1 - (1 - P)^m$, where P is the detection probability for a single observation. Inasmuch as the false-alarm probability is not zero in our experiment, the same formula is applied here to false-alarm responses as well, and the resulting pairs of proportions are plotted in the ROC space.

Figure 4 shows ROC points for each of the observers of Group I at each stage of observation. Each point can be converted to a value of d', which will agree with the values listed in Table 4, by taking the absolute value of the difference between the corresponding normal deviates, as scaled along the top and right side of the figure. To compare the data with the prediction at hand, we have taken the point [P(S | n) = 0.16, P(S | s) = 0.57] as representative of the three observers at stage 1. This point, symbolized by a diamond, taken together with the formula just given, generates the other diamond-shaped points shown for subsequent stages of observation. The values of d' predicted are 1.17, 1.44, 1.64, 1.81, 1.95, and 2.10. These predicted values consistently underestimate the average results shown in Table 4, but, of course, even a close correspondence of values of d' would be of little significance, given the large discrepancy between the locations of predicted and obtained points as seen in Fig. 4. The formula $P_m = 1 - (1 - P)^m$ predicts that the false-alarm proportion will increase steadily over successive stages; in fact, this proportion decreases for Observers 1 and 2, and remains relatively constant for Observer 3. It seems clear that the observers have not combined binary decisions by making a positive response for the sequence of observations whenever a single observation indicated a positive response.

In passing, let us simply note the ROC curves produced at each stage of observation by the three observers of Group II with a rating response. They are shown in Fig. 5 (a, b, c).

Cumulation with and without Interim Responses

Let us examine the results of Condition 4, in which responses were made only at the conclusion of the trial. Generally lower values of d' at stage 6 of Condition 4 than at stage 6 of Condition 5 would suggest that the interim responses of Condition 5 facilitated the cumulation of sensory information by aiding memory of previous observations.

Table 6 compares the values of d' at stage 6 for the two conditions. The first two rows show that, on the average, the observers in both groups performed as efficiently without as with the interim responses. Some individual differences are apparent, but we can conclude, at least, that the present results do not establish the efficacy of interim responses in the kind of task under study.

Let us briefly consider these results in connection with the previously-noted increase in detectability from the initial to the final conditions of the experi-

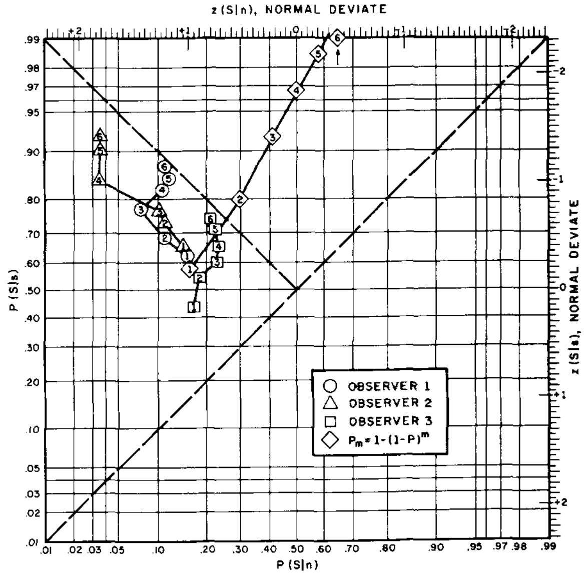

Fig. 4. ROC data from Group I in Condition 5. Points are shown for each of three observers at each of six observation stages. Also shown are results predicted by a commonly-considered rule for combining binary decisions (see text).

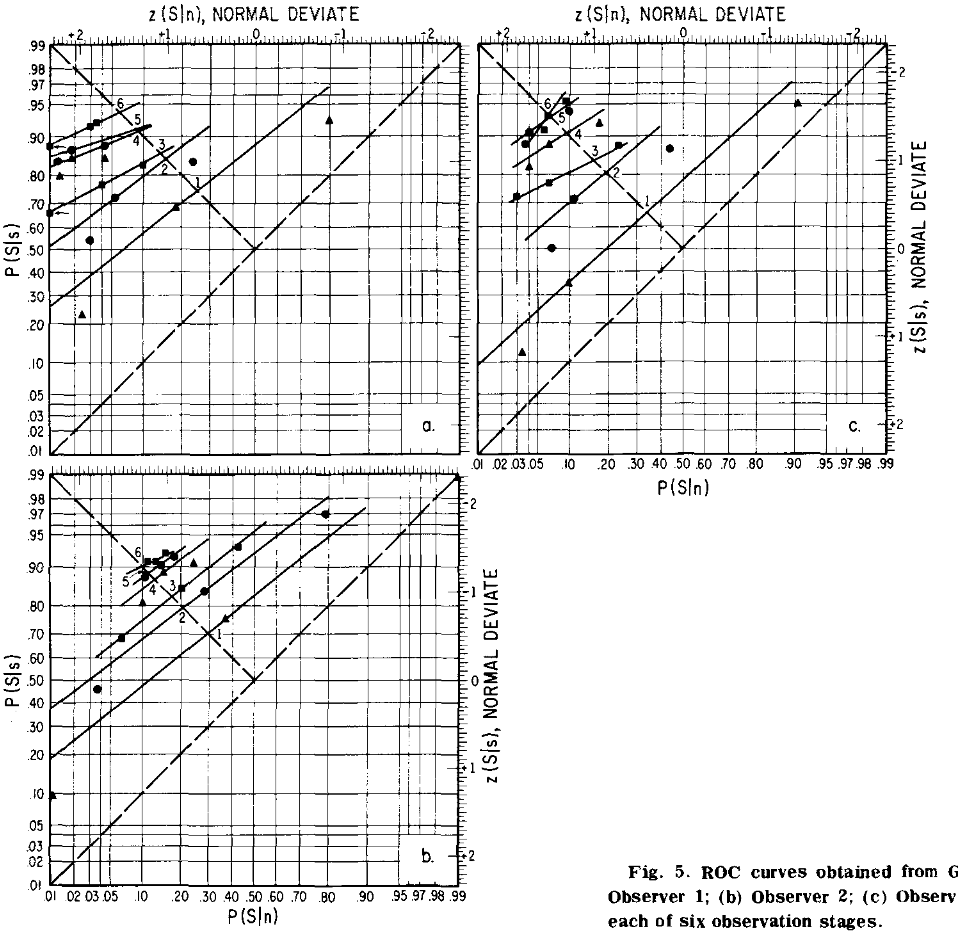

Fig. 5. ROC curves obtained from Group II in Condition 5: (a) Observer 1; (b) Observer 2; (c) Observer 3. Curves are shown for each of six observation stages.

ment, or, more exactly, from Conditions 2 and 3 to Conditions 5, 9, and 10. The fact that both groups reached as high a d' at stage 6 of Condition 4 as at stage 6 of Condition 5 could mean either of two things. One possibility to consider is that early in Condition 4 the observers became more adept at detecting signals, perhaps because of their experience in cumulating observations. We are inclined to discount this possibility; we find no clear evidence for day-to-day improvement within Condition 4, or for improvement from one block of trials to the next within the first session of this condition. A more likely alternative is that a modification of the equipment between Conditions 3 and 4 brought about inadvertently an increase in the effective signal-to-noise ratio. The effect, in any case, is not serious: the average variation in d' for one observation among Conditions 5, 9, and 10, and presumably also Condition 4, is the equivalent of about 0.33 dB.

We can note in the bottom two rows of Table 6 that all six observers in Condition 4 exceed the prediction of $\sqrt{m}$ improvement for stage 6, if Condition 3 is taken as the base. If stage 1 of Condition 5 is taken as the base, then, in Condition 4 as in Condition 5, Group I's average falls short of the prediction and Group II's average exceeds it. Group I does not do as well as Group II in Condition 4, which is consistent with the results of Condition 5.

THE DECISION TO TERMINATE A SEQUENCE OF OBSERVATIONS

Having examined some processes involved in the growth of detectability with increasing observation time, let us consider now the nature of the decision to terminate observation. Under the deferred-decision procedure the observer is limited by a maximum allowable number of observations on each trial, but he can terminate the trial sooner if he chooses. Because a cost is assessed on each observation he must balance conflicting goals to maximize the quality of his decision about signal existence and to minimize the number of observations. The observer can, of course, select a balance between these goals and determine the appropriate number of observations in advance of the trial—that is, he can use what we term a predetermined-observation *process*, as he must under the predetermined-observation *procedure*. However, the deferred-decision procedure allows him

Table 6. Comparison of Results Obtained Without and With Interim Responses. Entries are Values of d' at Stage 6

	Group I: Yes-No				Group II: Rating			
Observer	1	2	3	Average	1	2	3	Average
Condition 4, without interim response	2.22	3.81	1.38	2.47	3.08	3.20	2.36	2.88
Condition 5, with interim response	2.36	3.22	1.44	2.34	3.12	2.60	3.04	2.92
$\sqrt{m}$ prediction based on Condition 3	1.76	3.70	1.35	2.23	2.74	2.40	1.37	2.18
$\sqrt{m}$ prediction based on Condition 5	3.23	3.53	1.96	2.91	3.28	2.56	2.06	2.62

to determine how many observations to take as the trial progresses—he can therefore use a sequential-analysis process, letting his decision to terminate depend upon sensory information obtained during the trial.

Which of these two processes our observers used is a question to keep in mind throughout this section, for almost all of the analyses of data to be presented contribute to its answer. Three related topics are discussed individually. One is the relative efficiency of the deferred-decision and fixed-observation procedures. Another is the location of the observer's decision criteria, for "yes" and "no" responses, as a function of the stage of observation. Finally, we discuss the various effects of changes in the signal-presentation probability.

Sequential Analysis versus Predetermination

Predetermination of the number of observations on each trial is not as inefficient a process as it might seem at first glance. It is a flexible process that can vary according to changes in the signal-presentation probability, the signal strength, the values of the decision outcomes, and the cost of taking an observation. Used in optimal fashion, the predetermination process is more efficient than the standard approach of classical statistics, in which one asks "how large a sample is required to produce a certain result?" and obtains an answer without regard to the a priori probabilities of the hypotheses or the values of the decision outcomes.

Indeed, in some situations, it can be difficult to discern whether an observer is predetermining the number of observations or making a sequential analysis. Depending on the signal strength, the two processes in their optimal forms can lead to similar average numbers of observations, with differences in detectability, or to similar detectabilities and differences in the average number of observations (see Birdsall & Roberts, 1965b). Moreover, the observer might adopt a hybrid process; he might, for example, predetermine four observations but sometimes take three or five.

In the present experiment, the distribution of terminations over stages of observation indicates that all observers used the sequential-analysis process. Figure 6 (a, b) shows these distributions for the two groups of observers at each signal-presentation probability. We have separated trials on which a "yes" response was made from trials on which a "no" response was made; separating trials contingent upon the stimulus presented leads to very similar distributions. In general, the distributions are wider than we would expect to result from predetermination; certainly, no very strict predetermination was made by these observers.

We can note that the observers are quite similar, and that changes in the signal-presentation probability lead to systematic changes in the distributions. The previously-noted difference between the groups is reflected in this figure: the observers in Group II terminate sooner; their distributions are generally to the left of those of the observers in Group I.

The Economy of the Deferred-Decision Procedure

In some practical detection settings, requirements for simplicity and low implementation cost may dictate use of the fixed-observation procedure. In other settings one might be willing to underwrite the added complexity of the deferred-decision procedure if superior detection performance results. In principle, superior performance will result because the observer can take advantage of those sequences of observations in which the evidence happens to be very persuasive at an early stage of observation. A rule of thumb derived from Wald's (1947) work is that a given level of detectability is reached by a sequential analysis in about one-half as many observations as required in the fixed-observation procedure. How great a savings in time was made by our observers?

The answer is supplied in Fig. 3. The squares plotted in the lower half of the figure represent performance in the deferred-decision condition with P(s) =0.50 (Condition 6); the coordinate values of these points are given in the first two rows of Table 5. We can see that $d' \approx 2.8$ was reached by the observers

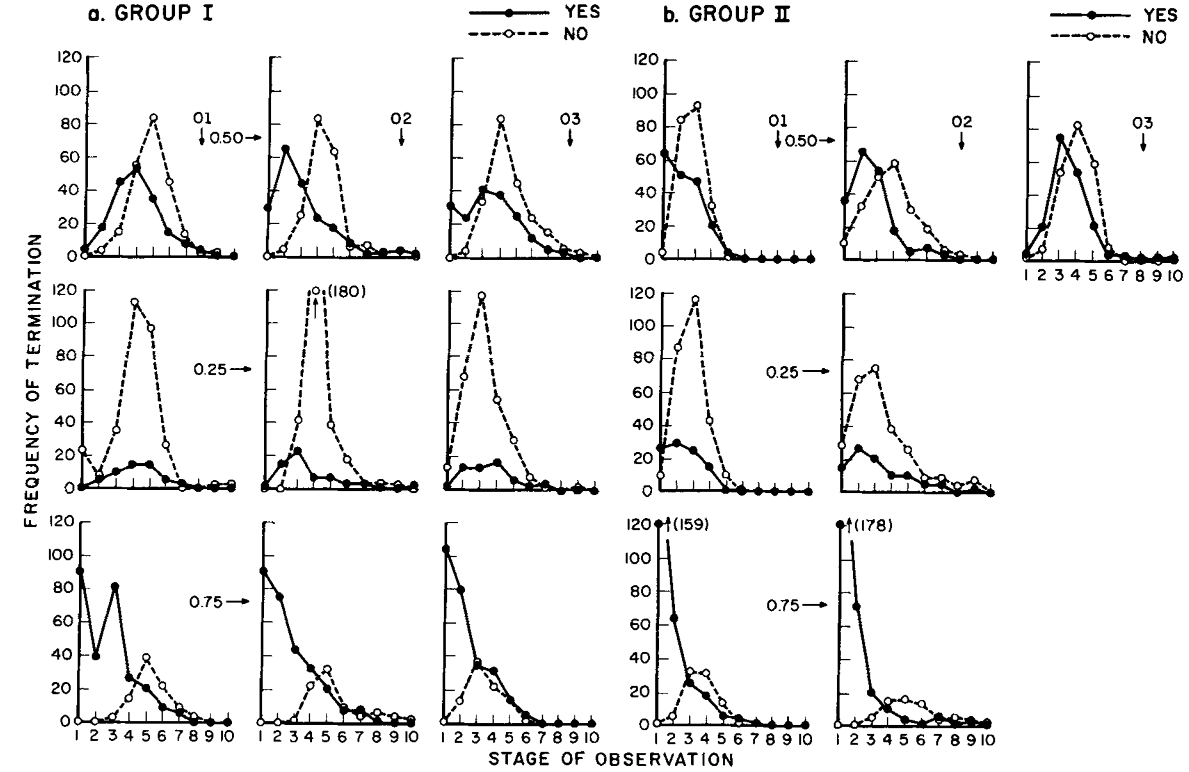

Fig. 6. The distribution of frequency of termination at each observation stage: a. Group I, b. Group II. The three observers in each group and the three conditions of signal-presentation probability are shown separately. In each panel of the figure, the two distributions are conditional upon the response made.

of Group I, on the average, in 4.1 observations. Extrapolating the fixed-observation results along the dashed line indicates that about 10 observations would have been required under the fixed-observation procedure to yield that high a value of d'. Group II reached d'≈ 2.8 in 3.2 observations with deferred decision, and in about 5.5 observations with a fixed number of observations. The savings of time in these two cases is close to 50%.

We can make a cut the other way, to determine the gain in detectability of the deferred-decision procedure for a given number of observations. For Group I the gain in d' at 4.1 observations, and for Group II the gain in d' at 3.2 observations, is approximately from 2.0 to 2.8.

Again it appears that these observers made a sequential analysis; the squares plotted in Fig. 3 would have fallen on the dashed lines if the observers were predetermining the number of observations.

Decision Criteria as a Function of Observation Stage

The decision criteria used by the observer at each stage of observation can be determined from the response-analysis-characteristic (RAC) graph, which is a plot of the probability of a signal given a "yes" response, $P(s|S)$, versus the probability of a signal given a "no" response, $P(s|N)$. The RAC graph will be derived and discussed in more detail in a later paper; here we simply list some of its properties. The RAC analysis is independent of values and costs and of the average number of observations. If performance is optimal, the RAC points will be independent of the a priori probability of a signal. Of major interest now are the results that RAC points at termination will lie on the negative diagonal if the decision criteria, for "yes" or "no" responses, are symmetrical; and that the RAC points will move toward the major diagonal if the two criteria become more lenient as the maximum allowable number of observations is approached.

Actually, the implications for decision criteria of the limited data of the present experiment are more easily seen if we convert from the RAC graph to what can be termed the "observation space." In the observation space the observer's basis for a decision ("yes," "no," or "continue") is plotted as a function of time. His basis for a decision is termed the "log odds ratio" and denoted L. Before the first observation, the only basis for a decision is the signal-presentation probability: $L = L_0 = \ell n\,[P(s)/P(n)]$. The probabilities of the causes, s and n, before and after a single observation x are related by Bayes theorem:

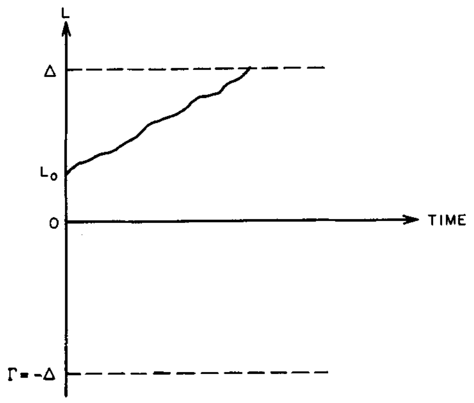

Fig. 7. The observation space, showing the basis for a decision L as a function of time. Δ and Γ are the decision boundaries for "yes" and "no," respectively: $\Delta = \ell n\,[P(s|S)/1\text{-}P(s|S)]$ and $\Gamma = \ell n\,[P(s|N)/1\text{-}P(s|N)]$. $L = L_0 = \ell n[P(s)/1\text{-}P(s)]$ before the first observation. As the observations proceed, $L = L_0 + \Sigma \ell n[\ell(x)]$, where $\ell(x)$ is the likelihood ratio.

L (based on x) $= L_0 + \ell n[\ell(x)]$, where $\ell(x) = f(x|s)/f(x|n)$ is the likelihood ratio, or the sensory datum. As the sequence of observations proceeds, $L = L_0 + \Sigma \ell n[\ell(x)]$. The decision criteria appear as boundaries in the observation space; when L crosses a boundary the appropriate ("yes" or "no") decision is made (see Birdsall & Roberts, 1965a, b).

Figure 7 depicts this observation space. The horizontal axis is time and the vertical axis is the decision basis L. The figure assumes $P(s) > P(n)$ and shows L crossing the boundary for a "yes" response, symbolized as Δ. The boundary for a "no" response is given as $\Gamma = -\Delta$.

In examining data we can plot L only if the observer reports L after each observation. It is a relatively simple matter, however, to determine the decision boundaries used by the observer, for they can be calculated from his "yes" and "no" responses. Specifically, the decision boundaries Δ and Γ are functions of the quantities plotted in the RAC graph, namely $P(s|S)$ and $P(s|N)$, respectively. The quantity Δ is the natural log of $P(s|S)$ divided by its complement: $\Delta = \ell n\,[P(s|S)/1 - P(s|S)]$. Similarly, $\Gamma = \ell n\,[P(s|N)/1 - P(s|N)]$. In plotting the decision boundaries used in the present experiment, we have made another conversion simply as a matter of convenience. Instead of plotting Δ and Γ on the ordinate, we have scaled the ordinate in units of $\ell n\,[P/(1-P)]$ so that $P(s|S)$ and $P(s|N)$ can be plotted directly.

Figure 8 shows the decision boundaries calculated from the data of the present experiment. It contains the data from all six observers in the three deferred-decision conditions. The values obtained for the "yes" boundary are plotted as squares; values obtained for the "no" boundary are plotted as circles. (Fewer than 6 x 3 = 18 squares and circles are plotted at each stage because values of 0 and 1 are not included and because one observer was absent from two conditions.) The results for different observers and conditions are not discriminated here because of insufficient data; we shall consider individual observers in reporting a second experiment in a following paper.

Points plotted as diamonds in Fig. 8 represent approximate medians for the sets of squares and circles, and the diamonds are connected by lines. We can see something of a trend: after about the third to the fifth observation, the decision boundaries tend to converge. The dashed lines in the figure bound the region where points occur and also indicate a convergence of the decision boundaries. These observers, roughly speaking, follow the optimal process; the trend of their boundaries is similar to the trend of the optimal boundaries (Birdsall & Roberts, 1965b). If the observers strictly predetermined the number of observations, the boundaries would, of course, converge abruptly at that number.

Effects of Signal-Presentation Probability

We might expect that the observers would take fewer observations, on the average, in the two conditions with an asymmetric signal-presentation probability, $P(s) = 0.25$ and 0.75, than in the condition with $P(s) = 0.50$. Prior information about the stimulus alternatives is provided by asymmetric probabilities, and the basis for decision L should deviate from zero, in the direction of the response boundary corresponding to the more likely stimulus alternative, before the first observation (see Fig. 7). A corollary is that a higher value of d' will be obtained at $P(s) = 0.50$ as a result of the greater average number of observations. The

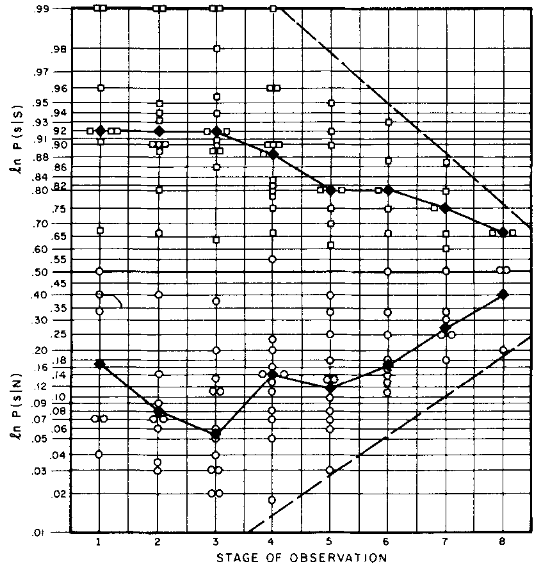

Fig. 8. Decision criteria, or boundaries, for "yes" response (top) and "no" response (bottom)for all six observers in the three deferred-decision conditions. (See the text for a discussion of the quantities plotted.)

Table 7. The Average Numbers of Observations in Deferred-Decision, Contingent Upon Stimulus and Response, as a Function of Signal-Presentation Probability

		Group I: Yes-No Training				Group II: Rating Training			
Observer		1	2	3	Average	1	2	3	Average
P(s) = 0.50	signal	4.2	3.2	3.7	3.7	2.3	2.7	3.6	2.8
	noise	4.9	4.4	4.5	4.6	2.7	3.7	4.1	3.5
	yes	4.1	3.1	3.5	3.6	2.2	2.6	3.5	2.8
	no	4.9	4.5	4.6	4.7	2.8	3.8	4.1	3.6
P(s) = 0.25	signal	4.2	3.9	3.5	3.9	2.4	3.2		2.8
	noise	4.1	4.3	3.2	3.9	2.8	3.4		3.1
	yes	4.1	3.9	3.7	3.9	2.4	3.3		2.9
	no	4.1	4.2	3.2	3.8	2.8	3.4		3.1
P(s) = 0.75	signal	2.7	2.6	2.4	2.6	1.9	1.8		1.9
	noise	5.1	5.2	3.3	4.5	3.4	4.5		3.9
	yes	2.6	2.6	2.2	2.5	1.8	1.7		1.8
	no	5.3	5.4	3.5	4.7	3.7	5.3		4.5

ROC curves (on probability scales) consistent with this expectation would diverge from the usual linear form; the middle point would be farther from the positive diagonal than the two end points and the curves would appear as rectangular hyperbolas.

Table 5 shows that both groups, on the average, took 0.2 fewer observations at P(s) = 0.25 than at P(s) = 0.50, and approximately one less observation at P(s) = 0.75 than at P(s) = 0.50. However, in the first comparison (between 0.25 and 0.50) the average result is a distortion of the performance of individual observers: only two of the five observers who served in both conditions took fewer observations at 0.25 than at 0.50. In the case of the difference between 0.75 and 0.50, the average result is a more valid representation; all five observers show a difference consistent in direction with the average result. We have not anticipated the differential effects of 0.25 and 0.75, but we shall discuss this bias shortly.

Table 5 also shows that values of d' at P(s) = 0.50 are not substantially higher than at P(s) = 0.25 and 0.75. In fact, in keeping with the empirical numbers of observations just noted, the average d' at 0.50 is slightly higher than at 0.75, and slightly lower than at 0.25. The variation in d' with signal-presentation probability is not large enough to alter the form of the ROC curve: Figure 9 (a, b) shows that the ROC curves from the deferred-decision procedure are very nearly linear.

Consider now breakdowns of the average numbers of observations contingent upon the stimulus presented and the response made. These data are shown in Table 7. It can be seen that the results for the two kinds of contingency are very similar, so we shall discuss only the response-contingent results. The observers adjusted their decision behavior when the signal-presentation probability was changed, and the adjustment was in the appropriate direction. On the whole, as compared with P(s) = 0.50, when the signal was less likely the observers took more observations to respond "yes" and fewer to respond "no," and when the signal was more likely they took fewer observations to respond "yes" and more to respond "no." Deciding thus quickly in favor of the more likely stimulus alternative and slowly in favor of the less likely stimulus alternative is consistent with a sequential-analysis process and inconsistent with the predetermination process.

A decision bias is apparent in these data. At P(s) = 0.50, where we might have expected the observers to take equally as many observations to say "yes" as to say "no," they took approximately one more observation to respond "no" than to respond "yes." A difference in this direction has also been observed in studies of reaction time (e.g., Bindra, Williams, & Wise, 1965). At P(s) = 0.75, approximately 2.5 more observations were taken before a "no" than before a "yes." Symmetrical decision behavior was obtained in this experiment with P(s) = 0.25; in this case the numbers of observations preceding a "yes" and a "no" were essentially the same.

CONCLUSION

The deferred-decision task in signal detection represents many practical detection tasks and everyday perception more accurately than does the fixed-observation task commonly used in psychophysics. The deferred-decision task provides a framework for studying the trading relationship between time and accuracy of performance—a relationship largely ignored in experimental psychology though central to most sensory, cognitive, and motor performances.

The data of this preliminary experiment show human observers to be capable of using the optimal observation processes, though a less efficient process is used under certain conditions of initial training. The

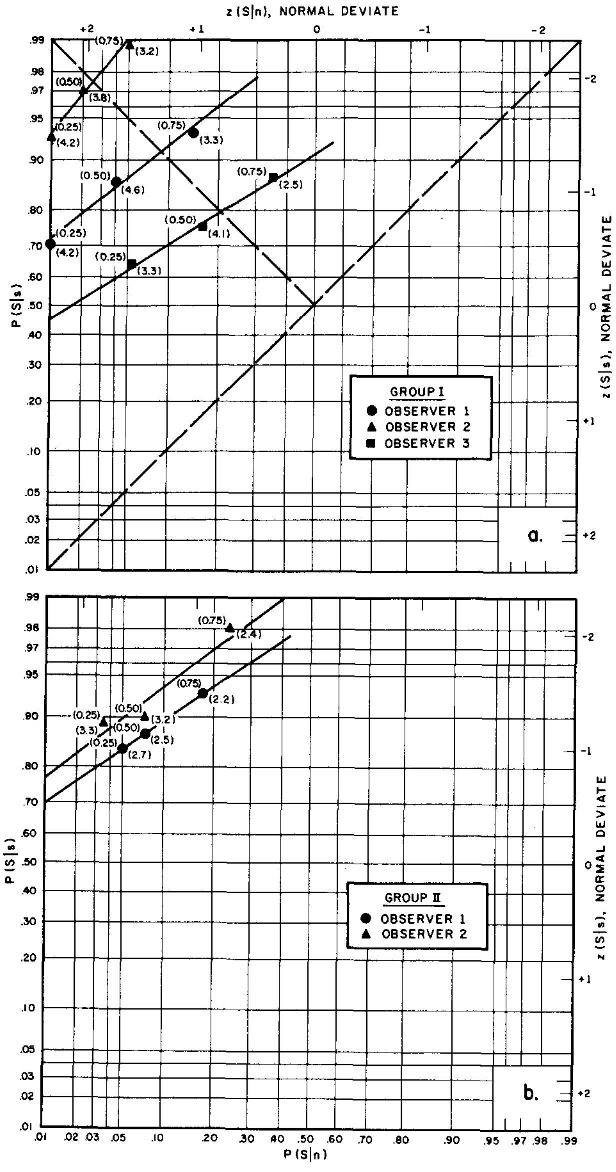

Fig. 9. ROC curves from the deferred-decision procedure: a. Group I, b. Group II. The numbers in parentheses above the points give the signal-presentation probabilities; the numbers in parentheses below the points give the average numbers of observations.

results also show that human observers are capable of using the optimal decision processes, though they give consistent evidence of a particular decision bias.

It is clear that the optimal models available for the deferred-decision task are sufficiently good approximations to human behavior to warrant more investigation in psychophysics of their detailed, quantitative predictions. Furthermore, the quantitative deviations of human from optimal behavior that have already been observed are sufficiently reliable within and among observers to justify application of the models and experimental results in practical detection situations.

References

Bindra, D., Williams, J. A., & Wise, J. S. Judgments of sameness and difference: Experiments on decision time. *Science*, 1965, 150, 1625-1627.

Birdsall, T. G., & Roberts, R. A. On the theory of signal detectability: An optimum nonsequential observation-decision procedure. *IEEE Trans. on Information Theory*, 1965a, IT-11, No. 2, 195-204.

Birdsall, T. G., & Roberts, R. A. Theory of signal detectability: deferred-decision theory. *J. Acoust. Soc. Amer.*, 1965b, 37, 1064-1074.

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

Peterson, W. W., Birdsall, T. G., & Fox, W. C. The theory of signal detectability. *Trans. IRE Professional Group on Information Theory*. 1954, PGIT-4, 171-212.

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

Wald, A. *Sequential analysis*. New York; London: Chapman and Hall, 1947.

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

Notes

[1] Dr. David M. Green advised us through all phases of this project. John B. Brown designed the laboratory apparatus; Paul F. Coughlin constructed and maintained it. Linda S. McElroy, Donna L. Darley, and Curt Freed programmed the PDP-8 computer to control the experimental procedure and to provide summaries of data. Mrs. McElroy collected the data and performed analyses of them, with the assistance of Joshua Coran and Susan Levin. Dr. Joseph Markowitz contributed to the analysis and evaluation of the results. Dr. Trieve Tanner, contract monitor, facilitated our work in several ways. None of these persons is responsible for the outcome; however, their contributions were highly capable and essential, and these contributions are gratefully acknowledged. This work was supported by the Human Performance Branch of the NASA-Ames Research Center.

[2] We shall present a probable reason for this increase in detectability after examining the results of Conditions 4 and 5. Fortunately, though we would have preferred greater stability, this change in detectability has no effect on our ability to answer the questions under study. Conditions 4 through 8 are of primary interest, and we shall see that detectability was as nearly constant over Conditions 4 through 10 as can be expected under any circumstances.

[3] Swets, Birdsall, Green, and Markowitz (paper in preparation).

(Accepted for publication November 7, 1966.)