THE UNIVERSITY OF MICHIGAN RESEARCH INSTITUTE
ANN ARBOR

THE THEORY OF SIGNAL DETECTABILITY AS AN INTERPRETIVE TOOL FOR PSYCHOPHYSICAL DATA

Technical Memorandum No. 78

Electronic Defense Group
Department of Electrical Engineering

By: Wilson P. Tanner, Jr.

Approved by: A. B. Macnee

AFCCDD TN 60-13

Contract No. AF19(604)-2277

Operational Applications Laboratory
Air Force Cambridge Research Center
Air Research and Development Command

May 1960

ensm

UMR1030

#6016 SIGNAL DETECTION AND PSYCHOPHYSICS - 308 pages
July 11-22, 1960 W. P. Tanner, Jr. in charge

J. P. Egan	"Operating Characteristics, Signal Detectability, and the Method of Free Response"	54 pages
W. P. Tanner	"The Theory of Signal Detectability as an Interpretive Tool for Psychophysical Data"	27 pages
T. G. Birdsall	"Detection of a Signal Specified Exactly with a Noisy Stored Reference Signal	23 pages
F. R. Clarke	"Confidence Ratings, Second-Choice Responses, and Confusion Matrices in Intelligibility Tests"	48 pages
J. A. Swets	"Decision Processes in Perception"	77 pages
F. R. Clarke	"The Theory of Signal Detectability"	23 pages
J. P. Egan	"Operating Characteristics in Speech Communications"	18 pages
R. K. Rift	"A Survey of the Theory of Probability"	19 pages
J. A. Swets	"A Brief Introduction to Decision Theory"	19 pages

308

TABLE OF CONTENTS

	Page
PREFACE	iv
ABSTRACT	vii
1. INTRODUCTION	1
2. STATEMENT OF THE PROBLEM	3
3. ASSUMPTION OF EXPECTED VALUE MAXIMIZATION	10
4. THE FOURIER SERIES BAND-LIMITED ASSUMPTION	13
5. THE FOURIER TRANSFORM BAND-LIMITED ASSUMPTION	17
6. COMPARISON OF THE SAMPLING THEOREMS	19
7. THE PHYSICAL APPROXIMATION	21
8. SUMMARY AND CONCLUSIONS	25
REFERENCES	26

PREFACE

I have written this paper to make explicit and clear the philosophy underlying the application of the theory of signal detectability to the study of psychophysics. I have undertaken this task because frequently criticisms of the applications have been brought to my attention. Usually it appears to me that the criticisms stem from the idea that we are trying to do something different from what we actually are, and consequently the criticisms are not relevant. By stating our position explicitly, I hope to make it possible for these criticisms to be aimed more directly to relevant points.

Two points in particular have attracted criticism. One is our use of the expected value criterion in our model; and the other is the use of the efficiency variable, η, dependent upon calculations derived from a particular finite sampling plan. In the text I have tried to make it clear why we are using these two concepts as we are. For our purpose it is unnecessary to think or believe that the expected value criterion is the one which people really use in everyday life. In fact we are aware that there are many situations in which this is an unreasonable criterion, and also of the large quantity of data which support the facts upon which the criticism is predicated. We contend that, while the criticism as made is supported by strong experimental evidence, it is not applicable.

The criticism of the use of the efficiency variable, η, has been more difficult to handle: partly because those making the argument have been exceedingly persistent and partly because their arguments have been presented in more elegant form. It has not always been easy to iden-

tify exactly how the problem they have solved differs from the problem we use, nor is it easy to determine exactly what the issue is. I sometimes suspect that the criticism is predicated on the notion that we intend to produce efficiency statements which can be used in engineering applications. This is a purpose which I have considered only superficially. Our purpose is to use the efficiency statements to tell us how to interpret data to develop a model which will describe the sensory behavior of a human observer. Of course, the model could be presented as it develops without any explanation of the mental processes involved in its development. This would make it more difficult to evaluate and more difficult to criticize. It seems wiser to make the underlying philosophy explicit.

The argument over sampling theorems has produced its profits. It has forced those involved to turn their attention to underlying philosophies: a subject matter which scientists frequently try to avoid. Differences in philosophy frequently lead to apparently unresolvable arguments since it is not always recognized that the points of view differ because they are predicated on different basic assumptions. The arguments should be concerned with the assumptions underlying the conclusions drawn from them.

The discussions of the sampling theorems have increased my confidence that, for our purposes, the use of η is sound. In particular, I am impressed with some recent ideas presented by Dr. Claude Shannon which I interpret as forming the basis for an argument which supports our use of the finite sampling plan. I would like to express my appreciation to Dr. Shannon for permitting me to read a rough draft of his paper and for discussing this matter in detail with me.

T. G. Birdsall of the Electronic Defense Group and Julian H. Bigelow of the Institute for Advanced Study have spent hours discussing this subject with me and to them I am deeply grateful for helping me formulate the problem. In addition I acknowledge my debt to the many persons who have attended informal meetings with me for the purpose of thorough reviews and rehashing of the problems. These include M. V. Matthews and E. E. David of Bell Telephone Laboratories, John A. Swets and David Green of the Massachusetts Institute of Technology, J. C. R. Licklider of Bolt, Beranek and Newman, and Frank R. Clarke and Allan Macnee of the Electronic Defense Group. In writing this paper I have leaned heavily on the discussions we have had; and while I must acknowledge that much of the discussion in this paper stems from their ideas, I must also take full responsibility for the statements which follow. I suspect that some of these people will not agree with all of the statements.

ABSTRACT

The theory of signal detectability is examined from the standpoint of determining a set of satisfactory assumptions for the purpose of developing an interpretive tool for use in psychophysical experiments. It is concluded that the assumption that the observer attempts to maximize the expected value of the outcome of the experiment is satisfactory for this purpose, and that a set of physical conditions can be established which justify a computation of the detectability of a signal in noise based on a finite sampling plan involving 2WT amplitude values over the open interval, 0 to T.

THE THEORY OF SIGNAL DETECTABILITY

AS AN INTERPRETIVE TOOL FOR PSYCHOPHYSICAL DATA

1. INTRODUCTION

In order to use a mathematical model in a scientific investigation, it is necessary to show that there is a satisfactory agreement between the conditions of the phenomena under investigation and either the assumptions of the model or an equivalent set of assumptions. This can be done in at least two ways: by altering the conditions of the investigation to agree with the assumptions of the model; or by adding to, or modifying, the assumptions of the model so that it more nearly agrees with the conditions of the investigation. Usually both methods are required to establish an adequate agreement. It is the purpose of this paper to examine the techniques which have been employed in establishing agreement between psychoacoustic experiments and the theory of signal detectability.

The theory of signal detectability is based, as all mathematical theories are, on precisely stated assumptions. While it is likely that the theory was developed to increase understanding about some interesting and urgent practical problems, the assumptions were chosen not only because of the application but also to permit mathematical manipulation. It is a rare event that both purposes can be satisfied simultaneously; and it is the mathematicians' tendency to prefer com-

promise in favor of permitting manipulation. In almost every case, it is necessary to make some compromise if one is going to apply a mathematical model to either an experimental or "real-life" problem.

In applying the model of signal detectability to psychoacoustics the problem is made even more difficult since psychoacoustics was not among the interesting and urgent problems the mathematicians had in mind when they developed the theory. The initial studies appeared during World War II as a result of the need for detecting radar signals embedded in noise. In this context Siegert (Ref. 1) presented his concept of the ideal observer.

After the war interest continued and in 1954 Peterson, Birdsall, and Fox (Ref. 2) and Van Meter and Middleton (Ref. 3) presented independent developments which described a far more sophisticated ideal observer than Siegert's.

The discussion in this paper will be based on the paper of Peterson, Birdsall, and Fox since it is presented in a more useful form. It is presented in two parts: the first part presents the general theory and the second considers some special cases. The general theory demonstrates that the theory of signal detectability is a special case of the theory of testing statistical hypotheses and decision theory. The second part consists of a series of studies in which some specializing assumptions are made to tailor the theory to handle certain hypothetical cases in a quantitative way. This part brings into the context of the theory of signal detectability the relation between such parameters as the signal energy and the noise energy, and the separation between the statistical hypotheses condi-

tional upon the existence of noise alone and those conditional upon the existence of signal plus noise.

The general theory appears directly applicable to the study of psychoacoustics. However, because of its generality, it is not a very powerful tool. It does not contain the mechanism for quantitative prediction and, without quantitative prediction, it does not lend itself to experimental use. Specializing assumptions are required which tailor the theory to agree acceptably to the conditions encountered in psychoacoustic experiments. The agreement between these assumptions and the experimental conditions must be carefully evaluated, for the success of the use depends on this agreement. The evaluation is based partly on experimental evidence and partly on faith. The faith in turn is based on convincing, although not conclusive, arguments.

This paper presents certain logical considerations supporting the convincing arguments for the adequacy of the agreement between the specializing assumptions employed and the conditions encountered in the psychoacoustic laboratory.

2. STATEMENT OF THE PROBLEM

The concept of the ideal observer can be illustrated by a block diagram (Figure 1). The task of this observer is to accomplish an optimum mapping from an input space onto an output space. According to the theory of signal detectability, this is accomplished by computing the likelihood ratio associated with the input and comparing this

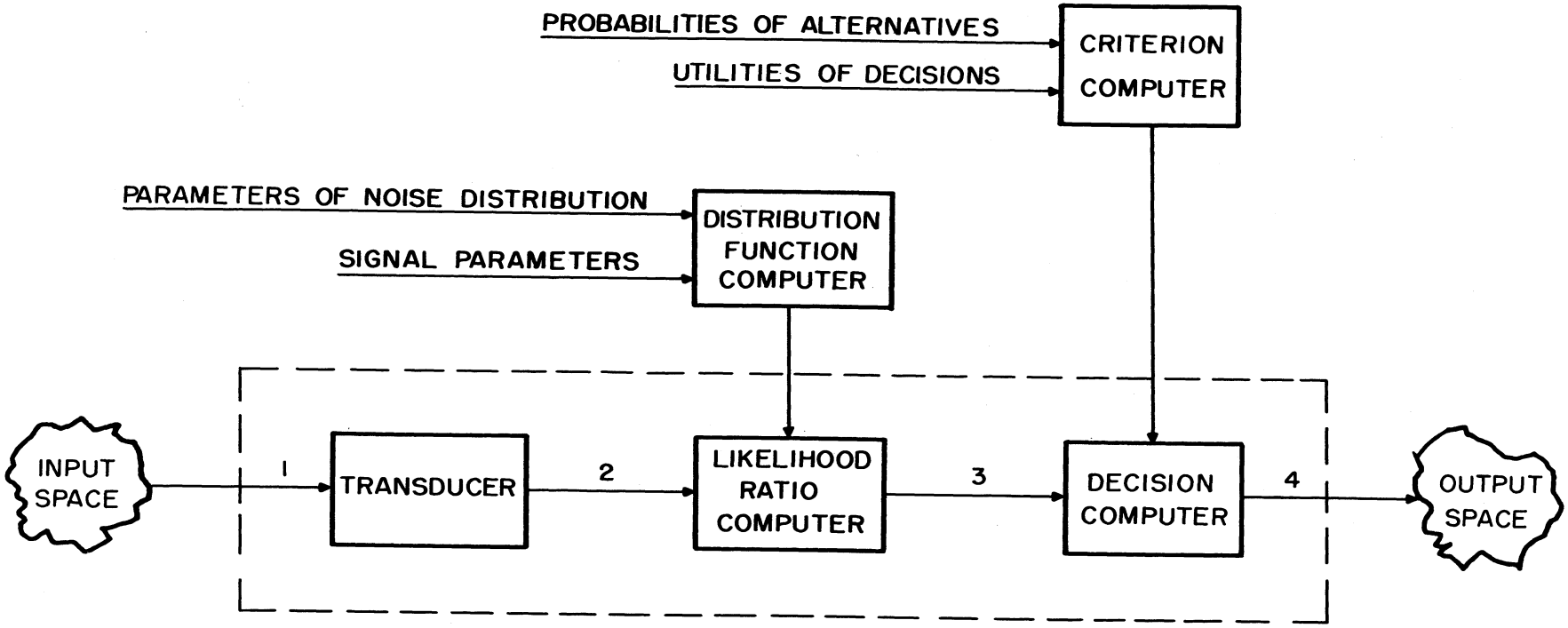

FIG. 1 BLOCK DIAGRAM OF IDEAL OBSERVER

with a number. If the likelihood ratio turns out to be greater than this number, then the input is said to include both signal and noise. If it is less than this number, then the input is said to include only noise.

This is a general statement applying to all detection problems within the framework of the theory. The statement in its general form does not furnish the framework for quantitative predictions of the performance of the ideal observer, i.e., what percentage of detections and what percentage of false alarms might be expected in any given detection task. It is necessary to examine the statement to see how it must be expanded in order to permit such predictions. It is also necessary to consider the problem of why these predictions are desired.

The likelihood ratio is defined as $\ell(x) = \frac{f_{SN}(x)}{f_N(x)}$ where $f_{SN}(x)$ if the likelihood, or probability density, of the hypothesis signal plus noise leading to the input x; and $f_N(x)$ is the likelihood of the hypothesis noise alone leading to the input x. In order to compute the likelihood ratios it is necessary to assign numbers to the likelihoods $f_{SN}(x)$ and $f_N(x)$ for all possible values of x. In other words, it is necessary to know the distributions of x given that the hypothesis signal plus noise is true and given that the hypothesis noise alone is true. In order to compute these distributions, the ideal observer includes a distribution computer.

These distributions depend upon the parameters of the signal, the parameters of the noise, and the way in which the signal and noise are combined. Only if one is willing to specify these conditions in sufficient detail to permit calculation can predictions of false-alarm

rate and detection rate be made. It is, therefore, necessary to make specializing assumptions about the particular environment in which the ideal observer is going to operate.

In the general theory, the rule for decision is to state that the signal exists whenever $\ell(x) \geq W$, otherwise state that noise alone exists. W is specified only to the extent that it is a number - an arbitrary number. Before predictions can be made of detection rates and false-alarm rates it is necessary to specify W as a particular number. There are a number of ways in which W might be specified and each depends on the particular purpose for which the ideal observer is designed. Six ways of choosing a number for W are reviewed by Birdsall (Ref. 4). These include such methods as maximizing correct decisions, maximizing expected value, maximizing detection rate for a given, fixed false-alarm rate and maximizing the preservation of information. To make predictions of false-alarm rate and detection rate, it is necessary to assume a method for computing W. For this computation, the ideal observer has a decision computer and this, in turn, needs to know the method for computing W and the values of the variables necessary to make the computation.

Thus, in order to make quantitative statements based on the model, it is necessary to make specializing assumptions to permit the distribution computer and the criterion computer to perform their respective functions. Any one of a number of assumptions will permit the operation in either case. This choice is not completely arbitrary; some conditions are imposed on the selection by the purpose one has in mind for making the assumptions.

Before attempting to select the particular assumptions, it is necessary to consider exactly why it is desirable to have a model which leads to these quantitative predictions. In this paper two reasons, leading to different criteria for selecting the specializing assumptions, will be considered. One of these is that one may want a model which is descriptive or explanatory. A model of this sort plays the role of a scientific theory describing the relations between classes of observable phenomena. The assumptions for a model of this sort must lead to predictions which permit experimental verification. The other is that one may want a model which he can use as an interpretive tool in considering the implications of experimental data with reference to particular questions he is asking. In this case, he may find it reasonable to make assumptions which can be fairly well satisfied only after appropriate experimental manipulation. The two cases might lead to quite different assumptions.

If a psychoacoustician is interested in a descriptive model which explains how people interpret acoustic stimuli in the course of their everyday experience, he must concern himself with the types of signals and the types of noise encountered in what he hopes are typical environments. He must also conduct experiments to find out which of the methods for the selection of W appear to be consistent with behavior. Based on the evidence, one might make a cautious statement which says: "People behave in a way which can be described as selecting W by a particular method." That they actually choose a number may not be determinable. One might also find experimentally that there are lower bounds which appear to be placed on W (that a

threshold exists), but this, of course, is an experimental question.

On the other hand, the psychoacoustician may be interested in studying the auditory equipment available to observers in listening to acoustic signals. He may find that the analyses he wishes to perform can best be carried out if he uses signals and noises which are atypical to those of everyday environments. This may permit greater agreement between the physical conditions of the experiment and the assumptions made to permit the computations. The fact that the physical conditions are atypical with regard to everyday environments does not necessarily degrade the quality or the usefulness of the answers to the questions he is asking.

He may also select a particular method as his assumption for computing W and again there is no reason why this method has to be typical of those encountered in everyday environments. It need only be a reasonable assumption for the class of experiments being performed. If it is necessary to train observers to behave in a particular way in experiments in order to get at the answers to particular experimental questions, then this training is permissible. That this behavior is atypical to everyday environments is not a valid criticism of the assumption made for this purpose.

The block diagram of Figure 2 illustrates the experimental schema of the second purpose. The diagram, from Tanner and Birdsall (Ref. 5) describes a procedure for using optimum and non-optimum models to draw conclusions about the human observer's hearing mechanism. The non-optimum models are degraded from the optimum by introducing statistical uncertainty into the form of the signal as it is transmitted.

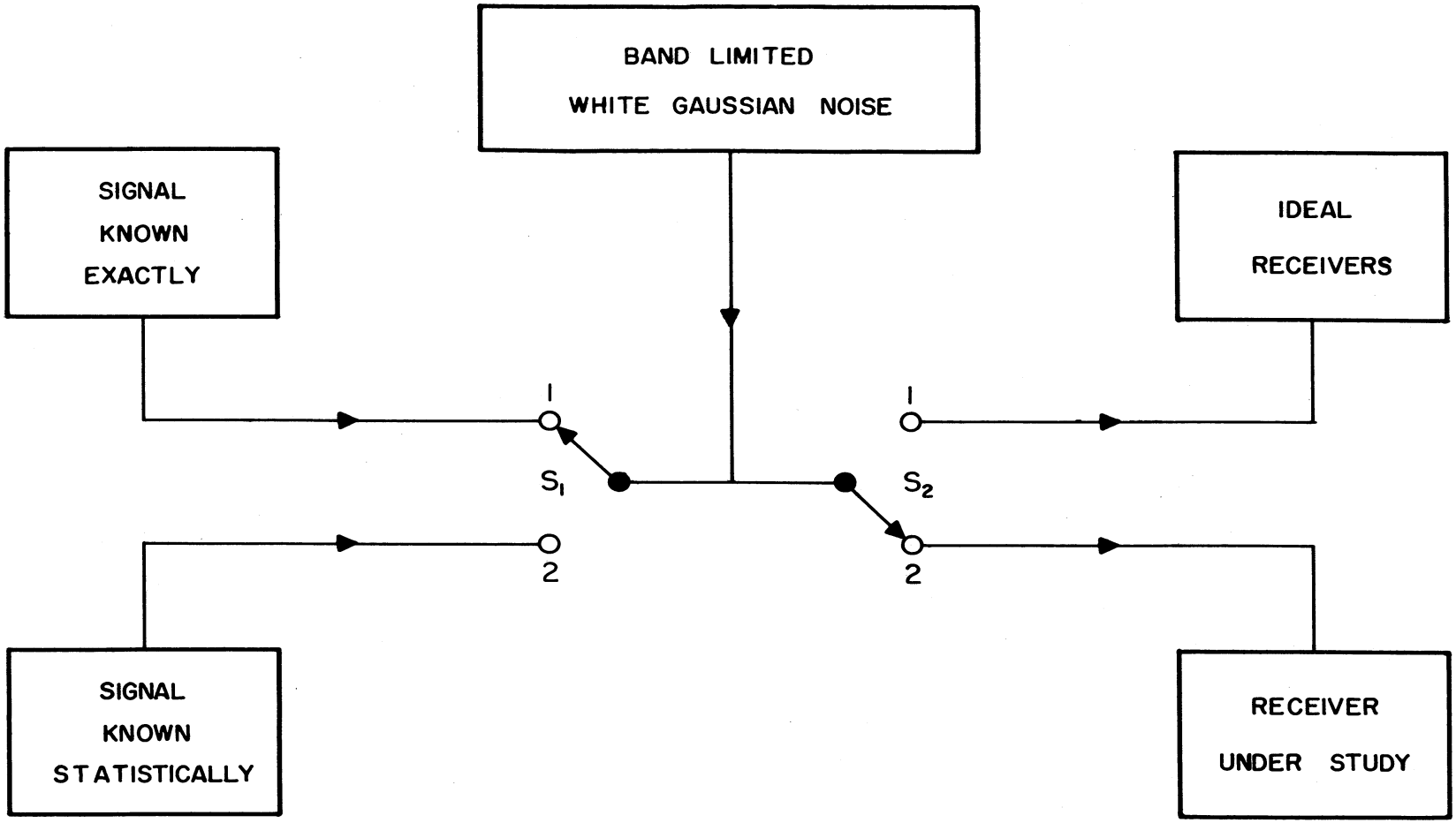

FIG. 2

COMPOSITE BLOCK DIAGRAM OF CHANNELS FOR PSYCHOPHYSICAL EXPERIMENT.

Under matched conditions (using the same energy and yielding the same performance), the human observer is said to introduce the same degree of uncertainty into a channel using an ideal transmitter as the non-optimum transmitter does in the channel which has an ideal receiver. Conclusions involving this type of reasoning require measures of efficiency and, in order to compute efficiencies, it is necessary to have a standard of reference. Assumptions are required to permit this computation. These assumptions should be of such a nature that either they, or an equivalent set, can be reasonably approximated by physical conditions in the laboratory.

The purpose of this discussion is to select and justify the specializing assumptions necessary to bring the theory of signal detectability into agreement with the laboratory situation so that this theory can be used as a tool to aid in interpreting data to answer questions about the performance of the hearing mechanism. It is hoped that the answers to these questions will form a basis for developing a descriptive model of the hearing mechanism. This description may then be incorporated into a more comprehensive model describing or explaining behavior in everyday environments. For the present, however, attention will be restricted to the use of the model as an interpretive tool.

3. THE ASSUMPTION OF EXPECTED VALUE MAXIMIZATION

One of the first problems encountered in developing techniques for using the theory of signal detectability as a tool for psychoacoustic studies was the determination of the observers' control of the cut-off

value (the number W). As his false-alarm rate varied, did it look as if he was guessing or behaving as if he was varying a cut-off value?

To study this problem it was desirable to trace out ROC curves. The most logical procedure appeared to be to give the observer a basis for computing W in his experiments. An examination of the various criteria reviewed by Birdsall (Ref. 4) suggested that the expected value maximization would be one of the easiest to explain to an observer. The point which must be kept in mind is that the purpose of introducing a function which is to be maximized is to permit the necessary manipulation for collecting data permitting an ROC curve to be traced.

The number W is referred to as β for this special case to distinguish it from the general weighting function. In order to maximize the expected value, the input is said to consist of signal plus noise whenever the following condition is satisfied:

$$\ell(x) \geq \beta = \frac{P(N)}{P(SN)} \quad \frac{V_{N \cdot CA} + K_{N \cdot A}}{V_{SN \cdot A} + K_{SN \cdot CA}}$$

where P(N) and P(SN) are the a priori probabilities of noise alone and signal plus noise respectively; V is a gain and K a cost or loss. The subscripts N and SN refer to the truth of the hypotheses noise and signal plus noise; and the subscripts A and CA to the event falling into the criterion A or falling into the criterion CA, the complement of A leading to the acceptance of the hypothesis noise alone.

By changing either the a priori probabilities or the values and costs, it is possible to shift the observer's position on his ROC

curve, provided that the observer is attempting to maximize the expected value of the outcome of his performance. The first problem is to establish for an observer an environment in which it is desirable to maximize expected value. Then, if it can be established through experiment that the observer is reacting to the changes in a consistent way, even if he is not actually maximizing the expected value of the outcome, the purpose is achieved.

The first problem is to determine those conditions which make the expected value of the outcome a desirable function to maximize. First of all, the expected value of a game is realized when both the player and the opponent have infinite resources. If the resources of the player are finite but sufficiently large to permit a great enough number of plays, such that the law of large numbers is applicable and, if the expected value of the game is greater than zero for the observer, then the expected value is a desirable function to optimize.

These conditions can be established in the following way. An observer is paid a basic rate, say one dollar an hour during which he makes 400 observations. For each correct decision he receives one tenth of a cent, for each incorrect decision he loses one tenth of a cent. Thus, his maximum loss in a one hour period is 40 cents and his basic pay rate is sufficient to keep him in the game infinitely long. Since the task always involves detecting a signal with some energy, the game can always be made to be favorable to the observer.

Since it is possible to satisfy the conditions, it remains only to test experimentally whether or not the observer, either through training or through an already developed ability, can react to changes

in a priori probabilities or to changes in the pay-off matrix. Some of the early experiments support the contention that a human observer can do this. Data reported by Smith and Wilson (Ref. 6), Munson and Karlin (Ref. 7), Tanner and Swets (Ref. 8), and Tanner, Swets and Green (Ref. 9) all support the contention that the behavior of the human observer can be manipulated by changing the expected value of the game.

The assumption that the expected value is maximized in a carefully designed laboratory experiment is used because it appears to accomplish the purpose for which it is intended. It is not important whether people seem to attempt to maximize expected value in everyday environments. The linearity of the utility of money is not involved. The fact that the pay per correct decision is an apparently negligible amount is also unimportant. There are thousands of observations involved and the observer can increase his pay rate by as much as twenty cents an hour. The assumption is adopted because it appears to work in the laboratory experiment.

4. THE FOURIER SERIES BAND-LIMITED ASSUMPTION

The reason for wanting to make an assumption permitting computation of the distributions of the likelihood ratio, or some monotonic function of the likelihood ratio, is to be able to state the performance of an ideal receiver under conditions which are at least very closely related to those of the laboratory experiment. This performance depends on the separation between the two hypotheses: that for signal plus noise and that for noise alone.

In undertaking this problem, Peterson, Birdsall and Fox assumed a sampling theorem which permitted them to apply discrete statistics to the analysis. Slepian (Ref. 10) argued that their results depended on the particular form of sampling theorem they employed and showed that a different assumption would lead to drastically different results. David and Matthews (Ref. 11) computed the separation between two distributions as a function of the number of sampling points, using an assumption essentially the same as Slepian's, showing that, as the number of sampling points increases beyond that of Peterson, Birdsall, and Fox, the separation increases slightly.

It is the purpose of this section to analyze and evaluate these assumptions with reference to their accuracy and usefulness in psychoacoustic experiments in which the model is employed as a tool to assist in the interpretation of the data.

Peterson, Birdsall, and Fox analyzed the case of a signal specified exactly embedded in a noise which was assumed to be additive, Fourier series band-limited white Gaussian noise. In their analysis, both the noise waveforms and the signal plus noise waveforms were assumed to be band-limited in the same way.

In this analysis both the signal and the noise are defined as voltage waveforms which can be defined over an open interval of time, 0 to T. According to the sampling theorem, an input waveform consisting either of noise alone, $n(t)$, or signal plus noise, $s(t) + n(t)$, can be described precisely by $2WT$ amplitude values where W is the bandwidth of the signal plus noise waveforms or of the noise waveforms. It is to be noted that the description of the waveform applies only to the interval 0 to T. Nothing is said about its form outside of the interval or about its method of generation.

The 2WT values can be taken in a number of ways. However, Peterson, Birdsall, and Fox work with 2WT equally spaced independent values, permitting the application of the following statistical theorems:

1) The likelihood ratio of a sequence of independent measures, taken under the same conditions, is the product of likelihood ratios of the individual measures in the sequence.

2) Since they work with logarithms of likelihood ratios, the theorems applying to the means and variances of the distributions of sums of random variables.

The analysis shows that the distributions of the natural logarithms of the likelihood ratio under the condition noise alone and under the condition signal plus noise are both normal having equal variances. The separation between the means, divided by the standard deviation, is $(\frac{2E}{N_o})^{1/2}$, where E is the signal energy and N_o is the noise power per unit bandwidth.

Fox (Ref. 12) points out that "under the restrictions: 1) the populations N and SN are finite dimensional; and 2) that the functions of time in these populations be (real) analytic, it is possible to prove that sampling plans utilizing arbitrarily small sample intervals can be found, all of which yield the same error probabilities or ROC curves." He further points out that the proof depends on the assumption of errorless measurements. He says "It is not at all uncommon that the assumption that errorless measurements are possible should lead to physically ridiculous conclusions."

Pushed to its limit, a proof of this nature permits one to incorporate information based on energy which exists outside of the observation interval. Consider for example the conditions outlined as follows. An observation interval, 0 to T, is defined. The populations N and SN are defined over the interval 0' to T' which completely includes the interval 0 to T. According to the proof, a sampling plan on the interval 0 to T can be devised which will define precisely the waveform on the interval 0' to T', and presumably the results would be the same as if the observation had been over the interval 0' to T'. If the sampling theorem employed by Peterson, Birdsall, and Fox applies to an observation over the interval 0' to T', then it would be possible to increase information by using more than 2WT measures over the interval 0 to T. In fact, one would expect that at least 2WT' measures would be required. The separation between the hypotheses would be greater than $(\frac{2E}{N_o})^{1/2}$ if the signal energy is calculated as that energy contained within the interval 0 to T. If it is calculated on the interval 0' to T', then the separation would again be exactly $(\frac{2E}{N_o})^{1/2}$.

Whether or not the detectability of a signal can be expressed by the number $(\frac{2E}{N_o})^{1/2}$ depends on what assumptions are made with reference to the way in which one can use knowledge of events happening outside of the observation interval. If the assumptions apply only to the observation interval, then $(\frac{2E}{N_o})^{1/2}$ appears to be an acceptable quantity. Results which predict greater detectability depend upon using the observations within the interval to describe the waveform outside of the interval. In the study of Peterson, Birdsall, and Fox, there was no basis for using information outside of the interval. The populations N and SN are defined only over the open interval 0 to T. While this

analysis can be extended, the applicability of the extension to conditions outside of the interval can be seriously questioned.

5. THE FOURIER TRANSFORM BAND-LIMITED ASSUMPTION

Slepian (Ref. 10) showed that, for the case of a noise signal in a noise background, the results of an analysis based on sampling theorems depend critically on the particular form of the assumptions employed. Particularly important are the relations between the way the "signal-noise" and the "noise-noise" fall off at the skirts of the band. Taking one's assumptions literally, and assuming perfect measurement, it is possible to establish a set of conditions which lead to perfect detection. He argues that attention should be redirected to the more interesting cases where perfect measurement is not assumed, and only finite detectability is possible.

David and Matthews (Ref. 11) performed some computations to show how detection increases as the number of measures is increased. In making these computations, they took literally the assumptions which extend the observation beyond the interval 0 to T. In the limit, the equations upon which David and Matthews computations were based can be shown to predict perfect detectability. Even though this is the case, their results indicate that detectability increases only slightly as the number of sampling points is increased beyond 2WT. From their graphs, one gets the impression that the separation between the two distributions is approaching a finite asymptote, only slightly greater than that predicted by the finite sampling plan employed by Peterson, Birdsall, and Fox. Certainly, if the separation is to become infinite,

it is approaching infinity very slowly.

The fact that these results do not appear to be consistent with the predictions based on carrying their assumptions to the limit might be attributed to the fact that the computer (IBM-704) which they employed has only a finite capacity to carry out errorless computations. The fact that error, no matter how small, was introduced into the computations may be equivalent to introducing error, no matter how small, into the measurements. If so, then David and Matthews have succeeded in constructing an excellent demonstration of Slepian's point.

The Fourier transform analysis describes the waveform as if it has always been in existence and always will continue to exist. The waveform is completely deterministic. Given a complete description of the waveform over any small interval, the description can be extended from $-\infty$ to $+\infty$. If the signal is present, one description applies, if noise alone is present, another applies. The noise is completely predictable. Pushing the assumption to the limit, the signal becomes detectable with certainty.

Extending the assumptions, and given absolutely precise measurements, it should then be possible to design an experiment in which, as the measures are increased in number, a description will evolve which converges on one of two descriptions. To be convincing, however, it will have to be a real experiment, it must be different from a set of computations determining the consequences of a carefully stated set of assumptions. Using the assumptions to construct an analytic tool, or for the basis of the development of a model is one thing; to claim that in the limit they apply to a set of real conditions (either laboratory or everyday life) is quite different. This claim must stand or

fall with the outcome of carefully conducted experiments.

6. COMPARISON OF THE SAMPLING THEOREMS

In comparing the sampling theorems, it is necessary to keep carefully in mind that Fourier analysis is a mathematical tool used to describe voltage waveforms. The mathematics incorporates precisely stated assumptions involving band limitations and processes extending over infinitely long times. To use Fourier analyses, it is not necessary to assert that the generating process agrees in detail with the assumptions; it is necessary to argue only that the analysis leads to results which are consistent with the purposes for which the analysis is being performed.

In order to compare the two sampling theorems, let it be accepted that neither Fourier series nor a Fourier transform band-limitation assumption can be precisely satisfied in the physical world. Even if a precise match could be achieved, it would be impossible to demonstrate the match with certainty based on experimental data. Thus, the first problem in the evaluation is to determine what happens to the analysis in the two cases if the assumptions are only approximated.

Both the analysis based on the Fourier transform assumption and that based on the Fourier series assumption assume that the measures are exactly precise and that they are taken at exactly precise points in time. Suppose that arbitrarily small error is permitted in the location of the point in time, what happens to the results based on the two

analyses? In the case where the Fourier transform band-limited assumption is incorporated, it is argued that the separation between the two hypotheses no longer goes to infinity in the limit. It appears to approach a finite asymptote, perhaps that suggested by the curves of David and Matthews. It is also possible that this asymptote applies to any finite set of measures and that certain detection can only be achieved based on an infinite number of measures.

The analysis based on the Fourier series assumption describes the separation of the two distributions based on a finite number (2WT) of exactly precise measures. If arbitrarily small error is permitted in these measures, the result differs only slightly. As the error becomes smaller, the separation between the two distributions approaches asymptotically the finite value predicted by the analysis based on the exactly precise measurements.

It appears that the necessity of introducing approximations, in attempting to match physically the assumptions of the analyses, does not do particular violence to the Fourier series analysis, while it does to the Fourier transform analysis. Because the Fourier series assumption appears to be less sensitive to approximation, it seems to be the more desirable of the two with which to work, particularly for the purpose of developing an analytic tool to be used as an aid in interpreting experimental data. In this case, one does not have to accept the conditions as they exist; one can choose his conditions to suit his purposes. How closely can one hope to create a set of conditions agreeing with the assumptions of the Fourier series analysis?

7. THE PHYSICAL APPROXIMATION

The problem is to create a set of physical conditions to which an analysis based on a precisely stated set of assumptions applies. If this can be achieved, it is not necessary to have the physical conditions agree closely with the assumptions underlying the analysis. The success in achieving a laboratory condition for which the theory is useful depends on the extent to which the analysis applies.

In order to evaluate the extent to which the analysis applies to the physical conditions, it is worth recalling to mind that the purpose of the application is to determine performance upper bounds which can be used as references against which the subjects' performance can be compared. These comparisons are valuable in interpreting the experimental results.

It is also worthwhile that the physical conditions are intended to be such that a particular analysis applies. If one chooses a different set of physical conditions, then this analysis might not apply. If one chooses a different type of analysis, then the physical conditions described below may not be satisfactory.

In this case, the task is to find a set of physical conditions to which an analysis with a known solution applies. One also might attempt to study a set of physical conditions to determine what types of assumptions would be required of an analysis applying to the particular set of physical conditions. As long as satisfactory agreement can be established, it makes little difference from which method of

approach it was evolved. However, it might be better to start with a problem which has been solved and attempt to manipulate the physical conditions than to start at the other end and try to manipulate the mathematics. A suitable mathematics may not exist for the set of physical conditions selected.

The analysis to be applied is that based on the Fourier series band-limited assumption. The assumptions of the analysis are:

1) That the waveforms arising from signal plus noise and the waveforms arising from noise alone are band-limited in the same way.

2) That over the open interval, 0 to T, both the signal plus noise waveforms and the noise alone waveforms conform to the conditions of Fourier series band-limitation.

3) That the noise is white Gaussian noise over the frequency band.

Both the second and third of the assumptions appear impossible to satisfy physically. The first probably can be satisfied.

A physical arrangement which conforms to the conditions necessary for the analysis to apply is illustrated in Figure 3. The noise and the signal are generated by a General Radio noise generator and a Hewlett-Packard audio oscillator respectively. The signal is gated, yielding a pulse which is a close approximation to a segment of a sine wave. The spectrum of the noise is approximately white from 50 cycles to 30 or 40 thousand cycles, and, except for the absence of rare infinite peaks, its amplitude distribution is nearly Gaussian. Since the signal is gated, almost all of its energy is concentrated in the inter-

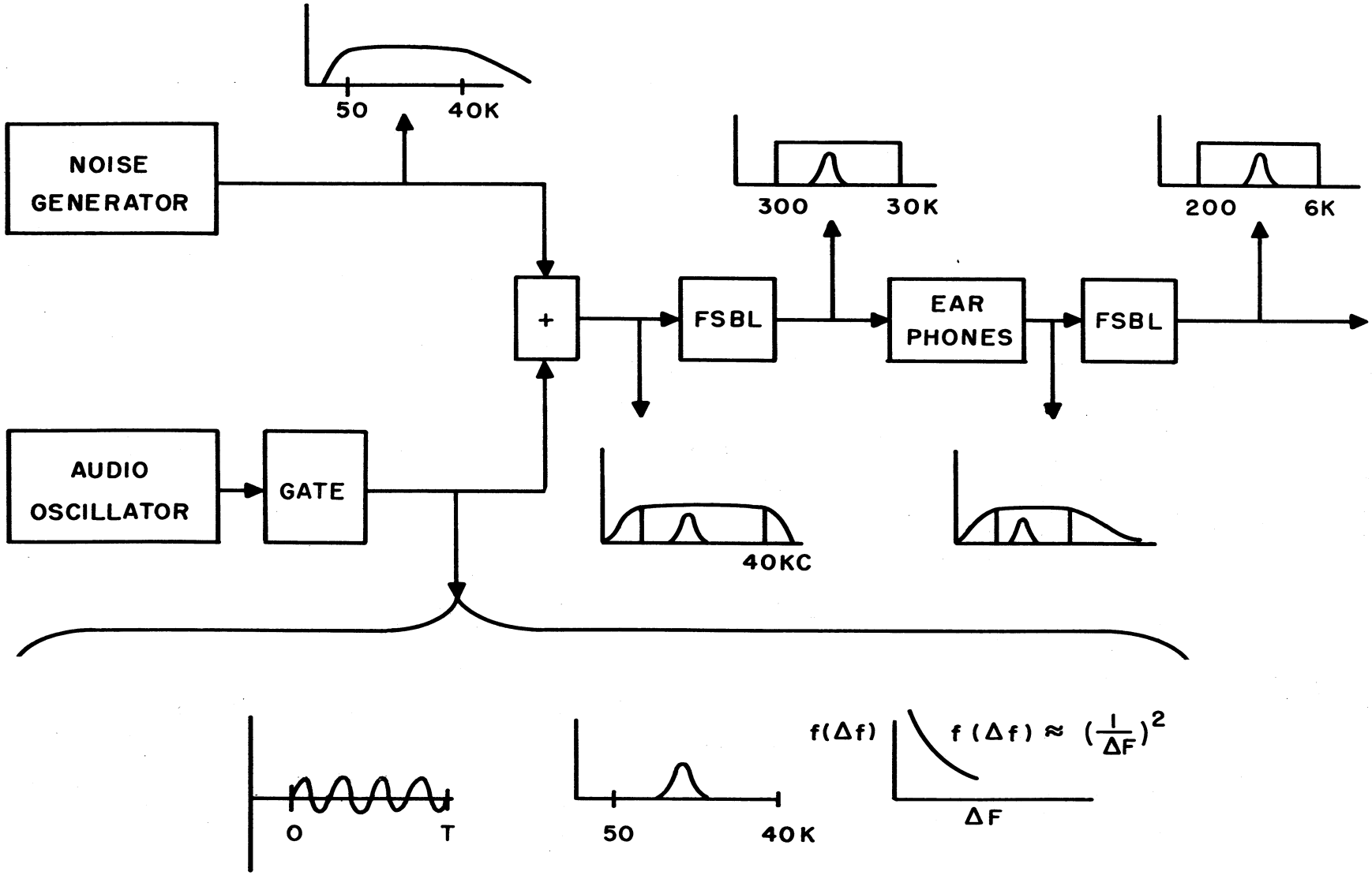

FIG. 3 BLOCK DIAGRAM ILLUSTRATING A LABORATORY ARRANGEMENT CONSISTENT WITH THE USE OF THE 2 WT SAMPLING PLAN

val, 0 to T. A major concentration of its energy is found at its center frequency. As one goes to either side, the frequency drops as $(\frac{1}{\Delta f})^2$. Thus, there can be only a negligible portion of the energy outside of that portion of the noise which is white. Any signal energy outside of this band could be only infinitesimally detectable because of the noise of the error of measurement.

The gated signal is then added to the noise, at which point the spectrum of the signal is well contained within the spectrum of the noise. Thus, both signal plus noise and noise alone appear to be band-limited in the same way. Assume now that both are placed through a Fourier series band-limiter, an unrealizable piece of equipment. This band-limiter is one whose output actually satisfies the assumptions of Peterson, Birdsall, and Fox during the open interval, 0 to T, and whose output is unanalyzable outside of the interval. An analysis of the detectability at the output of that filter is an analysis which considers all of the information at the input, with the exception only of that small portion which lies outside the filter transmission band. This is nearly zero and can be ignored.

If the output of the band-limiter is now fed to the earphones, the output of the earphones is essentially the same as if the band-limiter did not exist. The earphones are known to have a flat response approximately 200 cycles to 6000 cycles. Again, a Fourier series band-limiter creates the conditions for which the analysis applies and again removes almost no information from the waveform at the output of the adder. At each point the analysis leads to the same result, barring only minute differences. The earlier the analysis is performed, the more likely it is to be an over-estimate of the performance.

It may be possible to describe the waveform within the interval, 0 to T, with other analyses. For example, it may be possible to describe a set of generators which could generate this waveform if they had been operating forever. But one should not forget that this waveform was more likely to have been generated by equipment which is turned on in the morning and turned off at night. The fact that waveforms can be analyzed as if they are made up of a set of sine wave components leads to a useful mathematical tool. It does not limit the method of generating waveforms to the process of combining sine waves.

If one wishes to attribute to the physical waveforms the properties he has assumed for the purpose of analysis, then, as Slepian shows, he is likely to arrive at conclusions which are obviously ridiculous from a realistic standpoint.

8. SUMMARY AND CONCLUSIONS

The theory of signal detectability has been examined from the standpoint of determining a set of satisfactory assumptions for the purpose of developing an interpretive tool for use in psychophysical experiments. It was concluded that the assumption that the observer attempts to maximize the expected value of the outcome of the experiment is satisfactory for this purpose, and that a set of physical conditions can be established which justify a computation of the detectability of a signal in noise based on a finite sampling plan involving 2WT amplitude values over the open observation interval 0 to T.

REFERENCES

1. J. F. Siegert, Chapter 7 in J. L. Lawson and G. E. Uhlenbeck, Threshold Signals. New York; McGraw-Hill, 1950.
2. W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The Theory of Signal Detectability," Trans. Professional Group on Information Theory, Inst. Radio Engrs., 1954, PGIT-4, 171-212.
3. D. Van Meter and D. Middleton, "Modern Statistical Approaches to Reception in Communication Theory," Proc. Professional Group on Information Theory, Inst. Radio Engrs., 1954, PGIT-4, 119-145.
4. T. G. Birdsall, "The Theory of Signal Detectability," in H. Quastler (ed.) Information Theory in Psychology. Glencoe, Ill.; Free Press, 1954.
5. W. P. Tanner, Jr. and T. G. Birdsall, "Definitions of d' and η as Psychophysical Measures," J. Acoust. Soc. Amer., 1958, 30, 922-928.
6. M. Smith and Edna A. Wilson, "A Model for the Auditory Threshold and its Application to the Multiple Observer," Psychol. Monogr., 1953, 67, No. 9.
7. W. A. Munson and J. E. Karlin, "The Measurement of the Human Channel Transmission Characteristics," J. Acoust. Soc. Amer., 1956, 26, 542-553.
8. W. P. Tanner, Jr. and J. A. Swets, "A Decision-Making Theory of Visual Detection," Psych. Rev., 1954, 61, 401-409.
9. W. P. Tanner, Jr., J. A. Swets, and D. M. Green, "Some General Properties of the Hearing Mechanism," Tech. Rept. No. 30, Electronic Defense Group, The University of Michigan, March, 1956.

10. D. Slepian, "Some Comments on the Detection of Gaussian Signals in Gaussian Noise," Trans. Information Theory, Inst. Radio Engrs., IT-4, 1959, 65-68.

11. E. E. David and M. V. Matthews (Unpublished Memorandum)

12. W. C. Fox, "Signal Detectability: A Unified Description of Statistical Methods Employing Fixed and Sequential Decision Processes," Tech. Rept. No. 19, Electronic Defense Group, The University of Michigan, 1953.

Detection of a Signal Specified Exactly With a Noisy Stored Reference Signal[1]

T. G. Birdsall, The University of Michigan Research Institute

I. FOREWORD

Wilson P. Tanner, Jr.

In this paper Mr. Birdsall has undertaken the study of a problem the solution of which promises to extend the usefulness of the theory of signal detectability as a tool in psychophysical investigations. The problem is that of the behavior of a receiver which has a noisy memory: it "knows" the signal it is looking for only approximately, not precisely.

Up to the present time, mathematical studies of the detectability of signals have assumed perfect memories for the receivers. These studies consider cases wherein the receiver "acts" as if it could reproduce a template of a signal specified exactly at the transmitter. This template agrees with the transmitted signal in every detail. The waveform recorded on the template, its amplitude, starting time, and phase all agree precisely with the signal. The receiver can then compute a crosscorrelation between the template and the input waveform, which may consist of either noise alone or signal plus noise. This crosscorrelation constitutes the datum necessary for making the optimum decision regarding that waveform: either it contains a signal or it does not.

---

[1] This paper is based upon Technical Report No. 93, issued by the Electronic Defense Group of The University of Michigan in 1959. The research was supported in whole or in part by the United States Air Force under Contract No. AF49(638)-369, monitored by the AF Office of Scientific Research of the Air Research and Development Command.

It should be obvious to all familiar with statistics that, if the template were an inexact copy, or if it had noise added to it, a lower correlation would result on the average, and performance of the receiver would suffer. In this case both variables have an error component, while in the cases previously studied, one of the variables, the template recording, is noiseless.

One might say that Mr. Birdsall is investigating the case of the noisy template. If one knows that the receiver template is noisy, how should the inputs be processed? Should one compute crosscorrelations, or should a different analysis be made? It turns out that if the template is not very noisy, the receiver should compute the crosscorrelation, while if it is very noisy it should integrate the energy at the input.

While the study has implications in many fields, it is only the applications to psychophysics that are discussed here. It is obvious that human beings do not have perfect memories. They all have noisy templates. Mr. Birdsall's paper is directed toward the understanding of receivers with noisy templates, and since the human observer falls within that class it is hoped that the results obtained will help account for the form human data assume.

One important conclusion evolving from this study is that memory noise cannot be treated as noise added to the input. It has a nonlinear effect which must be taken into account in the interpretation of psychophysical data.

It is true that the scope of the specific example treated in the paper is limited. Even so, the resulting curves will aid in interpreting data and in achieving a better understanding of the problems encountered in psychophysical experiments.

II. INTRODUCTION

This analysis deals with the detection of a signal specified exactly (SSE) but not known exactly by the receiver. From the presentation standpoint the problem is identical to the elementary signal known exactly (SKE), but from the receiver standpoint the expected signal is distributed. The specific case analyzed is that wherein the receiver has a stored signal, $s(t)$ which differs from the true signal, $S(t)$, by band-limited white Gaussian noise. The reception is similarly corrupted by band-limited white Gaussian noise, which is independent of the internal storage noise. The analysis of the optimum receiver is made in Section 3. Evaluation of the receiver is made in Section 4 for the special case of $2WT = 1$, and in Section 5 the crosscorrelation receiver is evaluated and compared to the ideal for this special case. Section 8 is a discussion of the implications of the results.

III. ANALYSIS

This analysis is based on the theory of signal detectability. (Refs. 1, 2) The observation is of finite time duration, T sec, and all signals and noise waveforms have a finite series band-limit, W cps, on that time interval. The external noise is white and Gaussian with noise power per cycle, N_o, and the internal storage noise is also white and Gaussian with noise power per cycle, $\lambda N_o = n_o$.

The first step in the analysis is to derive the likelihood ratio. The receiver (or observer) must base its response on the total input, which is both the observation, $x(t)$, and the stored signal, $s(t)$.

The receiver has error-free storage of the parameters, N_o, W, T, and λ and on each observation can perform error-free operations on the specific x(t) and s(t) stored. The receiver does not have access to the true signal S(t) other than through the above listed items.

The probability density function of this input is found as follows. The noise density function for the observation is[2]

$$f_N[x(t)] = \left(\frac{1}{2\pi N_o W}\right)^{WT} \exp\left[-\frac{1}{N_o}\int_o^T x^2(t)dt\right], \qquad (1)$$

and in similar fashion, the density function for the observation when signal is present is

$$f_{SN}[x(t)] = \left(\frac{1}{2\pi N_o W}\right)^{WT} \exp\left\{-\frac{1}{N_o}\int_o^T [x(t) - S(t)]^2\, dt\right\}. \qquad (2)$$

The density functions for the storage signal, s(t), are the same for both hypotheses,

$$f[s(t)] = \left(\frac{1}{2\pi\lambda N_o W}\right)^{WT} \exp\left\{-\frac{1}{N_o\lambda}\int_o^T [s(t) - S(t)]^2\, dt\right\}. \qquad (3)$$

Now from the receiver's standpoint Eq. (3) specifies the distribution of the true signal, S(t), about the stored signal, s(t). If S(t) were known, the likelihood ratio would be the ratio of Eq. (2) to Eq. (1). In the case of a distributed signal this ratio must be averaged with respect to the distribution of the signal[3]

$$\ell[x(t)] = \int_S \exp\left[-\frac{1}{N_o}\int_o^T S^2(t)\, dt\right] \exp\left[\frac{2}{N_o}\int_o^T x(t)S(t)dt\right] dP(S) \qquad (4)$$

---

[2] See Peterson, Birdsall and Fox, Eq. (48)

[3] See Peterson, Birdsall and Fox, Eq. (56)

where, of course, for a fixed stored signal, s(t),

$$dP(S) = f[s(t)]dS. \tag{5}$$

It is shown in the appendix that the evaluation of Eq. (4), simplified, implies that the likelihood ratio is strictly monotone increasing with the quadratic form

$$\ell[x(t)] \text{ -smi- } \int_0^T x^2(t) + \frac{2}{\lambda} x(t)s(t) - \frac{1}{\lambda} s^2(t)dt. \tag{6}$$

Thus the receiver which computes the above quadratic will be an ideal receiver under the conditions of the problem. Several equivalent forms are given in Eqs. (7) and (8).

$$\ell[d(t) \text{ -smi- } \int_0^T \left[x(t) + \frac{1}{\lambda} s(t)\right]^2 dt - \frac{1+\lambda}{\lambda^2} \int_0^T [s(t)]^2 dt \tag{7}$$

$$\ell[x(t)] \text{ -smi- } \int_0^T \left[\lambda x(t) + \frac{\lambda s(t)}{\sqrt{1+\lambda} - 1}\right] \left[x(t) - \frac{s(t)}{\sqrt{1+\lambda} + 1}\right] dt \tag{8}$$

The receivers (all optimum) based on Eqs. (6), (7), and (8) are given in Figs. 1, 2, and 3.

IV. EVALUATION OF IDEAL, 2WT = 1

The evaluation of the ideal should be carried out with an ROC curve analysis, i.e., a comparison of the distributions of either the likelihood ratio or any monotone function of the likelihood ratio, under the two conditions of noise alone and signal plus noise.

To date, the author has been unable to obtain these distributions in closed form. In all of the detection literature the performance of optimum receivers has been obtained in closed form for only a handful of cases. In

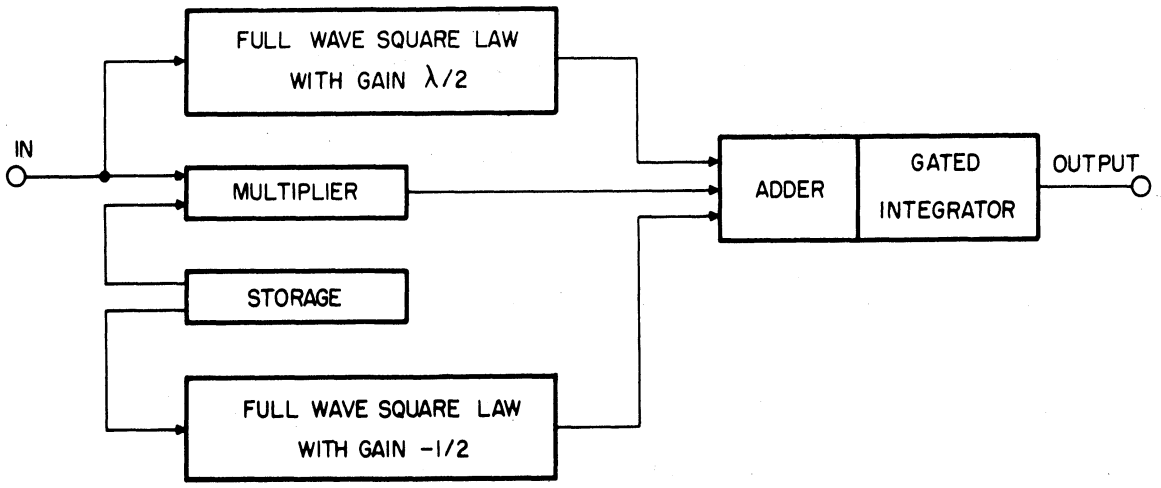

FIG. 1 RECEIVER OF EQUATION 6

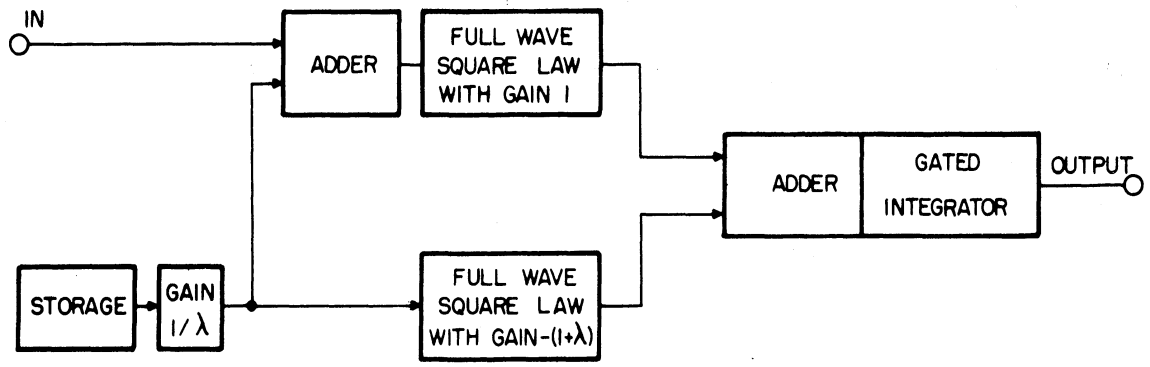

FIG. 2 RECEIVER OF EQUATION 7

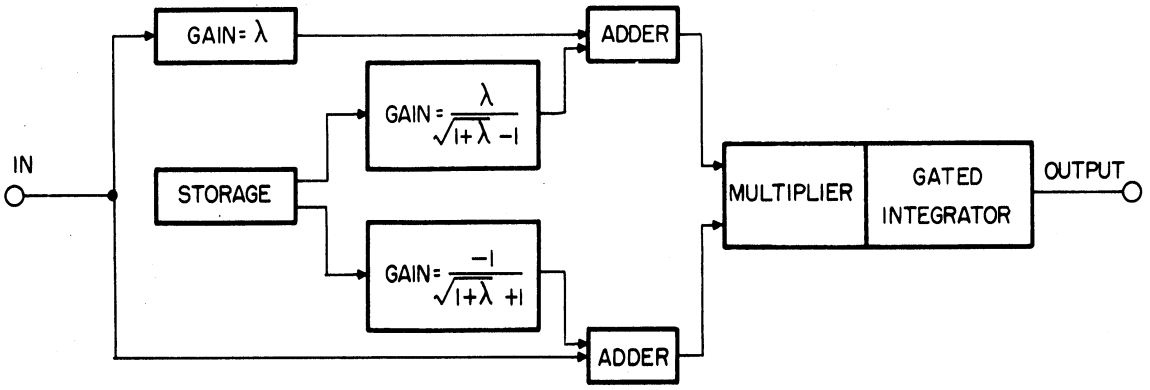

FIG 3a

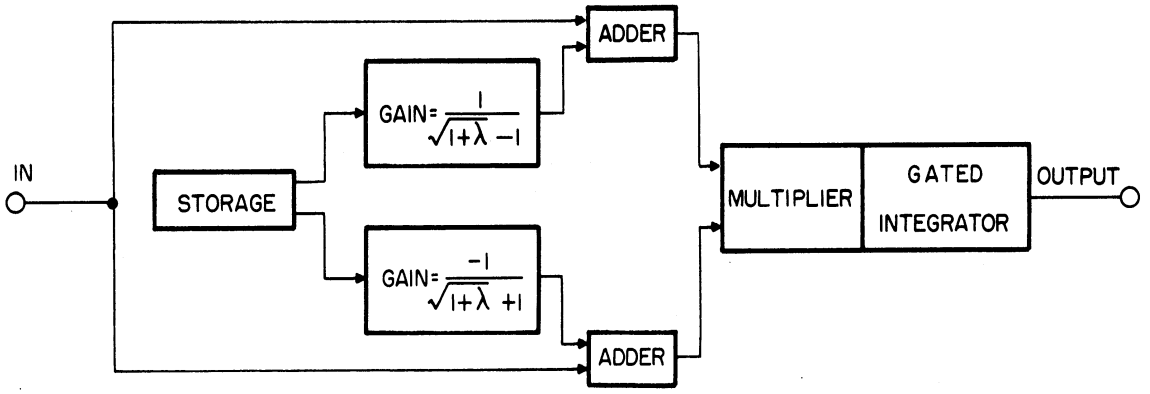

FIG. 3b

FIG. 3 RECEIVER OF EQUATION 8

(a) LOW OR MEDIUM INTERNAL NOISE VERSION

(b) HIGH INTERNAL NOISE VERSION

FA2

the others, approximations or analogs have been used to obtain numerical results. Such techniques could also be applied to this optimum receiver. A less direct approach was taken for this problem; namely, the performance in a two-alternative-choice-in-time has been determined and is given in this section. It should be mentioned that although such an evaluation is not as complete as an ROC analysis, it usually agrees with the results of an ROC analysis in the medium probability range (1% to 50% false-alarm range).

The evaluation assumed that the stored signal is the same for two observation intervals, $x_1(t)$ and $x_2(t)$, one which is due to noise alone and one which is due to signal plus noise. The receiver operation is to compare the outputs corresponding to the observations and indicate the interval with the larger output. The evaluation determines the signal strength necessary to obtain performance equivalent to that which would have been obtained had the signal been known exactly.

Equation (7) is the easiest relation to evaluate in this situation. Because it is assumed that the stored signal, $s(t)$, is the same for both observations, one concludes that

$$\ell[x(t)] > 2[x_2(t)] \Longleftrightarrow \int_0^T [x_1(t) + \frac{1}{\lambda} s(t)]^2 \, dt > \int_0^T [x_2(t) + \frac{1}{\lambda} s(t)]^2 \, dt. \quad (9)$$

It was assumed that $2WT = 1$, and hence each function of time can be represented by a single sample, and the integrals will be equal to $1/2W$ times the integrand at the sampled points. Hence, if the samples are denoted by dropping the "t",

$$\ell[x_1(t)] > \ell[x_2(t)] \Longleftrightarrow (x_1 + \frac{1}{\lambda} s)^2 > (x_2 + \frac{1}{\lambda} s)^2 \quad (10)$$

$$\Leftrightarrow (x_1 + \frac{1}{\lambda} s)^2 - (x_2 + \frac{1}{\lambda} s)^2 > 0 \tag{11}$$

$$\Leftrightarrow (x_1 - x_2)(x_1 + x_2 + \frac{2}{\lambda} s) > 0 \tag{12}$$

$$\Leftrightarrow \begin{cases} (x_1 - x_2) > 0 \text{ and } (x_1 + x_2 + \frac{2}{\lambda} s) > 0 \\ \text{or} \\ (x_1 - x_2) < 0 \text{ and } (x_1 + x_2 + \frac{2}{\lambda} s) < 0. \end{cases} \tag{13}$$

Now if $x_1(t)$ is due to signal plus noise and $x_2(t)$ due to noise alone, the probability that the above inequalities hold is the probability of a "correct decision", i.e., in this case, that the output of observation number one is greater than that of observation number two. It is obvious that the situation is completely symmetric and that the probability of a correct decision is the same as if the hypothesis had been reversed.

The variables, $(x_1 - x_2)$ and $(x_1 + x_2 + \frac{2}{\lambda} s)$, are independent Gaussian variables. The means and variances are as follows:

$$\begin{aligned} &\text{Normalize so that } \sigma^2(x_i) = 1. \\ &\text{Then} \qquad \sigma^2(s) = \lambda, \\ &\text{and} \qquad \mu(x_1) = S \\ &\qquad\qquad \mu(x_2) = 0 \\ &\qquad\qquad \mu(s) = S. \end{aligned} \tag{14}$$

$$\text{Hence } \mu(x_1 - x_2) = S - 0 = S$$

$$\sigma(x_1 - x_2) = \sqrt{1 + 1} = \sqrt{2} \tag{15}$$

so that

$$\Pr(x_1 - x_2 > 0) = \Phi\left(\frac{S}{\sqrt{2}}\right) \tag{16}$$

and

$$\Pr(x_1 - x_2 < 0) = \Phi\left(-\frac{S}{\sqrt{2}}\right),$$

where Φ is the normal, or Gaussian, distribution function.

Similarly,

$$\mu(x_1 + x_2 + \frac{2}{\lambda} s) = S + 0 + \frac{2}{\lambda} S = S \frac{\lambda+2}{\lambda}$$

$$\sigma(x_1 + x_2 + \frac{2}{\lambda} s) = \sqrt{1 + 1 + \frac{4}{\lambda^2}} = \sqrt{\frac{2(\lambda+2)}{\lambda}} \tag{17}$$

so that

$$\frac{\mu}{\sigma} = S \sqrt{\frac{\lambda+2}{2\lambda}} \ , \tag{18}$$

$$\Pr(x_1 + x_2 + \frac{2}{\lambda} s > 0) = \Phi\left(S \sqrt{\frac{\lambda+2}{2\lambda}}\right) , \tag{19}$$

and

$$\Pr(x_1 + x_2 + \frac{2}{\lambda} s < 0) = \Phi\left(-S \sqrt{\frac{\lambda+2}{2\lambda}}\right).$$

Combining Eqs. (16) and (19) as indicated in Eq. (13),

$$\Pr(\text{correct}) = \Phi\left(\frac{S}{\sqrt{2}}\right)\Phi\left(S\sqrt{\frac{\lambda+2}{2\lambda}}\right) + \Phi\left(-\frac{S}{\sqrt{2}}\right)\Phi\left(-S\sqrt{\frac{\lambda+2}{2\lambda}}\right). \tag{20}$$

Had there been no storage corruption, the situation would have been labeled "SKE", and the corresponding performance would have been obtained from Eq. (20) by letting $\lambda \to 0$.

$$\text{SKE} \quad \Pr(\text{correct}) = \Phi\left(\frac{S}{\sqrt{2}}\right) \tag{21}$$

Figure 4 presents Eq. (20) for no internal noise ($\lambda = 0$), for as much internal as external noise ($\lambda = 1$), and the limiting performance as the internal noise increases ($\lambda = \infty$).

To complete this analysis, the efficiency of this receiver should be computed (Ref. 3). The computations were carried out for efficiencies

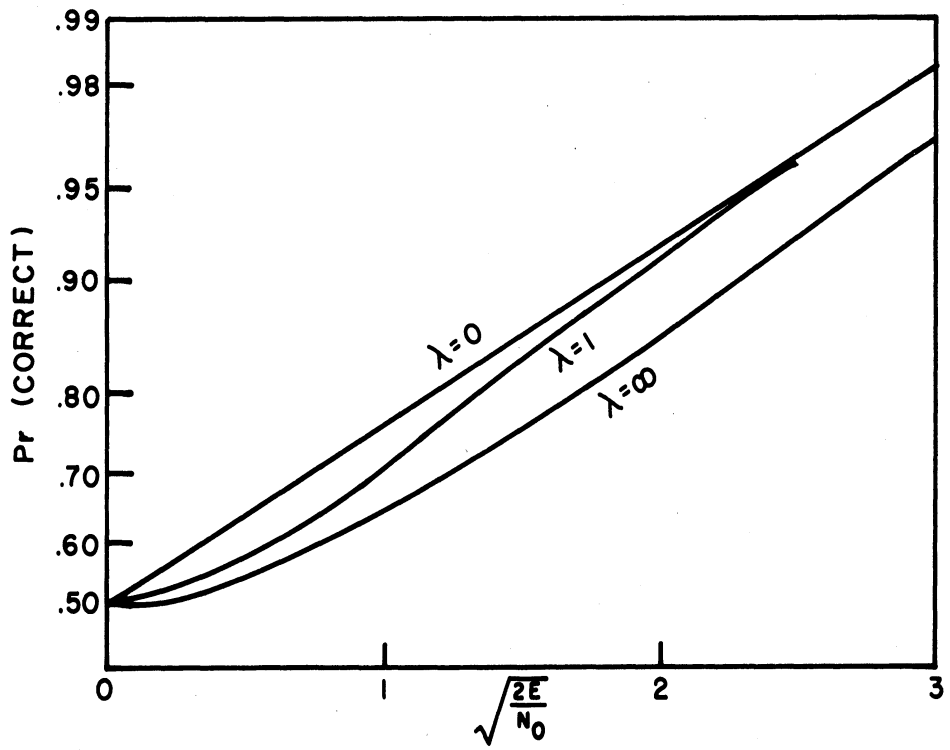

FIG. 4 IDEAL RECEIVER PERFORMANCE, TWO ALTERNATIVES FORCED CHOICE, 2WT = 1

above 0.10 for constant values of λ, and are shown in Fig. 5. Lines of constant performance are indicated on this figure to show the regions that would be encountered in normal psychophysical experimentation.

V. EVALUATION OF CROSSCORRELATOR, 2WT = 1

This is an extremely simple case, since the crosscorrelation simply supplies the sign for comparison. Specifically,

$$\int_0^T x_1(t)s(t)dt > \int_0^T x_2(t)s(t)dt \iff x_1 s > x_2 s \tag{22}$$

$$\iff \begin{cases} s > 0 & \text{and } x_1 > x_2 \\ \text{or} & \\ s < 0 & \text{and } x_1 < x_2 \end{cases} \tag{23}$$

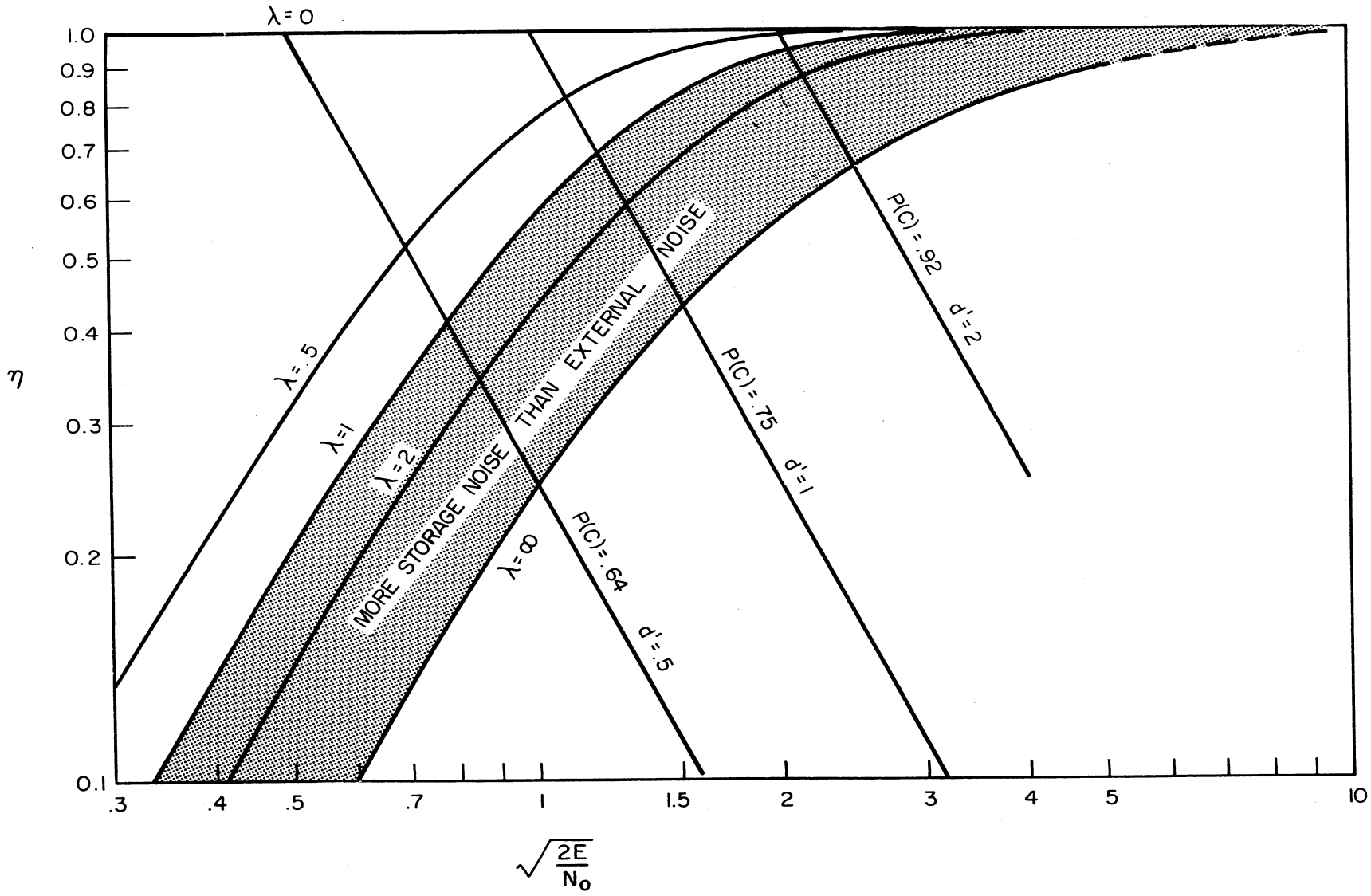

FIG. 5 RECEIVER EFFICIENCY, η, NOISY MEMORY LIKELIHOOD RATIO RECEIVER

when x_1 is due to signal plus noise, the probability that $x_1 > x_2$ is the same as the probability that $x_1 - x_2 > 0$, namely,

$$\Pr(x_1 > x_2) = \Phi\left(\frac{S}{\sqrt{2}}\right) \tag{24}$$

Under any conditions the probability that the stored sample is positive is

$$\Pr(s > 0) = \Phi\left(\frac{S}{\sqrt{\lambda}}\right) . \tag{25}$$

Hence,

$$\text{Prob(correct)} = \Phi\left(\frac{S}{\sqrt{2}}\right)\Phi\left(\frac{S}{\sqrt{\lambda}}\right) + \Phi\left(-\frac{S}{\sqrt{2}}\right)\Phi\left(-\frac{S}{\sqrt{\lambda}}\right) . \tag{26}$$

By comparing Eq. (20), for the ideal, with Eq. (26) one sees that where the term $\sqrt{\frac{\lambda+2}{2\lambda}}$ appeared for the ideal, the term $\sqrt{\frac{1}{\lambda}}$ appears for the cross-correlator. For small values of λ ($\lambda < 0.1$), these are roughly the same; for large values of λ, the term for the ideal rapidly approaches $\sqrt{\frac{1}{2}}$, while the crosscorrelator term descends toward zero. The curves of Figs. 4 and 5 apply with the following corrections.

TABLE I

λ for crosscorrelator	0	.40	.67	1.00	2.00
λ for ideal	0	.50	1.00	2.00	∞

A complete curve of this relation is given in Fig. 6. This shows the serious loss of efficiency when the crosscorrelator memory is noisy, since 2 db more storage noise than external noise has the same effect as 8.5 db on the ideal (noisy-storage) receiver. Figure 7 is included for comparison with Figure 4.

VI. COMPARISON COMPUTATIONS

Two further computations may be made for the sake of comparison. Had the signal actually been distributed on transmission with the same

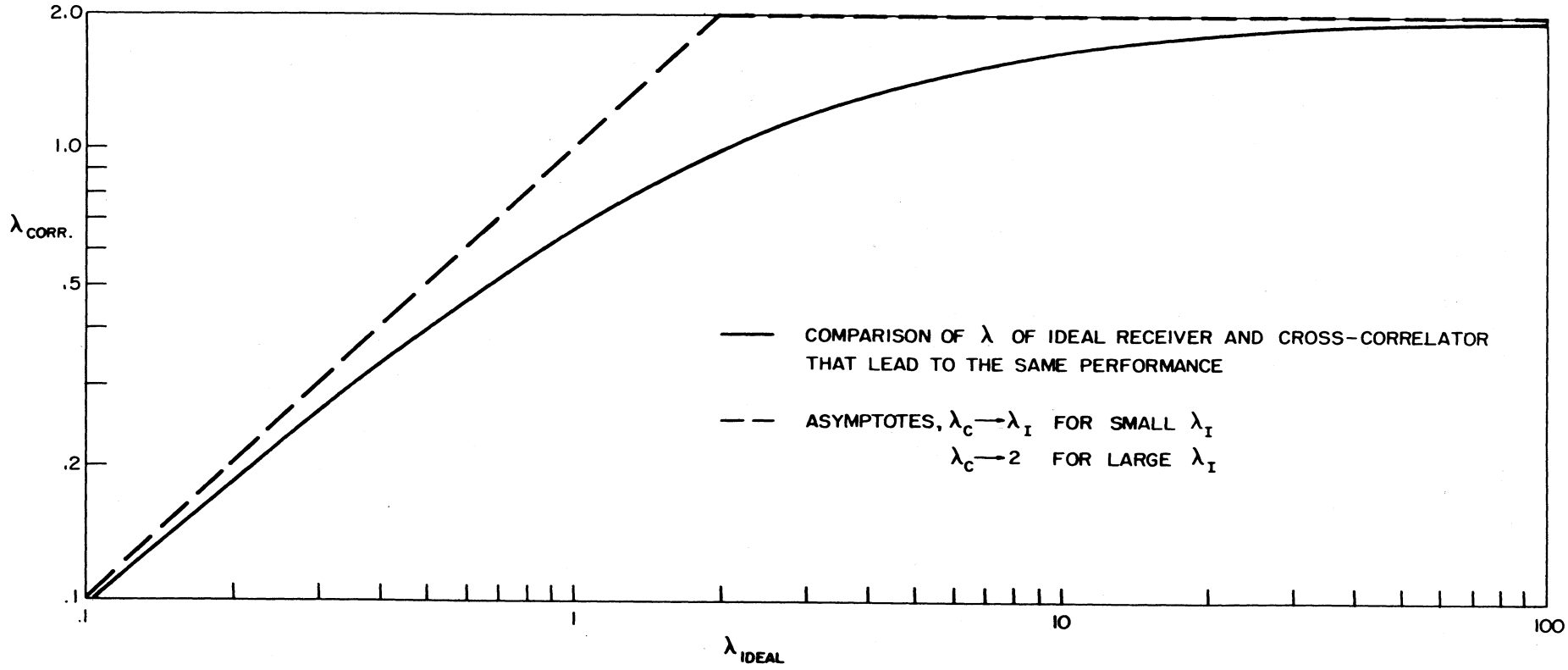

FIG. 6
λ FOR CROSSCORRELATOR VS. λ FOR IDEAL

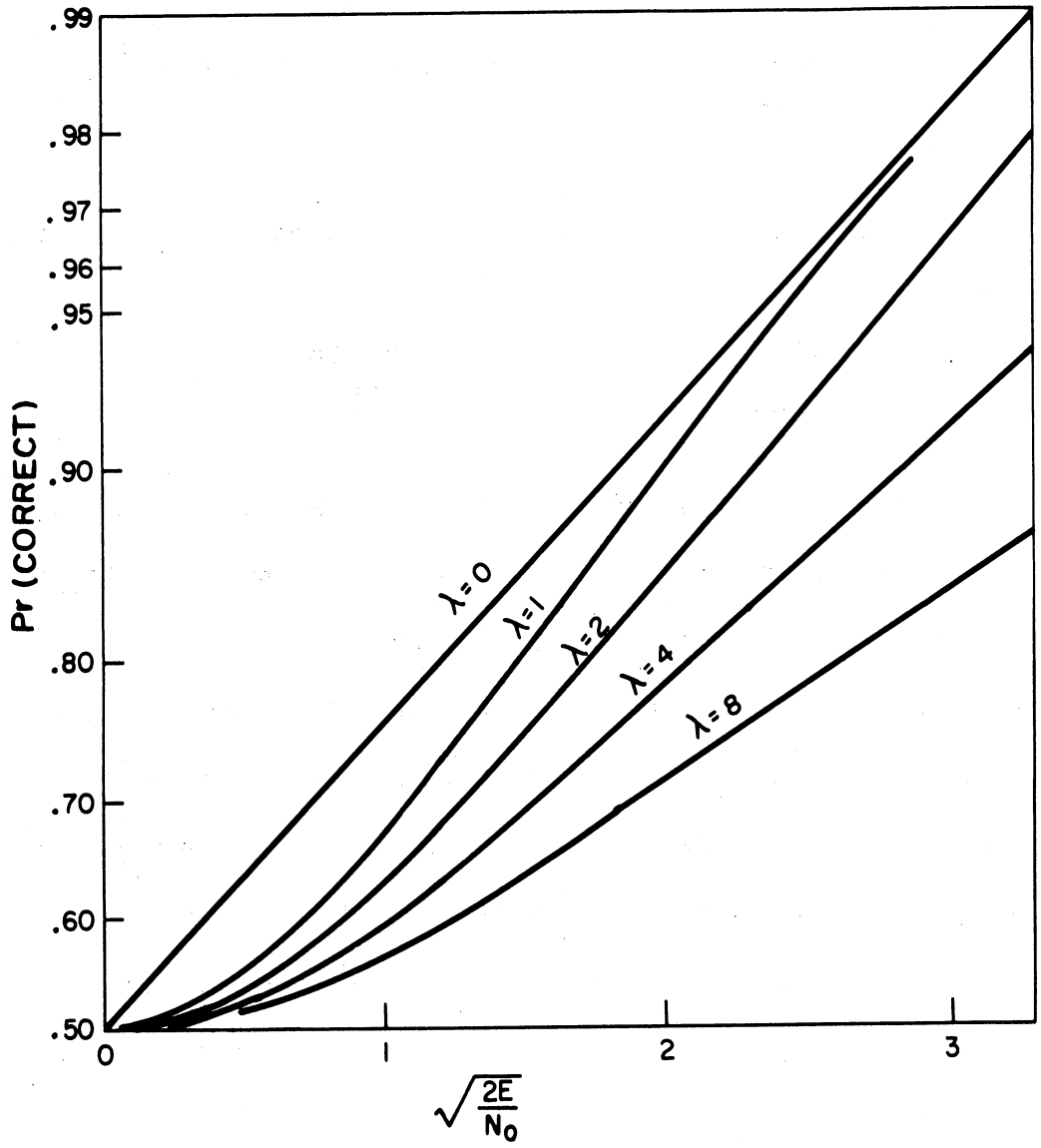

FIG. 7 CROSSCORRELATOR PERFORMANCE, TWO ALTERNATIVES FORCED CHOICE, 2WT = 1

distribution as that in storage, the ideal receiver of Section III would be the true likelihood receiver for this new case, and the only difference would be where the signal variance increases, sample variances also increase; namely, for a signal in the first interval,

$$\sigma^2(x_2) = 1 \qquad \sigma^2(x_1) = 1+\lambda \tag{27}$$

Equation (20) becomes:

$$\Pr(\text{correct}) = \left(\frac{S}{\sqrt{\lambda+2}}\right)\Phi\left(\frac{S}{\sqrt{\lambda}}\sqrt{\frac{\lambda^2+4\lambda+4}{\lambda^2+2\lambda+4}}\right)+\Phi\left(\frac{-S}{\sqrt{\lambda+2}}\right)\Phi\left(\frac{-S}{\sqrt{\lambda}}\Phi\sqrt{\frac{\lambda^2+4\lambda+4}{\lambda^2+2\lambda+4}}\right) \quad (28)$$

The radical $\sqrt{\frac{\lambda^2+4\lambda+4}{\lambda^2+2\lambda+4}}$ rises from 1.0 at $\lambda = 0$, to 1.15 between $\lambda = 2$ and $\lambda = 3$, then decreases slowly to 1.00 as $\lambda \to \infty$. Hence a good lower approximation is

$$\Pr(\text{correct}) = \Phi\left(\frac{S}{\sqrt{\lambda+2}}\right)\Phi\left(\frac{S}{\sqrt{\lambda}}\right)+\Phi\left(-\frac{S}{\sqrt{\lambda+2}}\right)\Phi\left(-\frac{S}{\sqrt{\lambda}}\right). \quad (29)$$

On close inspection, one observes that Eq. (29) is the exact equation for the crosscorrelator for this case. Hence, when the signal has the same variance as the stored signal and $2WT = 1$, the ideal receiver does only slightly better than the crosscorrelator. Thus it can be concluded that the improvement of the ideal over the crosscorrelator is much more important when the signal is actually stable but memory is poor, than when the signal is distributed and the lack of specificity is not due to poor memory.

A final calculation is for a receiver which has noise-free storage but λN_o joules per cycle <u>added</u> to the input. The detection index for such a receiver is

$$(d')^2 = \frac{2E}{N_o + \lambda N_o} = \frac{1}{1+\lambda}\,\frac{2E}{N_o} \quad (30)$$

so that the efficiency is

$$\eta = \frac{1}{1+\lambda} \,. \quad (31)$$

This receiver does not behave at all like those in the poor memory situation, and hence it can be concluded that, at least for this and similar cases, noisy memory does not act like "additional noise that can be reflected (invariantly) into the input."

VII. SUMMARY OF EQUATIONS FOR 2WT = 1

In this section the final performance equations are repeated, together with definitions of parameters. Since one man's normalization is another man's poison, two alternative notations are used.

First Normalization:

E - signal energy at receiver input and at receiver memory input.

N_o - noise power per cycle of white noise added to signals at receiver input.

λ - ratio of memory-noise power per cycle to N_o.

Signal Specified Exactly, Ideal Receiver (SSE, Ideal):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\right)\Phi\left(\sqrt{\frac{E}{N_o}}\,\Phi\sqrt{\frac{\lambda+2}{\lambda}}\right) + \Phi\left(-\sqrt{\frac{E}{N_o}}\right)\Phi\left(-\frac{E}{N_o}\sqrt{\frac{\lambda+2}{\lambda}}\right). \quad (20.1)$$

Φ is the normal or Gaussian distribution function $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$.

Signal Specified Exactly, Crosscorrelation Receiver (SSE, x-cor):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\right)\Phi\left(\sqrt{\frac{2E}{N_o\lambda}}\right) + \Phi\left(-\sqrt{\frac{E}{N_o}}\right)\Phi\left(-\sqrt{\frac{2E}{N_o\lambda}}\right) \quad (26.1)$$

Signal Known and Specified Statistically, Ideal Receiver (SKS, Ideal):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\sqrt{\frac{2}{\lambda+2}}\right)\Phi\left(\sqrt{\frac{2E}{N_o\lambda}}\sqrt{\frac{\lambda^2+4\lambda+4}{\lambda^2+2\lambda+4}}\right)$$

$$+\Phi\left(-\sqrt{\frac{E}{N_o}}\sqrt{\frac{2}{\lambda+2}}\right)\Phi\left(-\sqrt{\frac{2E}{N_o\lambda}}\sqrt{\frac{\lambda^2+4\lambda+4}{\lambda^2+2\lambda+4}}\right) \quad (28.1)$$

Signal Known and Specified Statistically, Crosscorrelation Receiver (SKS, x-cor):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\sqrt{\frac{2}{\lambda+2}}\right)\Phi\left(\sqrt{\frac{2E}{N_o\lambda}}\right) + \Phi\left(-\sqrt{\frac{E}{N_o}}\sqrt{\frac{2}{\lambda+2}}\right)\Phi\left(-\sqrt{\frac{2E}{N_o\lambda}}\right). \quad (29.1)$$

Second Normalization:

E - signal energy at receiver input.

N_o - noise power per cycle at receiver input.

e - signal energy at undistorted memory.

n_o - noise power per cycle of memory noise.

(SSE, Ideal):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\right) \Phi\left(\sqrt{\frac{E}{N_o} + \frac{2e}{n_o}}\right) + \Phi\left(-\sqrt{\frac{E}{N_o}}\right)\Phi\left(-\sqrt{\frac{E}{N_o} + \frac{2e}{n_o}}\right). \quad (20.2)$$

(SSE, x-cor):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\right) \Phi\left(\sqrt{\frac{2e}{n_o}}\right) + \Phi\left(-\sqrt{\frac{E}{N_o}}\right) \Phi\left(-\sqrt{\frac{2e}{n_o}}\right) . \quad (26.2)$$

(SKS, Ideal):

$$P(c) = \Phi\left(\sqrt{\frac{2eE}{En_o+2eN_o}}\right) \Phi\left(\sqrt{\frac{2e}{n_o}}\sqrt{\frac{E^2n_o^2+4EeN_on_o+4e^2N_o^2}{E^2n_o^2+2EeN_on_o+4e^2N_o^2}}\right)$$

$$+ \Phi\left(-\sqrt{\frac{2eE}{En_o+2eN_o}}\right) \Phi\left(-\sqrt{\frac{2e}{n_o}}\sqrt{\frac{E^2n_o^2+4EeN_on_o+4e^2N_o^2}{E^2n_o^2+2EeN_on_o+4e^2N_o^2}}\right) \quad (28.2)$$

(SKS, x-cor):

$$P(c) = \Phi\left(\sqrt{\frac{2eE}{En_o+2eN_o}}\right) \Phi\left(\sqrt{\frac{2e}{n_o}}\right) + \Phi\left(-\sqrt{\frac{2eE}{En_o+2eN_o}}\right) \Phi\left(-\sqrt{\frac{2e}{n_o}}\right). \quad (29.2)$$

In all cases, with no internal noise, both receivers are ideal and

(SKE, Ideal):

$$P(c) = \Phi\left(\sqrt{\frac{E}{N_o}}\right). \quad (32)$$

VIII. CONCLUSIONS

Several receivers have been discussed and evaluated. The one of primary concern in this paper is the receiver that is the optimum receiver when restricted to receivers with noisy memory and detecting a signal specified exactly in white Gaussian noise. The second receiver is the crosscorrelator, which would be the optimum receiver if the memory were perfect. The evaluations of Sections 4 and 5 are for the noisy-memory, signal-specified-exactly situation, for which the receiver under study is optimum and the crosscorrelator is not. In Section 6 the signal was actually distributed and both receivers evaluated; the memory in this case was considered noise-free, storing the mean of the signal distribution. The noisy-memory, signal-specified-exactly optimum receiver is also optimum for this case. In Section 6 a receiver with additional noise at the input but no memory noise was treated. It can be concluded that the performance of the "noisy-memory, signal-specified-exactly" optimum receiver detecting a signal under the conditions for which it was designed is not equivalent to the performance in any of the other receiver-signal combinations.

Two normalizations have been presented in Section 7, and the interpretation inherent with these normalizations deserves discussion. The first normalization, used in the analysis, considered the internal noise proportional to the external noise. Of course, this is possible for any fixed situation. However, when the efficiency, η, is plotted against the input signal quality $\frac{2E}{N_o}$ the rise in efficiency embodies the fact that the memory-signal quality is correspondingly increasing. The result is that for very small signals the efficiency is very low and rises rapidly as the signal level increases. In contrast, if the model being studied has a fixed-

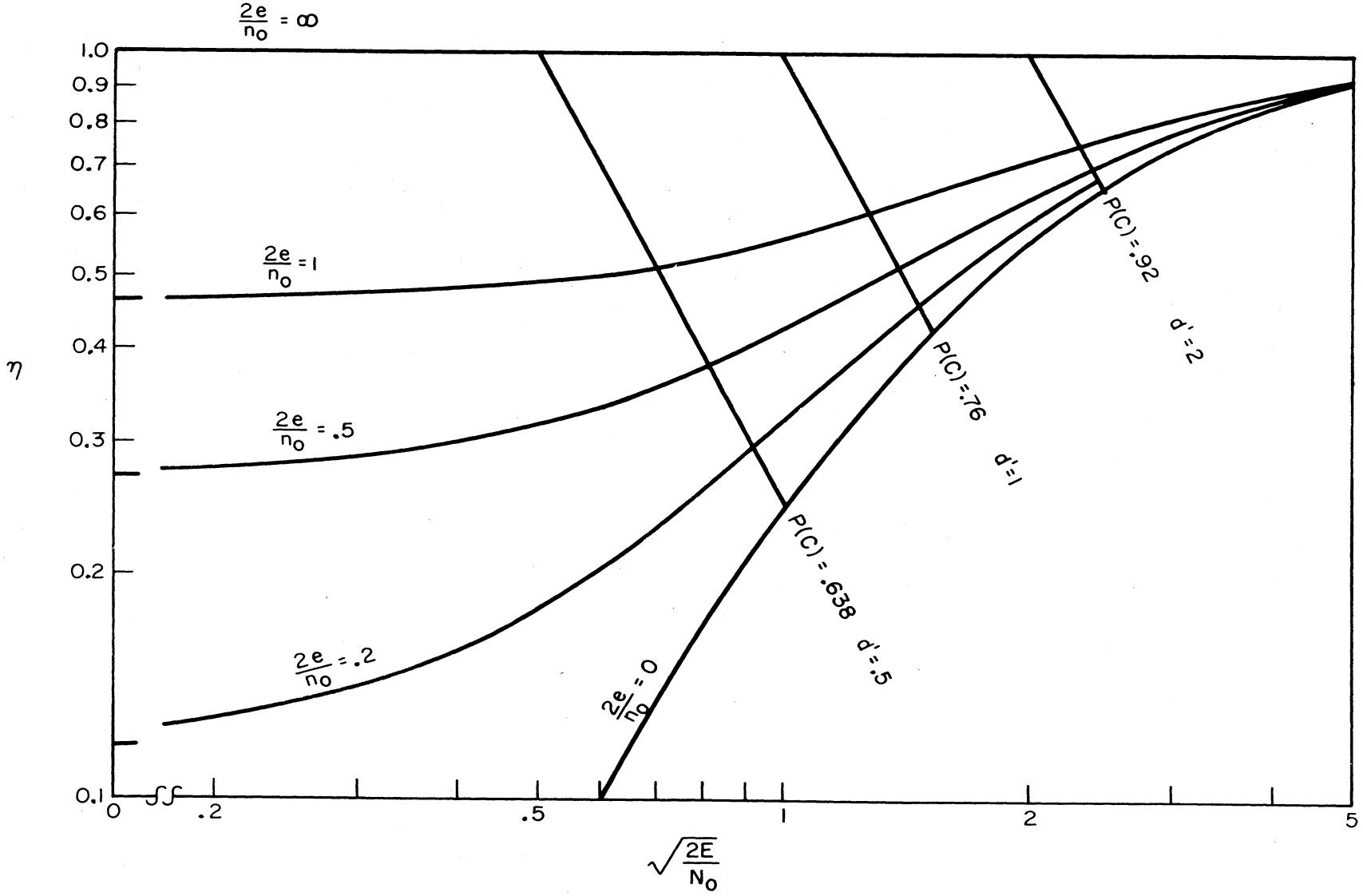

FIG. 8 RECEIVER EFFICIENCY, η, NOISY MEMORY LIKELIHOOD RATIO RECEIVER

quality memory signal $\frac{2e}{n_o}$ the increase of efficiency with signal level is quite different. Figure 8 shows this effect. The efficiency is constant for low level signals and then rises to 1 when the external quality becomes much larger than the internal quality.

In any application of a noisy-memory model the experimenter has one more variable to contend with. On the one hand, it is conceivable that this type of model could unify previously conflicting data. On the other hand, one would not expect the relation of internal quality to external quality of signal to be the proportional relation of Fig. 5 or completely independent as in Fig. 8.

The basic analysis of this problem has been only partially completed in this paper. The determination of the ideal receiver is completed in Section 3. However, the numerical evaluation has been completed only for the special case of 2WT = 1 and a symmetric forced choice in time where no memory degradation took place between presentations. Although the author feels this special case indicates the effect of "noisy-memory," a more complete evaluation might shed more light on this model.

IX. APPENDIX

Evaluation of Equation (4)

Insertion of Eqs. (3) and (5) into Eq. (4) yields

$$\ell(x) = \left(\frac{1}{2\pi\lambda N_o W}\right)^{WT} \int_S \exp\left\{\frac{1}{N_o}\int_o^T -S^2(t)+2x(t)s(t) - \frac{1}{\lambda}[s(t)-S(t)]^2 dt\right\} dS.$$

In evaluating this, a simplified notation is used to recognize the completion of the square in the exponent integrals; specifically, the time argument was omitted.

The negative of the integrand of the inner integral is

$$S^2 - 2xS + \frac{1}{\lambda}(s-S)^2 = S^2 - 2xS + \frac{s^2}{\lambda} - \frac{2sS}{\lambda} + \frac{S^2}{\lambda}$$

$$= S^2\left(1 + \frac{1}{\lambda}\right) - 2S\left(x + \frac{s}{\lambda}\right) + \frac{s^2}{\lambda}$$

$$= \left(S\sqrt{1 + \frac{1}{\lambda}} - \frac{x + \frac{s}{\lambda}}{\sqrt{1 + \frac{1}{\lambda}}}\right)^2 - \frac{\left(x + \frac{s}{\lambda}\right)^2}{1 + \frac{1}{\lambda}} + \frac{s^2}{\lambda}$$

$$= \left(1 + \frac{1}{\lambda}\right)\left(S - \frac{x + \frac{s}{\lambda}}{1 + \frac{1}{\lambda}}\right)^2 - \frac{\lambda x^2 + 2sx - s^2}{1 + \lambda} .$$

Hence,

$$\ell(x) = \left(\frac{1}{2\pi\lambda N_o W}\right)^{WT} \exp\left[\int_o^T \frac{\lambda x^2 + 2xs - s^2}{N_o(1+\lambda)}\, dt\right] \int_S \exp\left[-\frac{\lambda+1}{N_o\lambda}\int_o^T \left(S - \frac{\lambda x + s}{\lambda + L}\right)^2 dt\right] dS.$$

Now this should be examined in careful detail. The integral with respect to S is a Lebesgue integral in the 2WT dimensional function space. For each x(t) and s(t) the expression

$$\int_o^T \left[S(t) - \frac{\lambda x(t) + s(t)}{1 + \lambda}\right]^2 dt$$

represents the square of the distance from S(t) to the fixed point $\frac{\lambda x + s}{1 + \lambda}$. Hence the integral

$$\int_S \exp\left[-\frac{\lambda + 1}{N_o\lambda}\int_o^T \left(S - \frac{\lambda x + s}{\lambda + 1}\right)^2 dt\right] dS$$

is proportional to the probability of the region of integration because the integrand is proportional to the normal probability-density function in 2WT dimensional space, The region of integration is the whole space and hence has probability one, independent of the center $\frac{\lambda x + s}{1 + \lambda}$ of the distribution.

This integral has some (non-zero) value, k_1, which is not a function of S(t), s(t), or x(t), although it is a function of λ and N_o. This value need not be determined since the desired result is to show that the likelihood ratio can be written as

$$\ell(x) = k_1 \exp \int_o^T \frac{\lambda x^2 + 2sx - s^2}{N_o(1 + \lambda)} \, dt.$$

Thus $\ell(x)$ is strictly monotone increasing with the exponent

$$\ell(x) \text{ --smi-- } \int_o^T \left(x^2 + 2 \frac{sx}{\lambda} - \frac{s^2}{\lambda} \right) dt.$$

Confidence Ratings, Second-Choice Responses, and Confusion Matrices in Intelligibility Tests[1]

Frank R. Clarke, The University of Michigan

I. INTRODUCTION

In the standard intelligibility test, words (or some other message units) are read in random order over a noisy channel and the listener attempts to identify each stimulus item on the basis of the degraded signal available to him. The typical measure of the listener's performance is the per cent of the stimuli correctly identified. Aside from many clinical and engineering applications, this dependent variable, the intelligibility score, has been of great use in the study of speech perception. However, without extending the listener's task, it is possible to get a more complete description of his performance. By the use of small closed message sets it becomes feasible to construct confusion matrices which show the proportion of instances in which any given response is made to any given stimulus. This detailed analysis of the listener's identification responses has been used profitably in an analysis of perceptual confusions among English consonants (Ref. 1) and in the development of a rule for predicting the confusion matrix for any subset of items given the confusion matrix for its master set (Refs. 2,3).

As we attempt to extend our knowledge of the behavior of a listener in identifying degraded speech signals, it is apparent that the listener's

---

[1] This study is part of a thesis presented to Indiana University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. This research was supported in part by the U. S. Air Force under Contract No. AF 19(604)-1962, monitored by the Operational Applications Laboratory, Bolling Air Force Base, Washington 25, D. C. This is Report No. AFCRC-TR-58-54, ASTIA Document No. Ad 160712.

task in the intelligibility test, i.e., writing down a single response to each stimulus item, is such that some of the information about the stimulus which is available to him is discarded. For example, we know that on some occasions the listener is quite certain that this response is correct, and on others he is fairly uncertain of his response. It has been shown that the listener can assign ratings of confidence to his identification responses in a non-chance manner, and thus such ratings carry additional information about the stimulus items transmitted.

Still another possible information-bearing response of the listener has had little study. This is the listener's second-choice response. Often, after making his best guess as to the identity of the stimulus, the listener feels that he can make a reasonable second choice, with some confidence that this second choice will be correct in the event that his first choice is incorrect.

The experiments to be reported in this paper were designed to investigate certain aspects of a listener's ratings of confidence and his second-choice responses.

II. CONFIDENCE RATINGS APPLIED TO IDENTIFICATION RESPONSES

The experiment in this section will be used to provide the framework for the explanation of some of the concepts to be used throughout this paper. The data from this experiment will serve for making predictions and comparisons in later sections. This experiment deals with the use of a confidence rating by a listener to estimate the probability that his attempted identification of a transmitted message is, in fact, correct. The nature of these data is best understood by considering the manner in which they were obtained.

Procedure

Five highly trained listeners were utilized in this experiment. A closed set of stimulus words was used in intelligibility tests. This set consisted of the five spondees: duckpond, eggplant, greyhound, stairway, and vampire. Each of these words was represented equally often in a deck of 75 cards, and test lists containing these 75 items were constructed by shuffling this deck of cards. Each of these test lists was read by one talker (FC). The speech signals were mixed electrically with white noise (uniform spectrum level, 100 cps to 7000 cps), and then they were presented to the listeners over binaural headsets (PDR-10 earphones). The speech level was defined by the average (over all items read in the tests) of the peak deflections on a Daven VU meter of the signals as measured in the speech channel. The noise level was observed on a Daven VU meter placed across the noise channel. The speech-to-noise ratio was computed using these values with suitable corrections for attenuators in the circuit and correction for the reduction in the bandwidth of the noise at the earphones. A speech-to-noise ratio of approximately -16.5 db was chosen with the aim of achieving an average articulation score of about 60%. The overall level of the noise was approximately 82 db re 0.0002 dyne/cm^2. Items were read at the rate of one every 10 sec. On each stimulus presentation the listeners made two responses: 1) they attempted to identify the word read by the talker by writing down one of the five possible responses; and 2) they assigned a rating of 1, 2, 3, 4, or 5 to indicate how confident they were that their attempted identification of the stimulus item was, in fact, correct. The listeners were instructed to use the rating 1 in such a way that responses given this rating would have a probability of being correct lying within the range 0.90 to 1.00. Ratings of 2 through 5 were to indicate estimates of *a posteriori* probabilities within the

ranges 0.75 to 0.90, 0.55 to 0.75, 0.35 to 0.55, and 0.00 to 0.35, respectively. After each test of 75 items, the listeners scored and tabulated their own responses, and thus had constant feedback to aid them in their attempt to use the ratings as instructed. Prior to the tests for which data are here reported, the listeners had taken 299 intelligibility tests of typical length, 90 of which involved the use of the above mentioned set of spondees presented at the speech-to-noise ratio of -16.5 db. Twelve of these tests provided practice in the use of ratings. Twenty-two of the subsequent 24 tests met a pre-established criterion that the average articulation score for the group be between 50% and 70%. These tests provided the data for the following discussion.

Results and Discussion

Obtained proportions in this and the following experiments will be taken as estimates of underlying probabilities. The data for any individual listener may be presented in complete detail (except for any time-dependent effects) in a 5 x 25 matrix - five rows for each of the stimuli and 25 columns for each of the response pairs (one of five identification responses followed by one of five ratings). By summing entries in this table over rating categories, we may obtain a typical confusion matrix with rows corresponding to stimuli and the five columns corresponding to identification responses. For certain of the conclusions presented below, it is assumed that this latter matrix does not differ systematically from that which would have been obtained if the listener had made a single identification response and had not followed this response with a confidence rating. Intuitively this seems reasonable, and from previous studies (Refs. 4,5) it is known that the added confidence rating does not depress the listener's articulation score.

One concern of this study is the extent to which the additional rating response adds information over and above the information transmitted by the identification response. Implicitly, at least, we will be considering this in a general sense throughout the rest of this section. However, first let us examine the data in terms of the technical meaning of the term "information," using Shannon's measure of information transmitted (Ref. 6). To calculate the amount of information added by the rating response to that already transmitted by the identification response, it is necessary to determine the amount of information transmitted when both responses are considered jointly and subtract from this the information transmitted when only the identification response is considered. Since the degrees of freedom for the two tables differ, it is first necessary to apply the Miller-Madow correction for bias (Ref.7) to the calculated values of information transmitted (I_t).

TABLE I

	A I_t(S x I • R)	B I_t(S x I)	C A - B	D B/A	E p(C)
L1	0.689	0.553	0.136	0.803	0.566
L2	0.994	0.889	0.105	0.894	0.697
L3	1.063	0.905	0.158	0.851	0.696
L4	0.547	0.471	0.076	0.861	0.520
L5	0.686	0.527	0.159	0.768	0.566
Mean	0.796	0.669	0.127	0.835	0.609
Mdn.	0.689	0.553	0.136	0.851	0.566

Analysis of rate of information transmitted for each of the five listeners in this experiment. Column A shows information transmitted per stimulus item when identification responses and rating responses are considered jointly. Column B shows information transmitted per stimulus item when only identification responses are considered. Columns C and D are self-explanatory. Column E gives the per cent correct responses for each listener.

Table I summarizes the results for each listener in terms of calculated rates of information transmission. On the average, the rating response added 0.127 bit per item to the rate of information transmission for the identification response alone. Of the information carried by the identification response and the rating response considered jointly, 16.5% was contributed uniquely by the ratings.

In many communication situations the listener could easily utilize the additional information provided by the ratings. For example, if the listener is provided with a back-channel and is allowed to trade speed of transmission for accuracy, then, he could accept only those messages which he rated with a rating R_k or stricter, and ask for repeats on all others (Refs. 8,9). In Table I we see that the articulation score, p(C), for L1 is 0.566. Now, if the listener decided to accept his identification response only when his confidence rating was a 1 or a 2, and if whenever his confidence rating was a 3, 4, or 5 he were to ask the speaker to repeat the message, then, L1 would be expected to identify correctly 0.769 of the messages in the set; but it would take 4.5 times as long as needed to attain 0.566 identification. Note that this method of increasing accuracy of message reception only involves trading speed for accuracy; it does not depend on the listener having a knowledge of his confusion matrix. Of course, the articulation score may be improved without the use of ratings merely by repeating each message several times before moving on to the next message. At the end of each series of transmitted messages, the listener may then make a single response in attempting to identify the transmitted message. However, the addition of ratings in this situation also results in a marked improvement. (See Egan, Ref. 10, for details. Especially Case 2 in Chap. 3 and in Chap. 5, and Appendix D.)

The ability of the receiver to assign ratings in accordance with a posteriori probabilities is indicated in Fig. 1. The points in this figure show the obtained a posteriori probabilities associated with each of the rating categories, while the bars show the probability with which the listener used each category. These obtained a posteriori probabilities were calculated by dividing the number of times a listener assigned a given rating to a correct identification response by the total number of times that the listener used that particular rating. It will be noted that all listeners did a fairly good job of using the rating categories as directed. However, these are averages over the five stimulus items, and a more detailed analysis of the data indicated that the listeners tended to overestimate a posteriori probabilities for items of relatively low intelligibility and to underestimate these probabilities for items of relatively high intelligibility. This supports a similar finding by Pollack and Decker (Ref. 5).

Much of the recent work dealing with the confirming behavior of a receiver has had results reported in terms of the Type II ROC curve (Refs. 11,12). This curve, the receiver operating characteristic, is a plot of the probability that the receiver will confirm his identification response, given that it is correct, against the probability that he will confirm his identification response, given that it is incorrect. Where ratings are used, rather than a binary decision, the assumption is made that under some particular criterion, the receiver would accept as correct all identification responses to which a confidence of R_k or stricter had been given and would reject all responses in which he had less confidence. Figure 2 shows the data for the five listeners plotted as a Type II receiver operating characteristic. In this figure the proportions are

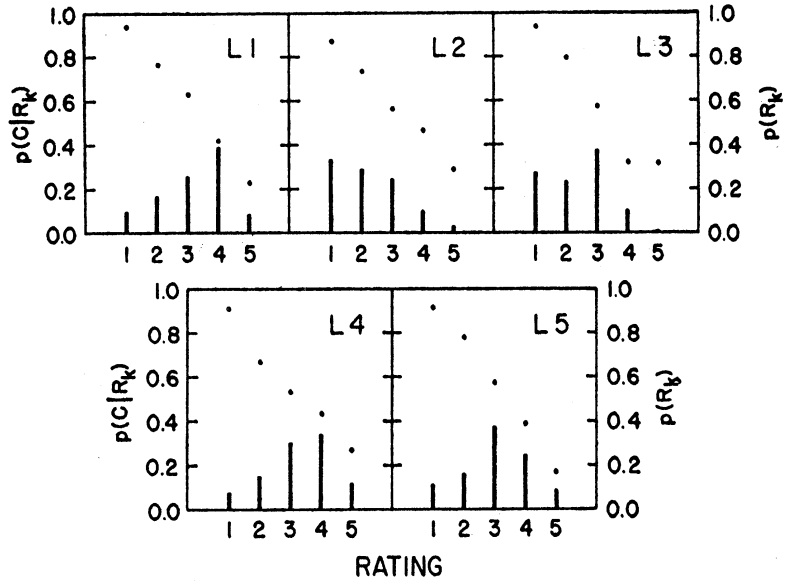

FIG. 1. The bars in this figure show the probability, $p(R_k)$, that the listener assigned the rating R_k to his identification response. The points indicate the probability that the listener was correct given that he assigned the rating R_k to his identification response. The data for each listener are shown separately.

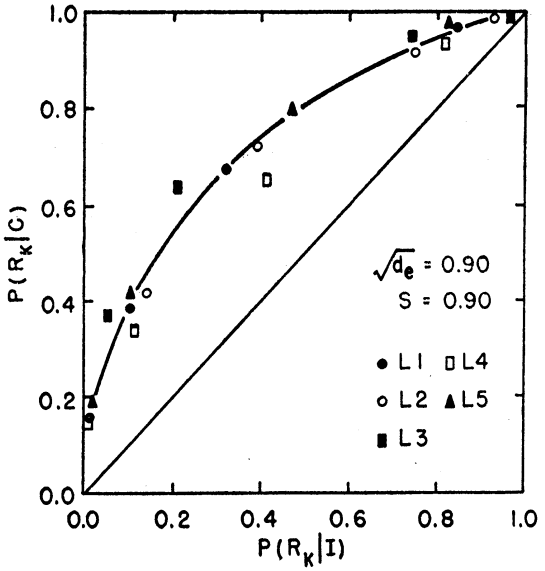

FIG. 2. A Type II ROC curve for five listeners obtained with a set of five spondees read at a speech-to-noise ratio of -16.5 db. The ordinate shows the probability of a listener accepting his identification response as correct given that his identification response is, in fact, correct. The abscissa indicates the probability of a listener accepting his identification response as correct given that it was incorrect. Here $P(R_K|C) = \Sigma_{k=1,K}\, p(R_k|C)$ and $P(R_K|I) = \Sigma_{k=1,K}\, p(R_k|I)$. The two parameters describing the solid curve drawn through the data are $(d_e)^{\frac{1}{2}}$ and s.

plotted on a linear scale. The solid curve drawn through the data points fits very nicely, and since it is a straight line when plotted on normal-normal probability paper, as in Fig. 3, a listener's performance may be specified by two parameters. The two parameters chosen to represent the line are $(d_e)^{\frac{1}{2}}$ and s, the slope of the line (Ref. 11). For each listener's data, the straight line which minimized the sum of the squares of the perpendicular distances from the points to the line was obtained. The average value of $(d_e)^{\frac{1}{2}}$ and the average value of s were used to construct the single line plotted in Fig. 2.

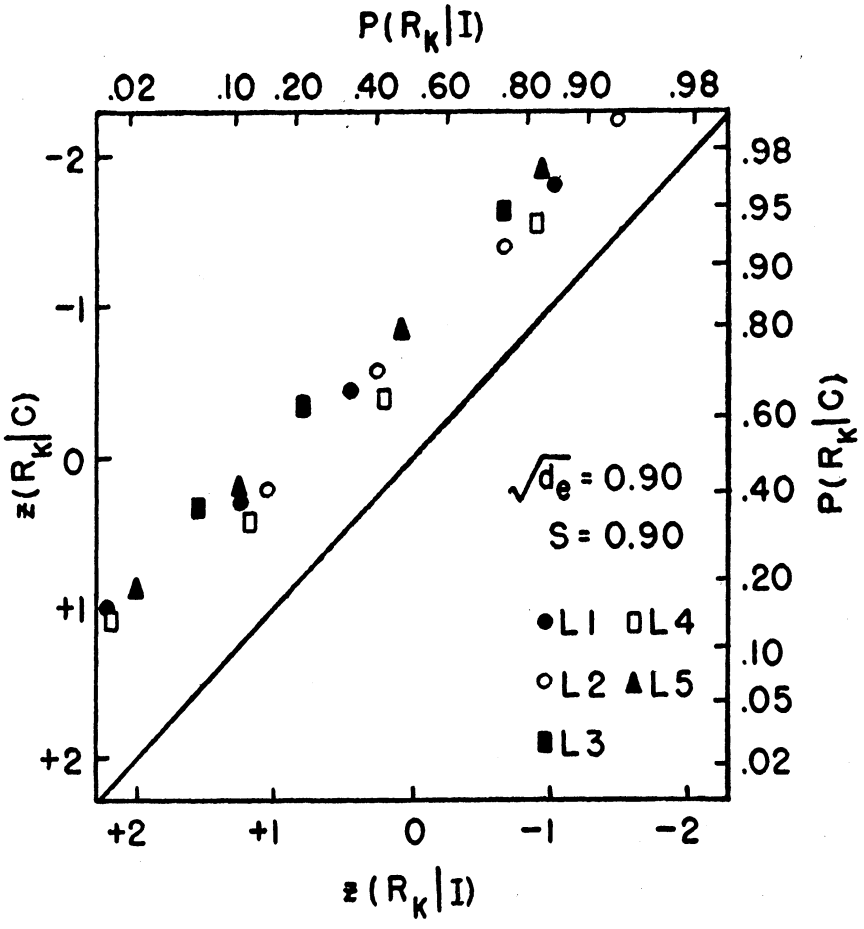

FIG. 3. The data of Fig. 2 plotted on normal-normal probability paper.

III. THE USE OF TWO IDENTIFICATION RESPONSES

We have seen in the previous section that requiring a listener to rate his confidence in his identification response is one method of recovering some of the information which is available to the listener but which is normally lost when he makes s signle identification response. Another method of attempting to tap this additional information is to require the listener to respond to each stimulus presentation with his best estimate of what item was transmitted and to follow this with his second-best estimate.

Little research has been carried out concerning a subject's second-choice behavior. Shannon (Ref. 13) required subjects to use first, second and additional "guesses" as a technique for estimating the entropy of printed English. Also, Quastler and Blank (Ref. 14) have used this technique to estimate information transmitted by visual displays. In neither of these instances was the emphasis on the analysis of the subject's second and subsequent choice behavior per se.

Swets, Tanner, and Birdsall (Ref. 15) investigated the second-choice beahvior of a viewer attempting to determine in which of four short intervals of time an increment in the intensity of a light occurred. They found the viewer's second-choice performance to be significantly better than that predicted by chance, and, in fact, they were able to make fairly good predictions of the obtained data using their "decision-making theory of visual detection."

Pollack (Ref. 16), in a verbal learning context, investigated the information content of the subjects' second, third and fourth choices. Here, when a subject made an error in attempting to recall an item he

was so informed and required to make another guess. Bricker and Chapanis (Ref. 17) carried out a similar experiment with tachistoscopically presented stimuli. In both of these studies, the information transmitted by the subject's second choice was compared with that which would be expected under the simplest of threshold models. Both studies purported to demonstrate that considerable information is transmitted by the subject's second and later choices. However, they did not analyze their data in such a way as to determine to what degree the information contained in the second choice (for example) is independent of that contained in the first choice. The fact that the second choice contains a considerable amount of information about the stimulus does not permit the conclusion that the second choice adds information to that already contained in the first choice. Certainly, one cannot conclude that the information transmitted by the first and second responses considered jointly is equal to the information transmitted by the first choice plus the information transmitted by the second choice. The present study will attempt to determine to what extent the information contained in the second choice does, in fact, add to that contained in the first choice.

The study reported in this section is largely empirical. Because of the paucity of data concerning second-choice behavior, as well as the complexity of the speech signal, there is no adequate theory which leads to the prediction of a listener's stimulus-by-second-choice confusion matrix given only his stimulus-by-first-choice matrix. A simple threshold model fails to account for the obtained data. An extension of the constant-ratio rule suggests that the listener would make his second-choice response as though he were responding to the stimulus item with a set of response alternatives limited by removing his first choice from his original set

of alternatives. However, in this study, this model also fails to account for the obtained data. A simplified version of the theory of signal detectability also fails to account for the observed data.

Procedure

The procedure for obtaining the data to be reported in this section was essentially the same as that described in section II, the only basic change being in the instructions to the listeners. Here, the listeners were instructed to make their best attempt to identify the stimulus item, and to follow this with their next best choice. Thus, after the stimulus item as spoken only once, the listeners made two identification responses. Both choices could not be the same. The same talker and listeners were used here as in obtaining the rating data reported in the previous section. Two additional listeners were used for some of this second-choice work. Two sets of five spondees each were used: Set 1, cupcake, padlock, pancake, starlight, and wildcat; and Set 2, duckpond, eggplant, greyhound, stairway, and vampire. Listeners were highly practiced on the second-choice task with these message sets before data were obtained. It will be noted that Set 2 is identical to that used to obtain the rating data; thus, direct comparisons between the two types of second responses are possible. The data in this section were obtained prior to those reported in the previous section.

Results and Discussion

Before examining data of more general interest, it is necessary to consider the possibility that any information transmitted by the listener's second choice is gained at the expense of information in his first choice. That is, when a listener's task is to give two responses to each stimulus item he might distribute the information available to him over both

responses; yet, if he were limited to a single response, he might well be able to convey the same amount of information in this single response. If this were occurring, it would almost certainly have the effect that the listener's articulation score for his best estimate of the stimulus item when he was required to make two responses would be lower than the articulation score for his best estimate when required to make only one response to each stimulus item. The equivalence of a listener's first-choice behavior when required to make only one response to each transmitted message and when required to make two responses to each message was investigated by comparing first-choice articulation scores obtained by each of the two procedures. Carefully controlled tests were conducted using a set of five items. On one-half of the tests the five listeners responded with a single best estimate of the transmitted message. On the remaining tests they responded with their best estimate followed by a second choice. Tests were counter-balanced, and talker variables were eliminated through the use of recorded tests. (Recorded speech was mixed with "live" noise.) Twelve tests were conducted under each condition. The mean and median articulation scores of the single-choice tests were 0.617 and 0.640, respectively. Corresponding mean and median articulation scores for the listener's first choice in the two-response tests were 0.631 and 0.640. Clearly, this observed difference is not statistically significant. It would appear safe to conclude that requiring a listener to make a second-choice response does not depress his first-choice performance (Ref. 15).

For the informational analysis we will first look at some of the data obtained using Set 2. As there were five stimulus items in this set, on any given trial a listener could respond with any one of the five

items as his first identification response and with any one of the four remaining items as his second identification response. Thus, the data for each listener may be presented in a 5 x 20 matrix having entries which show the proportion of instances in which a particular response pair occurred, given that a particular stimulus item was read. By appropriately collapsing this large matrix, we may obtain for each listener two separate matrices: 1) a stimulus-by-first-identification-response matrix, and 2) a stimulus-by-second-identification response matrix. For each of these three matrices, we may calculate the corresponding rate of information transmission. These are 1) the information transmitted by the listener's first and second responses considered jointly, $I_t(Sx1 \cdot 2)$; 2) the information transmitted by the listener's first identification response, $I_t(Sx1)$; and 3) the information transmitted by the listener's second choice, $I_t(Sx2)$. The information transmitted by the listener's second identification response which is independent of that transmitted by his first response is given by $I_t(Sx1 \cdot 2) - I_t(Sx1)$.

Table II summarizes the results obtained in this experiment using stimulus Set 2, the same set of items for which rating results were reported in section II. As all conditions except the listener's second-response task are as similar as possible to the conditions reported in the previous section, Table II may be directly compared with Table I. Comparison of the two sets of data shows that the average articulation scores for the group differ by only 1%; furthermore, for the two sets of data the average rates of information transmission for the stimulus-by-first-response matrix differ by only 0.021 bit per item. Thus, comparison of the information-carrying capacity of the two types of additional responses, ratings and second identification responses,

TABLE II

Analysis of the rate of information transmitted for each of the listeners in the second-choice experiment utilizing message set 2. Column A shows the rate of information transmission when the listener's first and second identification responses are considered jointly. Column B shows the rate of information transmission when only the listener's first identification response is considered. Column C shows the rate of information transmission when only the listener's second identification response is considered. Columns D, E, and F are self-explanatory. Column G shows the probability that the listener's first identification response is correct. Column H shows the probability that the listener's second identification response is correct, given that his first identification response was incorrect.

	A $I(S \times 1{\cdot}2)$	B $I(S \times 1)$	C $I(S \times 2)$	D A - B	E (A - B)/C	F B/A	G $p(C_1)$	H $p(C_2 \mid I_1)$
L1	0.556	0.527	0.100	0.029	0.290	0.948	0.559	0.376
L2	0.852	0.830	0.113	0.022	0.195	0.974	0.667	0.331
L3	1.002	0.950	0.461	0.052	0.113	0.948	0.713	0.495
L4	0.393	0.383	0.136	0.010	0.074	0.975	0.487	0.348
L5	0.562	0.548	0.227	0.014	0.062	0.975	0.576	0.404
Mean	0.673	0.648	0.207	0.025	0.147	0.964	0.599	0.391
Mdn.	0.562	0.548	0.136	0.022	0.113	0.974	0.567	0.376
L6	0.638	0.597	0.286	0.041	0.143	0.936	0.580	0.075

appears justifiable. In Table I, we see that the rating response carried on the average 0.127 bit per item over and above that carried by the identification response. In Table II we seen that the second identification response only averages 0.025 bit per item which is independent of the information carried by the first identification response. Thus, on the average, only 14.5% of the information carried by the second response was independent of that carried by the first response. Furthermore, unlike the case with the rating response, there is no simple way in which the listener can utilize the additional information carried by the second identification response and in most cases it would be of no practical use.

From column H of Table II, we see, given that the listener's first identification response was incorrect, the probability of his being correct on his second response is 0.391. Comparing this figure to that which would be expected by "chance," 0.250, we see that the subject's performance may seem paradoxically "good," considering the small amount of information which the second identification response adds to that transmitted by the first identification response. However, it must be realized that the fact that the listener's second choices are considerably better than "chance," when measured in terms of per cent correct, in no way implies that the second choice carries any information about the stimulus which is not already carried by the first response. For example, given a set of words with a nonhomogeneous confusion matrix, we could read the words in a background of noise to a listener who makes a single identification response. He could then pass his response to a second subject who was not allowed to listen to the stimulus item when it was read. This second subject could be required to make the second-identification response, that is, attempt to identify the stimulus on the assumption that the listener's identification was incorrect. If

the stimulus items had varying degrees of interconfusibility, and if the second subject had some knowledge of the various factors leading to confusions among items, we would expect him to make identification considerably better than "chance expectation" even though he never heard the stimulus item. Clearly in this case the second subject is making his response to a stimulus which is the response of the listener, and even though this second response may contain considerable information about the stimulus, it cannot contain any information independent of that contained in the listener's response.

In order to give some meaning, in terms of per cent correct, to the average of 0.025 bit per symbol added by the listener's second identification response, the listener's actual performance will be compared with that which might be expected on the assumption that the listener's second response was, in fact, a response made to his first response, and that it was not otherwise influenced by the stimulus. For this purpose, let us look only at those instances in which the listener's first response was incorrect. Then, for each listener we may construct a stimulus-by-first-choice matrix, given that the first choice was incorrect as well as a first-choice-by-second choice matrix for those instances in which the first choice was incorrect. As these are tables of conditional probabilities, we may treat them as transition matrices for a Markov chain; therefore, the product of these two matrices gives the expected stimulus-by-second-choice matrix for the listener under the assumption that the second response is made solely to the first response, and otherwise is uninfluenced by the transmitted stimulus item. By way of example, Table III compares the obtained stimulus-by-second-choice matrix for L1 with that computed by taking the product of the above-mentioned matrices. It will be noted that the matrix computed on the Markov

assumption looks verydifferent from a "chance matrix." Figure 4 indicates the relationship between the entries of the obtained stimulus-by-second-choice matrix and those computed by the Markov assumption for all listeners. Although a more restricted definition of "information" than Shannon's is utilized, the fact that 22 out of 25 of the obtained diagonal entries are higher than the computed diagonal entries implies that the 0.025 bit per item added by the listener's second response is highly significant in a statistical sense.

The data for L6 which are included in Table II, but not otherwise referred to thus far, are included as a check on the training and motivation of the other five listeners. This listener (the author) was an experienced and highly motivated subject and comparison of his data with those of the other listeners would imply that their training and motivation were sufficient to result in a meaningful test of the information-carrying capacity of the second identification response.

TABLE III

The upper entries in this table indicate the obtained stimulus-by-second-choice matrix for L1. The lower, parenthetical entries resulted from taking the product of this listener's stimulus-by-first-choice matrix, given that the first choice was incorrect and his first-choice-by-second-choice matrix, given that the first choice was incorrect.

	2nd response when 1st incorrect				
Stimulus	DP	EP	GH	SW	VP
DP	0.558 (0.440)	0.115 (0.157)	0.182 (0.218)	0.079 (0.094)	0.067 (0.091)
EP	0.230 (0.225)	0.390 (0.299)	0.198 (0.232)	0.037 (0.054)	0.144 (0.191)
GH	0.281 (0.319)	0.248 (0.242)	0.318 (0.257)	0.073 (0.067)	0.080 (0.116)
SW	0.301 (0.426)	0.145 (0.134)	0.120 (0.110)	0.241 (0.133)	0.193 (0.198)
VP	0.212 (0.212)	0.217 (0.287)	0.179 (0.207)	0.043 (0.071)	0.348 (0.223)

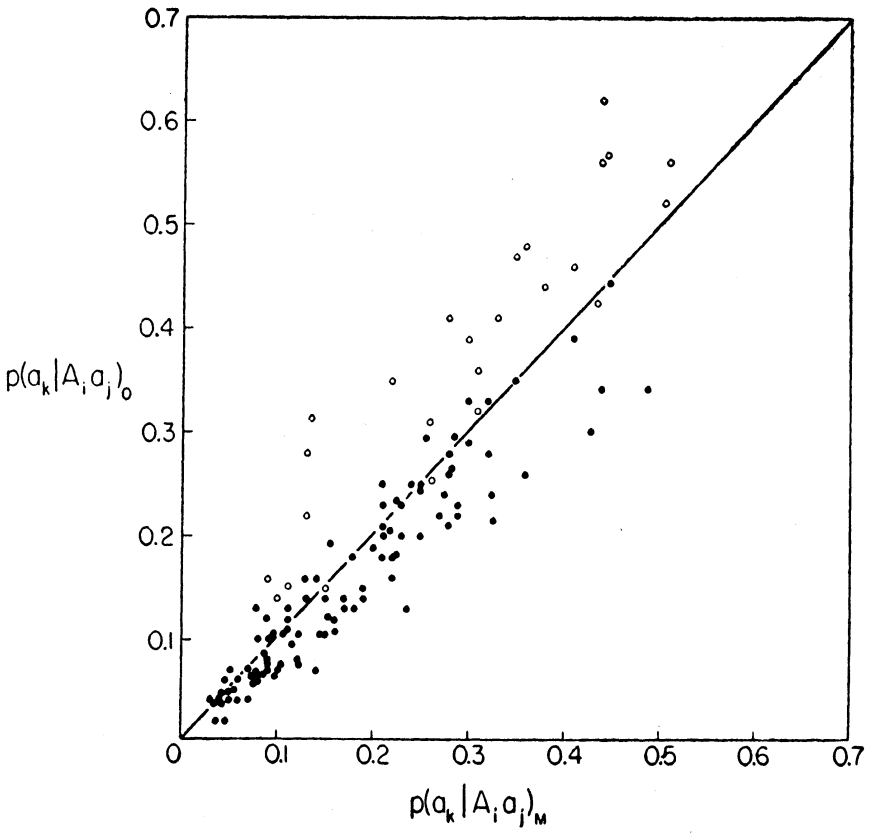

FIG. 4. The obtained probability of a second-choice response given that the first identification response was incorrect plotted against the corresponding predicted probability under a Markov assumption. The open circles indicate obtained and predicted probabilities of correct responses on the listener's second choice. The solid points show obtained and predicted probabilities of incorrect responses on the listener's second choice.

The above analysis was based upon the results obtained with stimulus set 2. Table IV summarizes the results for these listeners with stimulus set 1. These data clearly support those reported in Table II.

The data reported in this section may be summarized by making the following points: 1) On the average, of the information carried by the first and second responses considered jointly, over 97% is carried by the first choice alone. 2) On the average, of the information carried by the listener's second identification response, only about 7% is independent of information carried by the listener's first-choice response. 3) If the second response is a rating response, it will add about 6 times as

TABLE IV

Analysis of the rate of information transmitted for each of the listeners in the second-choice experiment utilizing message set 1. All entries are to be interpreted in the same way as those in Table II.

	A I(S x 1·2)	B I(S x 1)	C I(S x 2)	D A - B	E (A - B)/C	F B/A	G $p(C_1)$	H $p(C_2 \mid I_1)$
L1	0.526	0.516	0.295	0.010	0.034	0.981	0.569	0.386
L2	0.720	0.719	0.223	0.001	0.004	0.999	0.628	0.367
L3	0.970	0.911	0.367	0.059	0.161	0.939	0.701	0.452
L4	0.381	0.388	0.192	-0.007	-0.036	1.018	0.504	0.345
L5	0.555	0.537	0.255	0.018	0.071	0.968	0.570	0.412
L7	0.664	0.650	0.305	0.014	0.046	0.979	0.619	0.379
Mean	0.636	0.620	0.273	0.016	0.047	0.981	0.598	0.390
Mdn.	0.610	0.594	0.275	0.012	0.040	0.980	0.594	0.382

much information to that carried by a first identification response as will a second identification response. Furthermore, the additional information in the ratings is easily utilized by the listener, whereas that in the second identification response cannot easily be used in most situations.

It must be pointed out that the generality of these results may be limited. It is clear that the information content of the listener's second identification response will depend upon both the size of the message set and upon the signal-to-noise ratio at the listener's ear. If sets of only two items are used, the subject's second-identification response will contain as much information as his first choice, but the second choice can contain no information independent of that contained in his first. As the size of the message set is increased, it becomes possible for the second choice to contain less information than the first choice (though probably not more), but some of this information may now be independent of that contained in the first choice. As the signal-to-noise ratio is varied from minus infinity to plus infinity, the information in the second choice which is independent of that in the first choice would be expected to range from zero through a maximum and then back to zero again.

In order to make calculations of I_t, the number of probabilities to be estimated grows as the cube of the size of the message set. It was this consideration which limited the size of the message set investigated. The speech-to-noise ratio was chosen in an attempt to come near to maximizing the independent information content of the second response.

IV. A POSTERIORI PROBABILITIES AND THE RECEIVER OPERATING CHARACTERISTIC

We have seen in section II that, for a given message set at a particular speech-to-noise ratio, the listeners can be trained to give fairly

accurate estimates of the probability that their identification response is correct. There is the question of whether or not the listeners can make accurate estimates of a posteriori probabilities when the speech-to-noise ratio and the size of the message set change from test to test. The data presented in this section deal with this question. These data will also serve to give some indication of how the parameters of the receiver operating characteristic vary as a function of the size of the message set and the speech-to-noise ratio.

Procedure

In general, the procedure for obtaining these data was the same as that described in section II. The listener always made two responses to each item: an identification response, and a rating that indicated his degree of confidence in the correctness of his identification response. Five rating categories were used to indicate estimates of a posteriori probabilities in the following ranges: 1) 0.90 - 1.00; 2) 0.70 - 0.90; 3) 0.50 - 0.70; 4) 0.30 - 0.50; and 5) 0.00 - 0.30. Speech-to-noise ratios for the various tests ranged from -9db to -18db in 3-db steps. Closed message sets of two sizes were used: m = 4 and m = 16. Sixteen four-word message sets and four sixteen-word message sets were selected randomly from a list of 1000 monosyllabic words (Ref. 18). On each test, the listeners had copies of the message set before them, and they were required to respond to each item with a word from this list. The talker and five listeners had served in the experiments reported in the earlier sections of this paper. After two days of practice, which covered all message sets and all speech-to-noise ratios to be used in this experiment, the following basic design was replicated (using different sets of words) four times. On each of three consecutive days there were two 80-item intelligibility tests

utilizing a single message set of 16 words. There were also two intelligibility tests of 20 items with each of four of the message sets containing four words. The first day of each three-day period was considered as practice, and the results of the following two days were taken as data. Over these two days, for each message set there was one test at each of the four speech-to-noise ratios. The order of the tests was counterbalanced over the entire experiment. After each 80 stimulus presentations (either a single test with m = 16 or four tests with m = 4), the listeners scored the tests, and they calculated their a posteriori probability for each rating.

Results

Figure 5 indicates the degree to which the listeners were able to estimate a posteriori probabilities. In constructing this figure, all

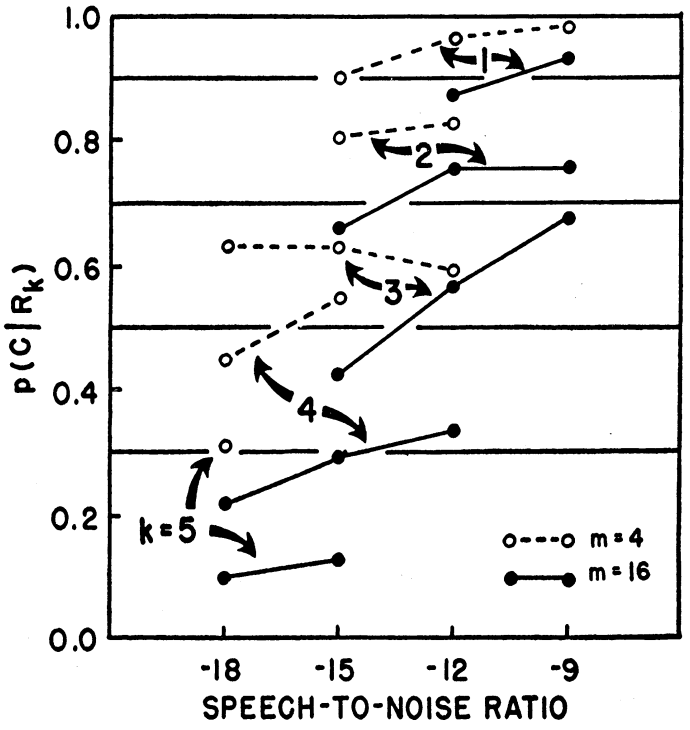

FIG. 5. Average obtained a posteriori probabilities for identification responses given confidence ratings of 1 through 5, as a function of size of message set and speech-to-noise ratio. The solid horizontal lines indicate the ranges of a posteriori probabilities for which the observers were instructed to use the rating categories.

data were treated as though from a single listener. This was necessary in order the the conditional probabilities for some of the lesser used ratings could be based on a large enough sample to assure a reasonable degree of reliability. Inherent in this averaging procedure is the danger of misrepresenting data for individuals. However, each listener's data were also analyzed separately and the average data are representative of their performance. Proportions having a denominator of less than 100 are not plotted in Fig. 5.

Despite changing speech-to-noise ratios and changing size of message sets from test to test, the listeners were able to do a fairly good job of partitioning their identification responses as instructed. By and large, the points for the various rating categories tend to fall within or very close to the boundaries which the listeners were instructed to use. These boundaries are represented by the solid horizontal lines in Fig. 5. However, comparison of their data with those reported in Fig. 1 shows that the listeners were able to do much better when speech-to-noise ratio and size of message set were not varied. As would be expected, at the high speech-to-noise ratios the listeners limited themselves primarily to high estimates of a posteriori probabilities and for low speech-to-noise ratios their estimates of a posteriori probabilities were generally low. This is clear from Fig. 5, when it is recalled that points are only plotted when the denominator is equal to or greater than 100.

It has been shown by many investigators that for many types of psychophysical judgments observers have a tendency to make their judgments relative to the range of stimuli encountered in the experiment.(Ref. 19) In view of this fact it is reasonable to expect that the listeners would tend to rate the "intelligibility" of any given item relative to the range of

"intelligibilities" encountered within any particular test. Such a tendency would account for the fact that the curves showing the relation between $p(C|R_k)$ and speech-to-noise ratio have for the most part, positive slopes. This "adaptation-level" factor would also be consistent with the observed fact that, for the larger message set relative to the smaller, the listeners overestimate their probability of being correct on their identification response. Apparently, there is an interaction between estimated a posteriori probabilities and both speech-to-noise ratio and size of message set. This interaction should not obscure the main finding that the listeners were fairly capable of estimating a posteriori probabilities over the entire range of conditions tested. It should be noted that Pollack and Decker (Ref. 5) did not find an interaction between the a posteriori probabilities associated with the listeners' ratings and speech-to-noise ratio.

Figures 6 and 7 show the obtained Type II receiver operating characteristics. The curves passing through the data would be straight lines if plotted on normal-normal probability paper. Thus, these curves may be characterized by the two parameters $(d_e)^{\frac{1}{2}}$ and s. There was no systematic variation in the slopes of these lines as a function of either message size or speech-to-noise ratio, and a slope of 0.9 was taken as representative of all of the data. However, $(d_e)^{\frac{1}{2}}$ appears to be a function of both of these variables as is seen most clearly in Fig. 8. These data suggest that at the lower speech-to-noise ratios the size of the message set is not an important variable in affecting $(d_e)^{\frac{1}{2}}$. Of course, when the speech-to-noise ratio is equal to minus infinity, $(d_e)^{\frac{1}{2}}$ must equal zero, and the curves of Fig. 8 must converge. Thus, the data suggest that as the speech-to-noise ratio increases the curves relating $(d_e)^{\frac{1}{2}}$ to speech-to-noise ratio for all sizes of message set climb together and then fan out at high speech-to-noise ratios.

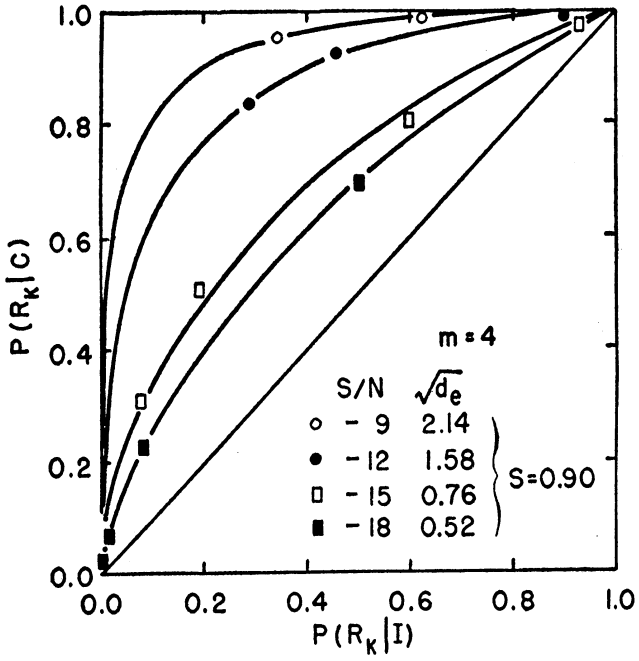

FIG. 6. Type II ROC curves for message sets of four items.

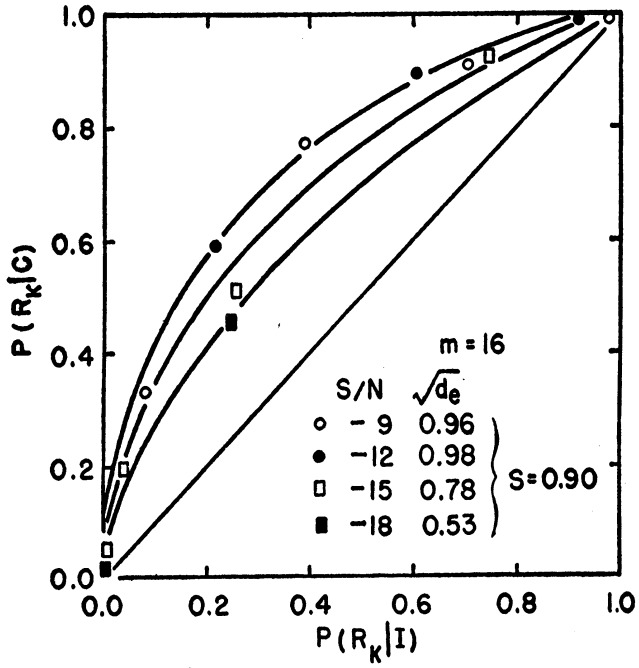

FIG. 7. Type II ROC curves for message sets of sixteen items.

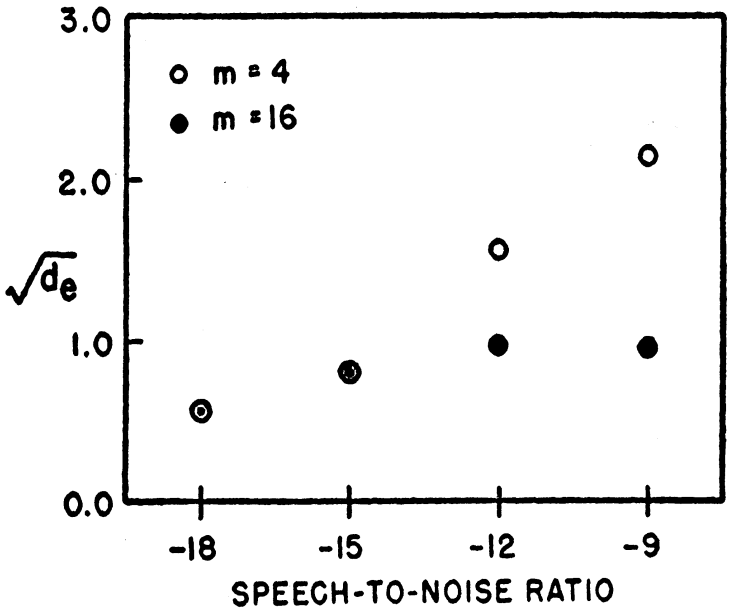

FIG. 8. The Type II ROC parameter $(d_e)^{\frac{1}{2}}$ as a function of size of message set and speech-to-noise ratio.

The P-10 Index

It has been shown (Refs. 4,20) that the intelligibility score is markedly affected by the number of alternatives from which the listener must choose his response on each stimulus presentation. This section will discuss some transformations of p(C) which were selected in an attempt to describe a family of intelligibility gain functions (with m as a parameter) by a single function. The P-10 index is suggested as a transformed score which shows promise of fulfilling this aim.

Because of its common use it is first necessary to examine the long-used "correction for guessing." As its name implies (and its formula verifies), this transformation of obtained proportions is based on a threshold model. This model states that a listener either "hears an item" in which case he identifies it correctly, or he "hears nothing" in which case he merely guesses at the identity of the item. It is further

assumed that if the listener must guess, each of the m alternatives in the set has the probability 1/m of being selected. With this model, the obtained proportion of correct responses with a set of size m, p_m, may be transformed to p_T, the proportion of instances in which the listener "heard the item." This transformation is accomplished by the formula

$$p_T = \frac{mp_m - 1}{m - 1} \qquad (1)$$

The degree to which this formula leads to a transformation of p_m which is independent of m may be evaluated by considering the intelligibility gain functions obtained from the experiment reported in the previous section. Figure 9 shows these gain functions. Scores are averaged over all five listeners; the curve for m = 16 represents data

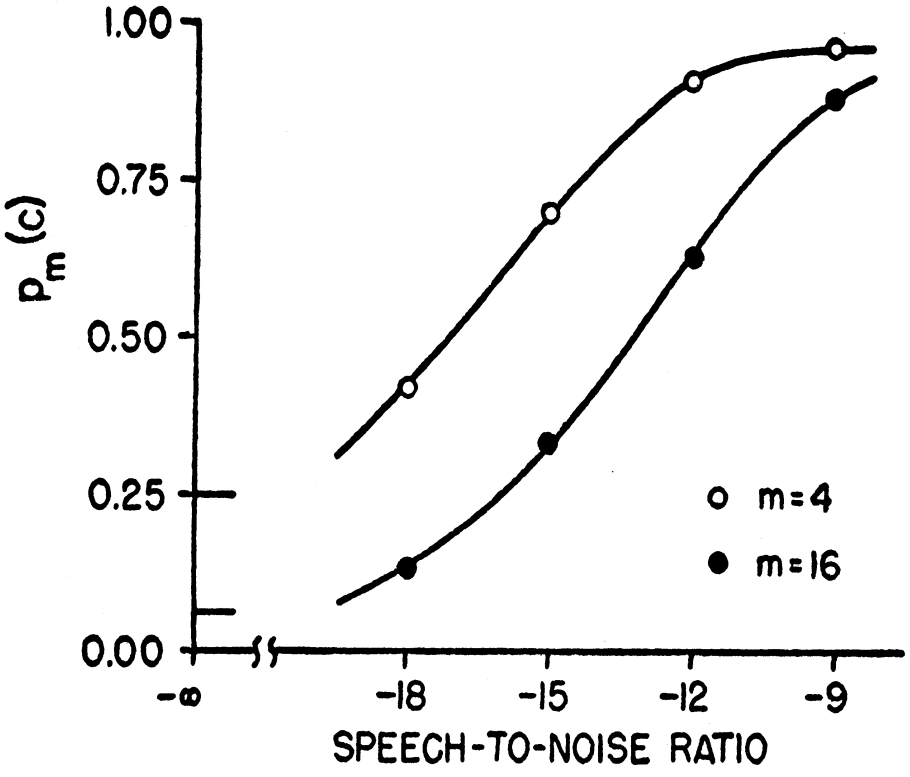

FIG. 9. Intelligibility gain functions for message sets of four items and for message sets of sixteen items.

pooled for the four message sets of 16 words each, while the curve for m = 4 is based on averages over the 16 message sets of four words each. When the obtained proportions of Fig. 9 are transformed by Eq. (1) the functions illustrated in Fig. 10 are obtained. Although the "correction for guessing" does reduce the separation between the functions of Fig. 9, it is clear that the two sets of points cannot be described by a single curve.

The constant-ratio rule (Refs. 2,3) is an empirical rule which states that when a subject is attempting to identify unordered stimuli in a background of noise, the ratio between any pair of entries in a row of the confusion matrix for some particular set of items will equal the ratio between the corresponding pair of entries in the confusion matrix for any subset of these items. Thus, for a message set of size m_1 whose confusion matrix is known, it is possible to predict the con-

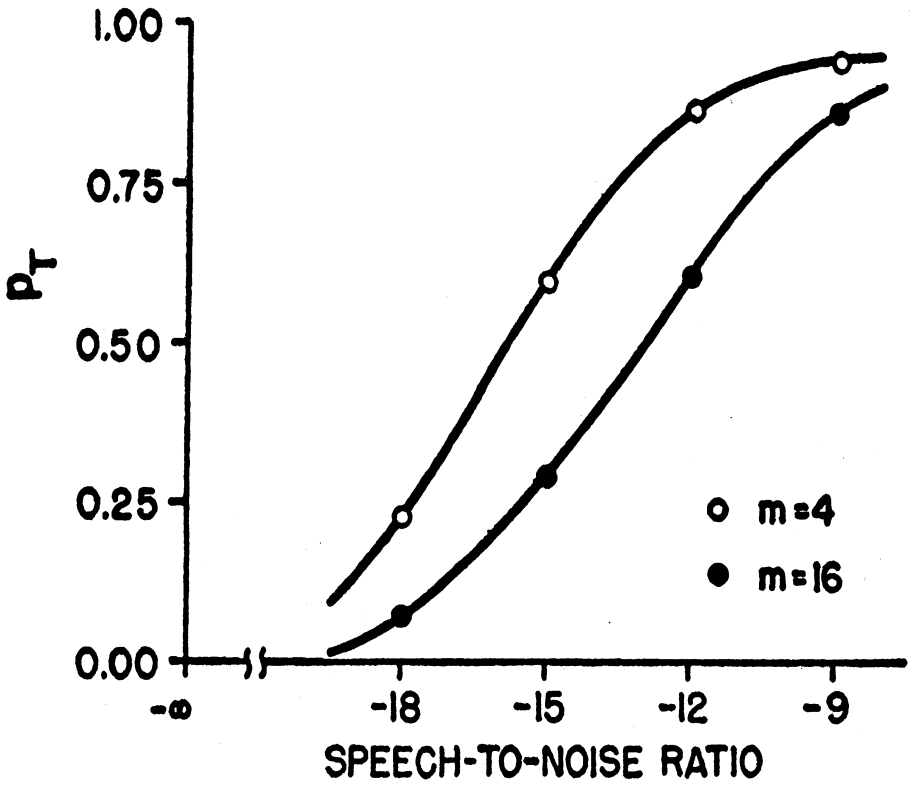

FIG. 10. The intelligibility gain functions of Fig. 9 after applying the traditional "correction for guessing" based on a simple threshold model.

fusion matrix (which yields the intelligibility score) for any subset of items. Egan (Ref. 10) suggested that the average intelligibility score for a number of possible subsets of size m_2 drawn from a master set of size m_1 may be approximated without having detailed knowledge of the confusion matrix for the master set. This is accomplished by assuming a uniform confusion matrix for the master set of items. This assumed uniform matrix has all diagonal entries equal to the articulation score for the master set, $p_{m1}(C)$, and all of the off-diagonal entries equal to $(1-P_{m1})/(m_1-1)$. Then, with such a matrix, regardless of the particular subset of size m_2 chosen from this master set, the constant-ratio rule predicts that the subset will have an articulation score given by the formula $p_{m2} = \frac{(m_1 - 1)p_{m1}}{m_2 - 1 + (m_1 - m_2)p_{m1}}$ (2). Egan pointed out that this approximation may be very poor for predicting the articulation score for any particular subset of size m_2. However, he has presented calculations which suggest that it may be an excellent estimate of the average articulation score for a large sample of subsets of size m_2.

This suggests that average articulation scores for various sizes of message sets could be equated by computing the articulation score that would have been expected had message sets of some particular size, say m = 10, been used. We shall denote this transformed score as P-10. Thus, from Eq. (2) we obtain $P\text{-}10 = \frac{(m - 1)p_m}{9 + (m - 10)p_m}$ (3). In this formula m may take any value; it need not be less than 10. The term p_m is the average articulation score obtained with a large sample of message sets, each of size m. P-10 is assumed to be the average articulation score which would have been obtained had a large sample of message sets of size m = 10 been drawn from the same population of messages.

Applying this transformation to the intelligibility scores of Fig. 9, we obtain the single function seen in Fig. 11. Whereas this result looks very promising, a note of caution must be expressed. When applied to the

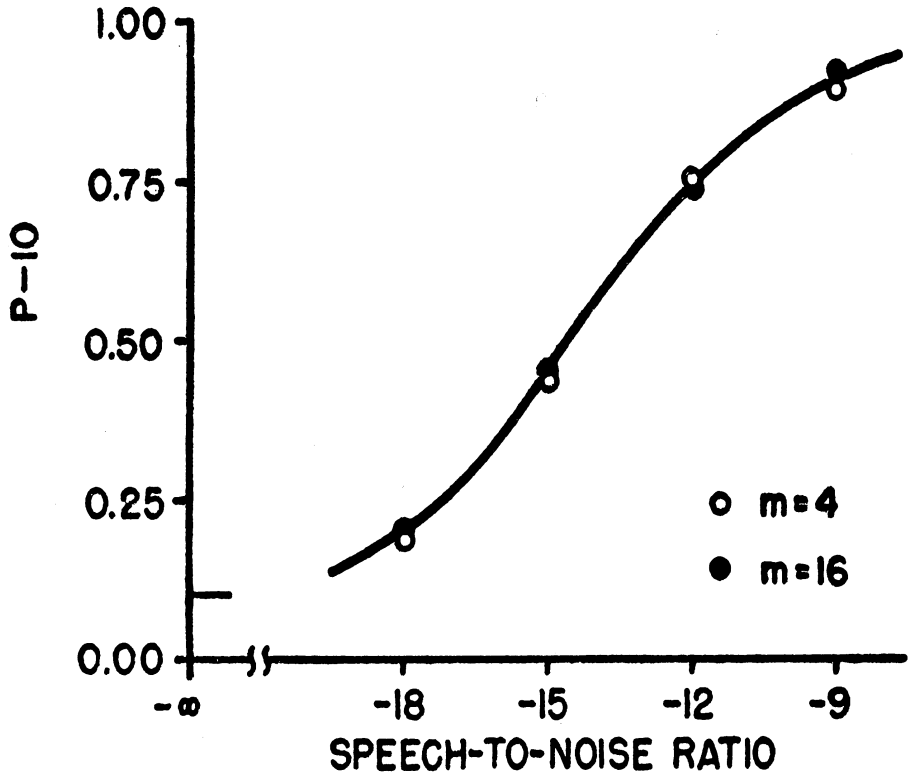

FIG. 11. The intelligibility gain functions of Fig. 9 after applying a transformation based on the constant-ratio rule.

data obtained by Miller, Heise, and Lichten (Ref. 20), the P-10 index does not do such an excellent job of reducing their family of curves to a single function. As the constant-ratio rule has been tested primarily with listeners who were highly practiced with the particular sets of messages under test, it may be that the P-10 index is only applicable when these conditions prevail.

Green and Birdsall (Ref. 21), on the basis of a theory of signal detectability, have proposed a set of transformations of the obtained data designed to yield a single function relating an inferred variable, d', to speech-to-noise ratio. For a detailed discussion of this model, the reader is referred to their technical report. The theory of signal detectability incorporating their simplifying assumptions does a somewhat less adequate job of reducing the data of Fig. 9 to a single function than does the P-10 index, although it does a somewhat better job of handling the Miller, Heise, and Lichten data.

V. PREDICTION OF MOC CURVES FROM ROC DATA

The monitoring task may be defined as follows (Ref. 22). A listener is provided with a subset of messages drawn from a known set of messages. On each stimulus presentation the listener's task is to state whether or not the transmitted item is one of the messages in the subset to be monitored. As defined here, the monitor's task is not to identify specifically the transmitted item; he must merely state whether the transmitted message is a member of the subset to be monitored, or whether it is a member of the complementary subset. Here, as in the identification task, the listener may vary the criterion under which he will accept a transmitted message as a message of the subset to be monitored.

Clearly, the identification task and the monitoring task are closely related. If a listener is willing to accept his identification response as being correct and is very confident that he is correct, then, if this message is one of the subset to be monitored, he should be very confident that the transmitted message was, in fact, a message of the subset to be monitored. Thus, in this instance, the listener should be willing to accept, with a high degree of confidence, the hypothesis that the transmitted message was a monitored message. If, on the other hand, the identification response made with a high degree of confidence is not of the subset to be monitored, then the listener should accept the hypothesis that the transmitted item was of the set to be monitored only with a very low degree of confidence. This reasoning provides a mechanism for predicting a monitor's (MOC curve) from an ROC curve. That this argument is an oversimplified one is made clear by the following example. Suppose that the subset of messages to be monitored consists of two messages A_1 and A_2. In the identification task, the listener on some particular stimulus presentation may not have much confidence in his identification response because, although he has great confidence that the message was either A_1 or A_2, he may not have the necessary information to decide which of the two was in fact transmitted. In this instance, then, the listener would have relatively low confidence in his identification response but high confidence in assigning the transmitted stimulus to the subset to be monitored. The above described method of moving from the ROC curve to the MOC curve does not take account of such occurrences.

Procedure

Five listeners were used in this experiment. None of these observers had served in any of the experiments previously described. The general experimental conditions matched those described in section II as closely

as possible. The same message set of five spondees and the same talker were used for both experiments. In this experiment the listener's task on each stimulus presentation was to accept or reject the hypothesis that the transmitted message was a message of the set to be monitored. On each test of 100 items the listeners attempted to maintain a constant criterion. Tests were run at four different criterion levels which were merely described by the labels "very strict," "strict," "medium," and "lax." Data were obtained following 33 practice tests. The listeners always scored their own tests. On any given monitor test, the subset of items to be monitored consisted of two words: either greyhound and stairway, or eggplant and vampire. These will be called MI and MII tests, respectively. Data were obtained on twelve MI tests and twelve MII tests, three of each conducted under each of the four criterion levels. Twelve standard intelligibility tests were also conducted. The various types of tests were presented in a counter-balanced manner over a period of six days. In all test lists,each of the items occurred twenty times. The method of prediction was as follows. Consider the MI tests. The two items GH and SW are to be monitored. The data of the listeners in section II can be used to plot theoretical MOC curves for these two M items. Consider the data for L1 in Table 1. Utilizing the assumptions mentioned above, we can compute for each level of confidence the proportion of instances in which L1 responded either GH or SW given that the transmitted stimulus was either GH or SW. We may also compute the proportion of instances in which L1 responded with either GH or SW given that one of the other three messages was transmitted. These proportions may then be cumulated from the strictest to the most lax confidence level. These cumulated proportions for the various confidence levels may serve as estimates of $p(Y|M)$ and $p(Y|S)$. Here, $p(Y|M)$

is the probability of the listener accepting the hypothesis that the transmitted message was one from the subset to be monitored given that it was from this subset, and $p(Y|S)$ is the probability of the listener accepting the hypothesis that the transmitted message was one from the subset to be monitored when it was one of the secondary messages.

The derived points for these theoretical MOC curves were plotted on normal-normal probability paper for each of the listeners in the experiment of section II. A single straight line (visual fit) was drawn through these points. These two curves, one for each of the two subsets to be monitored, serve as the predicted MOC curves for the experiment reported here. As pointed out above, these predicted curves were based upon a different crew of listeners than those used in the present experiment.

Results

Figures 12 and 13 shows the predicted curves and the obtained data points (solid dots). In both cases the obtained value of $\underline{s}$ differs from the predicted value. In Fig. 12 the obtained value of $\underline{s}$ is 1.02 while the predicted value is 0.73. In Fig. 13 the obtained and predicted values of $\underline{s}$ are 0.84 and 1.06, respectively. The predicted MOC curve in Fig. 12 has a $(d_e)^{\frac{1}{2}}$ of 1.68 while the obtained $(d_e)^{\frac{1}{2}}$ is 1.45. Things are somewhat better in Fig. 13 where the predicted and obtained values of $(d_e)^{\frac{1}{2}}$ are 1.24 and 1.22, respectively.

The observed discrepancies do not necessarily mean that the suggested model is inadequate. It could be that the other differences between the two experiments (different sets of listeners employed, slight changes in the talker's manner of pronouncing the words, etc.) led to the discrepancies. This problem of interpreting discrepant results was anticipated,

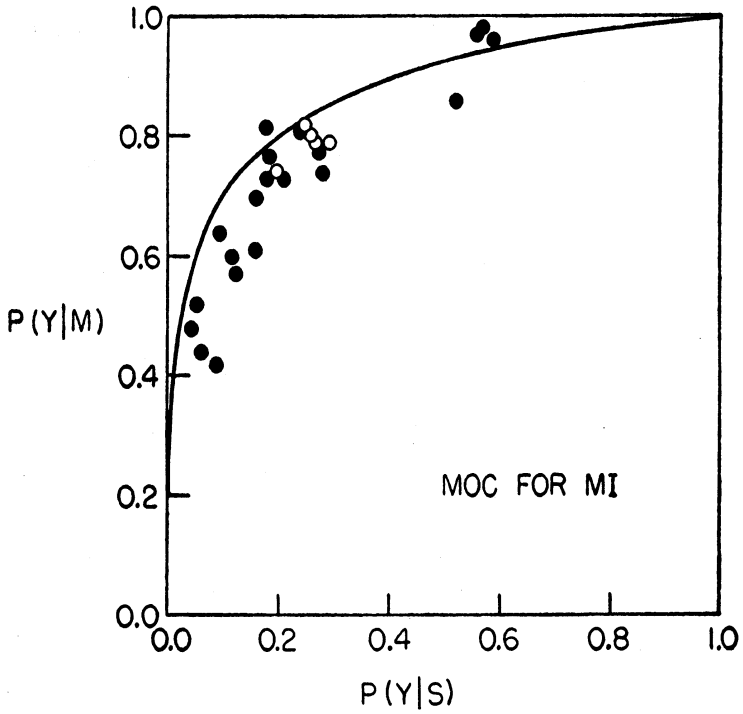

FIG. 12. The solid line shows the predicted MOC curve for the MI monitoring tests. This predicted curve is based on an earlier experiment. The solid points show the obtained data. The open circles are points predicted from articulation tests given in this experiment.

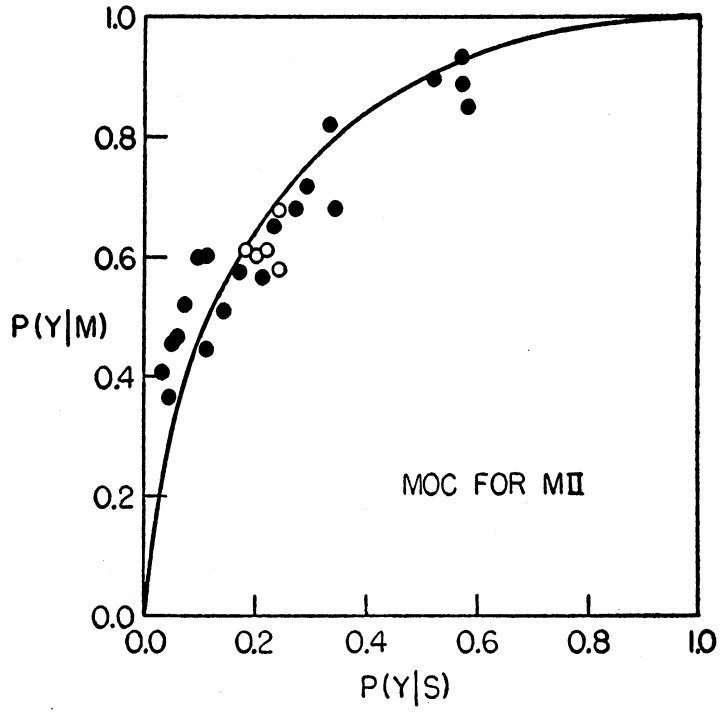

FIG. 13. The predicted MOC curve and obtained data for the MII monitoring tests.

and the articulation tests were conducted in order to partially discriminate between these two possibilities. The confusion matrices obtained by these articulation tests enable us to predict a single point on the MOC curve for each subject. If these points fall on or near the MOC curves predicted from data of the previous experiment, then it would appear that the proposed method of predicting MOC curves from ROC data is inadequate. If, on the other hand, these points fall in with the obtained MOC data, it would appear that the method of prediction is adequate at least for predicting $(d_e)^{\frac{1}{2}}$, though possibly not for $\underline{s}$. The open circles in Fig. 12 and Fig. 13 are MOC points predicted from the articulation tests of the current experiment. These lead us to the conclusion that the method of prediction is fairly adequate [at least for predicting $(d_e)^{\frac{1}{2}}$] but that uncontrolled differences in the two experiments led to the failure of prediction from one to the other. The most notable difference (one which it was impractical to control) is the very different level of training for the two groups of listeners.

APPENDIX I

PREDICTION OF SECOND-CHOICE RESPONSES BY A THRESHOLD MODEL

Threshold models may be as simple or as complicated as one wishes. The basic assumption is that some unit of an input to a sensory system is either above or below some hypothetical sensory threshold. If above threshold, the input unit may be identified by the organism with zero probability of error. If below threshold, no information about the unit is transmitted, and the observer has nothing more than his assumed a priori probabilities to guide him in attempting to identify the input unit.

Such a general statement of a trheshold model does not lead to any quantitative predictions. Typically, when dealing with the transmission of words to a listener, the following assumptions are added. (1) It is assumed that the input unit is the complete waveform of the word. (2) It is assumed that each word in the message ensemble has the same probability of being above threshold. (3) It is assumed that if a transmitted word is below threshold, the listener will respond in a perfectly random manner with one of the words in the message set (each with probability $1/m$). (4) To deal with second-choice behavior, we may add the assumption that the listener's second choice will be selected in a random fashion from the $m-1$ words in the message ensemble [each with probability $1/(m-1)$] which remain after the first choice is made.

To examine the consequences of these assumptions, let us consider the data for listener L1 which was obtained with message set 2. This listener's stimulus by first-choice matrix and stimulus by second-choice matrix are given in Table I-1.

Table I-1

Stimulus	First Choice DP	EP	GH	SW	VP	Second Choice DP	EP	GH	SW	VP
DP	.593	.096	.059	.084	.168	.227	.365	.180	.037	.190
EP	.180	.538	.072	.091	.119	.506	.180	.168	.059	.086
GH	.086	.207	.323	.247	.136	.341	.222	.215	.099	.123
SW	.017	.047	.104	.795	.037	.096	.363	.368	.049	.123
VP	.180	.104	.077	.094	.546	.402	.036	.267	.037	.158

Utilizing assumptions one, two, and three above, and the stimulus by first-choice matrix we may obtain an estimate of the probability that a word is above threshold. Taking the average of the diagonal entries in the stimulus by first-choice matrix we obtain as an estimate

$$p + (1 - p)/m = 0.559, \text{ or } p = 0.449.$$

Thus, with this simple threshold model we would estimate that, except for sampling error, the stimulus by first-choice matrix would have diagonal entries equal to $p + (1 - p)/m = 0.559$, and off-diagonal entries all equal to $(1 - p)/m = 0.110$. As the entries in Table I-1 are based on 405 responses per stimulus item, this theoretical matrix is clearly unreasonable. Utilizing the fourth assumption listed above, we would predict that the stimulus by second-choice matrix would have diagonal entries of 0.110 and off-diagonal entries of 0.2225. Considering the poor agreement of the obtained and the theoretical stimulus by first-choice matrix, it is not surprizing that this prediction is not borne out.

By dropping the second and third assumptions listed above, we may obtain a theoretical stimulus by first-choice matrix which does not appear too unreasonable. Specifically, we will now allow each of the words to have a different probability of being above threshold. Thus, for any

particular one of the words in the message set let us assume a probability, P_i, that this word will be above threshold when presented to the listener. Also, we will designate as p_i the probability that the listener will respond with the i^{th} word given that the transmitted item is below threshold. We shall relax the fourth assumption listed above, by assuming that if the listener responds with the j^{th} word in the list as his first choice, his probability of responding with the k^{th} word as a second choice is given by $\frac{p_k}{\sum_{i,i\neq j} p_i}$.

Consider the stimulus by first-choice matrix. Utilizing these new assumptions we have,

$$p(a_j|A_j) = P_j + (1-P_j)p_j \tag{1}$$

and

$$p(a_j|A_i) = (1-P_i)p_j,\ i \neq j. \tag{2}$$

There are various ways in which we might utilize all of the values in the stimulus by first-choice matrix to obtain estimates of the hypothesized probabilities. The following is one of the simpler.[1]

Let $(1-P_i) = 1/Q_i$. Then we may rewrite Eqs. (1) and Eqs. (2) as

$$p_j = Q_j[p(a_j|A_j) - 1] + 1 \tag{1'}$$

$$p_j = p(a_j|A_i)\ Q_i,\ i \neq j \tag{2'}$$

And summing each of Eqs. (2') over all i, $i \neq j$, we obtain

$$(m-1)p_j = \sum_{i,i\neq j} p(a_j|A_i)\ Q_i \tag{2''}$$

Combining each of Eqs. (1') and Eqs. (2'') we obtain

$$\sum_{i,i\neq j} p(a_j|A_i)\ Q_i + (m-1)\ Q_j\ [1 - p(a_j|A_j)] - m + 1 = 0 \tag{3}$$

---

[1] The particular form of this solution was suggested by reading an unpublished paper by Dr. Cletus J. Burke.

Thus, we have $\underline{m}$ simultaneous equations in $\underline{m}$ unknowns. From the $\underline{m}$ values of Q_i we can directly obtain the values of p_i and of P_i.

Solving this set of equations utilizing the data in Table I-1, we obtain the following values:

$P_1 = 0.492$ $\quad$ $p_1 = 0.198$

$P_2 = 0.422$ $\quad$ $p_2 = 0.200$

$P_3 = 0.158$ $\quad$ $p_3 = 0.196$

$P_4 = 0.745$ $\quad$ $p_4 = 0.195$

$P_5 = 0.425$ $\quad$ $p_5 = 0.211$

Substituting these values in Eqs. (1) and Eqs. (2), we obtain the theoretical stimulus by first-choice matrix shown in Table I-2.

Table I-2

		First Choice				
		DP	EP	GH	SW	VP
Stimulus	DP	.593	.102	.099	.099	.107
	EP	.115	.538	.113	.113	.122
	GH	.167	.168	.323	.164	.177
	SW	.051	.051	.050	.795	.054
	VP	.114	.115	.113	.112	.546

Before going on to the stimulus by second-choice matrix, it may be noted that the coincidence of the diagonal entries in the theoretical matrix and the diagonal entries in the obtained matrix is no tribute to the threshold model. This was forced by the particular manner in estimating the P_i's and the p_i's above. Also, the almost uniform distribution of the off-diagonal entries in each row of the theoretical matrix is not typical of

this approach, but is merely the result with this particular set of data.

The theoretical stimulus by second-choice matrix presented below was calculated using the values in the theoretical stimulus by first-choice matrix and the estimated parameters above. It was assumed that if the listener gave the j^{th} item as his first choice, his probability of giving the k^{th} item as his second response is given by $\frac{p_k}{\sum_{i,i\neq j} p_i}$

where $i \neq j$, and his probability is zero where $i = j$.

It will be seen that the theoretical stimulus by second-choice matrix of Table I-3 bears little resemblance to the obtained stimulus by second-choice matrix. This is also true when data from other observers are examined.

Table I-3

		Second Choice				
		DP	EP	GH	SW	VP
Stimulus	DP	.100	.225	.221	.219	.235
	EP	.220	.116	.217	.216	.232
	GH	.205	.208	.166	.204	.216
	SW	.235	.236	.230	.049	.247
	VP	.222	.222	.218	.218	.119

We have seen that the threshold model, both in its most common form and in a somewhat more sophisticated form, fails to account for second-choice data. If a threshold model is to handle such data (as well as that presented in section II) it would appear that the unit of analysis must be more elemental than the entire waveform of the word. There are several ways of doing

this, all of which require entire new sets of arbitrary assumptions. Further research in speech analysis may suggest a reasonable approach to a threshold model.

APPENDIX II

PREDICTION OF SECOND-CHOICE RESPONSES UTILIZING AN EXTENSION OF THE CONSTANT-RATIO RULE

Given knowledge of a listener's confusion matrix for some set of messages, the constant-ratio rule enables one to predict the listener's confusion matrix for any particular subset of these messages. This fact suggests a possible approach to predicting a listener's second-choice behavior. The listener's first choice is a selection from one of m alternatives. His second choice must then be selected from one of the remaining m-1 alternatives. If the listener's memory for the waveform of the signal in the background of noise is good, and if his first choice influences his second choice solely by limiting the number of alternatives available to him for a second choice, then it would be expected that the constant-ratio rule could be used to predict the listener's stimulus by second-choice matrix. By direct application of the constant-ratio rule and the above assumption, we have

$$p(b_k|A_i) = p(a_k|A_i) \sum_{j \neq k} \frac{p(a_j|A_i)}{1-p(a_j|A_i)} . \qquad \text{(II-1)}$$

Where $p(b_k|A_i)$ is the probability of the k^{th} identification response on the second-choice given the i^{th} stimulus item. The probability of the k^{th} identification response on the first-choice is given by $p(a_k|A_i)$.

A comparison of the obtained stimulus by second-choice matrix and the predicted matrix for L1 is given in Table II-1. The parenthetical

entries are the predicted proportions. It is clear from this table that at least one of the above two assumptions is not justified.

Table II-1

		Second Choice				
		DP	EP	GH	SW	VP
Stimulus	DP	0.227 (0.274)	0.365 (0.174)	0.180 (0.110)	0.037 (0.154)	0.190 (0.289)
	EP	0.506 (0.266)	0.180 (0.286)	0.168 (0.117)	0.059 (0.145)	0.086 (0.186)
	GH	0.341 (0.105)	0.222 (0.219)	0.215 (0.272)	0.099 (0.244)	0.123 (0.158)
	SW	0.096 (0.069)	0.363 (0.190)	0.368 (0.414)	0.049 (0.176)	0.123 (0.150)
	VP	0.402 (0.271)	0.136 (0.167)	0.267 (0.126)	0.037 (0.152)	0.158 (0.285)

APPENDIX III

PREDICTION OF SECOND-CHOICE RESPONSES BY A SIMPLIFIED THEORY OF SIGNAL DETECTABILITY

The theory of signal detectability may be used to make specific predictions of the entries in a stimulus by second-choice matrix given the stimulus by first-choice matrix. Unfortunately, for these non-orthogonal signals, the calculations for making such predictions are prohibitive when m is greater than two. However, with a simplified model, it may be possible to make a prediction of the average probability of being correct on the second-choice given the probability of being correct on the first-choice.

In this simplified model it is assumed that the m signals are orthogonal to one another and all of equal energy. At least one of these assumptions is unwarranted, for this simplified model predicts a uniform stimulus by first-choice matrix. However, it was felt that the inaccuracies introduced by these assumptions might average out and that this model might lead to adequate predictions of the average probability of being correct on the second identification response.

The theory of signal detectability in the case of five orthogonal signals of equal energy states that

$$p_5(C_1) = \int_{\infty}^{+\infty} f(x-d')[F(x)]^4 \, dx \qquad \text{(III-1)}$$

where $f(x)$ is a normally distributed random variable of unit variance, and $F(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-\frac{t^2}{2}} dt$. Thus, Eq. (III-1) gives the probability that a single draw from a normal distribution with mean d' and unit variance is larger than each of four draws from a normal distribution with mean zero and unit variance. Similarly,

$$p_5(C_2) = 4 \int_{-\infty}^{+\infty} f(x-d')[1-F(x)][F(x)]^3 \, dx \qquad \text{(III-2)}$$

Consequently,

$$p_5(C_2) = 4[p_4(C_1) - p_5(C_1)] \qquad \text{(III-3)}$$

Birdsall and Peterson (Ref.23), using approximate integration techniques, have solved Eq. (III-1) for various values of d' and m. They concluded that Eq. (III-1) is closely approximated by

$$p_m(C_1) = F(a_m d' - b_m) \qquad \text{(III-4)}$$

where $F(x)$ is the (area) normal distribution function, i.e.,

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

The value of b_m is obtained by setting $d' = 0$ in which case $p_m(C_1) = F(-b_m) = 1/m$. On the basis of their approximate integrations of Eq. (III-1), Birdsall and Peterson computed a_m for several values of m.

Substituting appropriate values of a_m and b_m in Eq. (III-4) we obtain

$$p_5(C_1) = F(0.835\ d' - 0.842) \qquad \text{(III-5)}$$

and

$$p_4(C_1) = F(0.827\ d' - 0.674) \qquad \text{(III-6)}$$

To predict $p_5(C_2)$ from knowledge of $p_5(C_1)$, it is only necessary to solve Eq. (III-5) for d' and then solve Eq. (III-6) for $p_4(C_1)$. Then, by Eq. (III-3) we may obtain the predicted value for $p_5(C_2)$.

Predictions using this procedure are compared with data taken from section III in the following table.

	$\bar{p}(C_1)$	$\bar{p}(C_2)$	$\overset{\Delta}{\bar{p}}(C_2)$
L_1	0.559	0.166	0.244
L_2	0.667	0.110	0.216
L_3	0.713	0.142	0.196
L_4	0.487	0.179	0.260
L_5	0.567	0.175	0.244
L_6	0.580	0.200	0.244

Here, $\bar{p}(C_1)$ is the average probability of being correct on the first choice, $\bar{p}(C_2)$ is the average probability of being correct on the second choice, and $\overset{\Delta}{\bar{p}}(C_2)$ is the predicted value of $\bar{p}(C_2)$.

REFERENCES

1. G. A. Miller and P. E. Nicely. J. Acoust. Soc. Am., 27, 338, 1955.
2. F. R. Clarke. J. Acoust. Soc. Am., 29, 715, 1957.
3. F. R. Clarke and C. D. Anderson. J. Acoust. Soc. Am., 29, 1318, 1957.
4. J. P. Egan and F. R. Clarke. J. Acoust. Soc. Am., 28, 1267, 1956.
5. I. Pollack and L. R. Decker. J. Acoust. Soc. Am., 30, 286, 1958.
6. C. E. Shannon and W. Weaver. The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill., 1949.
7. G. A. Miller. Information Theory in Psychology (ed. H. Quastler), Free Press, Glencoe, Ill., 1955.
8. J. P. Egan, F. R. Clarke and E. C. Carterette. J. Acoust. Soc. Am., 28, 536, 1956.
9. E. C. Carterette. J. Acoust. Soc. Am., 30, 846, 1958.
10. J. P. Egan. Message Repetition, Operating Characteristics, and Confusion Matrices in Speech Communication, Hearing and Communication Laboratory, AFCRC-TR-57-50, 1957.
11. F. R. Clarke, T. G. Birdsall and W. P. Tanner, Jr. J. Acoust. Soc. Am., 31, 629, 1959.
12. I. Pollack. J. Acoust. Soc. Am., 31, 1031, 1959
13. C. E. Shannon. Bell Sys. Tech.,J., 30, 50, 1951.
14. A. A. Blank and H. Quastler. Notes on the Estimation of Information Measures, Control Systems Laboratory, University of Illinois, R-56, 1954.
15. J. Swets, W. P. Tanner, Jr., and T. G. Birdsall. The Evidence for a Decision-Making Theory of Visual Detection, Electronic Defense Group, The Univ. of Mich., TR-40, 1955.

16. I. Pollack. Assimilation of Sequentially-Encoded Information: IV, HRRL Memo. Report No. 25, 1952.

17. P. D. Bricker and A. Chapanis, Psychol. Rev., 60, 181, 1953.

18. J. P. Egan. Articulation Testing Methods II, Office of Scientific Research and Development, Report No. 3802, 1944.

19. H. Helson. Amer. J. Psychol., 60, 1, 1947.

20. Miller, Heise, and Lichten. J. Exper. Psychol., 41, 329, 1951.

21. D. M. Green and T. G. Birdsall. The Effect of Vocabulary Size on Articulation Score, Electronic Defense Group, The Univ. of Michigan, TR-81, 1958.

22. J. P. Egan. J. Acoust. Soc. Am., 29, 482, 1957.

23. T. G. Birdsall and W. W. Peterson. Probability of Correct Decision: In a Forced Choice Among M Alternatives, Electronic Defense Group, The University of Michigan, QPR No. 10, 1954.

Decision Processes in Perception[1]

John A. Swets, Massachusetts Institute of Technology

Wilson P. Tanner, Jr. and Theodore G. Birdsall, University of Michigan

About five years ago, the theory of statistical decision (Ref. 1) was translated into a theory of signal detection. (Ref. 2) Although the translation was motivated by problems in radar, the detection theory that resulted is a general theory, for like the decision theory, it specifies an ideal process. The generality of the theory suggested to us that it might also be relevant to the detection of signals by human observers. Beyond this, we were struck by several analogies between this description of ideal behavior and various aspects of the perceptual process. It seemed to us to provide a framework for a realistic description of the behavior of the human observer in a variety of perceptual tasks.

Part I of this paper begins with a brief review of the theory of statistical decision and then presents a description of the elements of the theory of signal detection appropriate to human observers. Part II reports the results of some experimental tests of the applicability of the theory to the detection of visual signals.

---

[1] This paper is based upon Technical Report No. 40, issued by the Electronic Defense Group of The University of Michigan in 1955. The studies it reports were supported by the Electronic Defense Group under a contract with the U. S. Army Signal Corps, and by the Vision Research Laboratory, also of The University of Michigan, under Contract Nos. Da-36-039-SC 52654 and Nobs 53242. Our thanks are due Dr. H. B. Blackwell and Dr. W. M. Kincaid of the Vision Research Laboratory for their assistance in these studies. The preparation of this paper was supported by the U. S. Air Force under Contract No. AF 19(604)-1728 monitored by the Operational Applications Laboratory of the Cambridge Research Center. We are indebted to Dr. D. H. Howes for many helpful suggestions concerning the presentation of this material.

The theory and some illustrative results of one experimental test of it were briefly described in an earlier paper (Ref. 3). The present paper contains a more nearly adequate description of the theory, a more complete account of the first experiment and the results of four other experiments. It brings together all of the data collected to date in vision experiments that bear directly on the value of the theory.[2]

I. THE THEORY

Statistical Decision Theory

Consider the following game of chance. Three dice are thrown. Two of the dice are ordinary dice. The third die is unusual in that on each of three of its sides it has 3 spots, whereas on its remaining three sides it has no spots at all. You, as the player of the game, do not observe the throws of the dice. You are simply informed, after each throw, of the total number of spots showing on the three dice. You are then asked to state whether the third die, the unusual one, showed a 3 or a 0. If you are correct - that is, if you assert a 3 showed when it did in fact, or if you assert a 0 showed when it did in fact - you win a dollar. If you are incorrect - that is, if you make either of the two types of error possible - you lose a dollar.

How do you play the game? Certainly you will want a few minutes to make some computations before you begin. You will want to know the probability of occurrence of each of the possible totals two through twelve in the event that the third die shows a 0, and you will want to

---

[2] Reports of several applications of the theory in audition experiments are available in the literature; for a list of references, see Ref. 4.

know the probability of occurrence of each of the possible totals five through fifteen in the event that the third die shows a 3. Let us ignore the exact values of these probabilities, and grant that the two probability distributions in question will look much like those sketched in Figure 1.

Realizing that you will play the game many times, you will want to establish a policy which defines the circumstances under which you will make each of the two decisions. We can think of this as a criterion or a cutoff point along the axis representing the total number of spots showing on the three dice. That is, you will want to choose a number on this axis such that whenever it is equalled or exceeded you will state that a 3 showed on the third die, and such that whenever the total number of spots showing is less than this number, you will state that a 0 showed on the third die. For the game as described, with the a priori probabilities of a 3 and a 0 equal, and with equal values and costs associated with the four possible decision outcomes, it is intuitively clear that the optimal cutoff point is that point where the two curves cross. You will maximize your winnings if you choose this point as the cutoff point and adhere to it.

Now what if the game is changed? What, for example, if the third die has 3 spots on five of its sides, and a 0 on only one? Certainly you will now be more willing to state, following each throw, that the third die showed a 3. You will not, however, simply state more often that a 3 occurred without regard to the total showing of the three dice. Rather, you will lower your cutoff point: you will accept a smaller total than before as representing a throw in which the third die showed a 3. Conversely, if the third die has 3 spots on only one of its sides and 0s on five sides, you will do well to raise your cut-

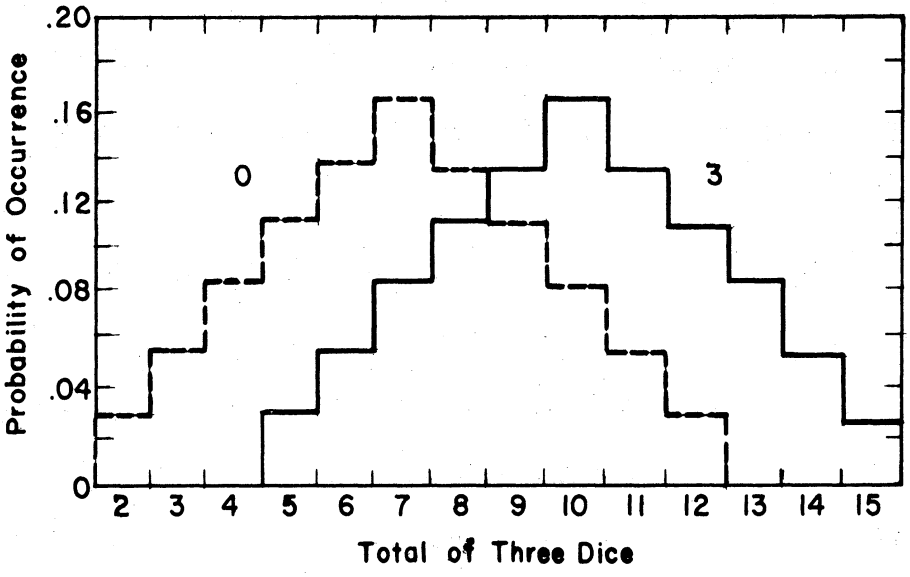

FIG. 1 THE PROBABILITY DISTRIBUTIONS FOR THE DICE GAME

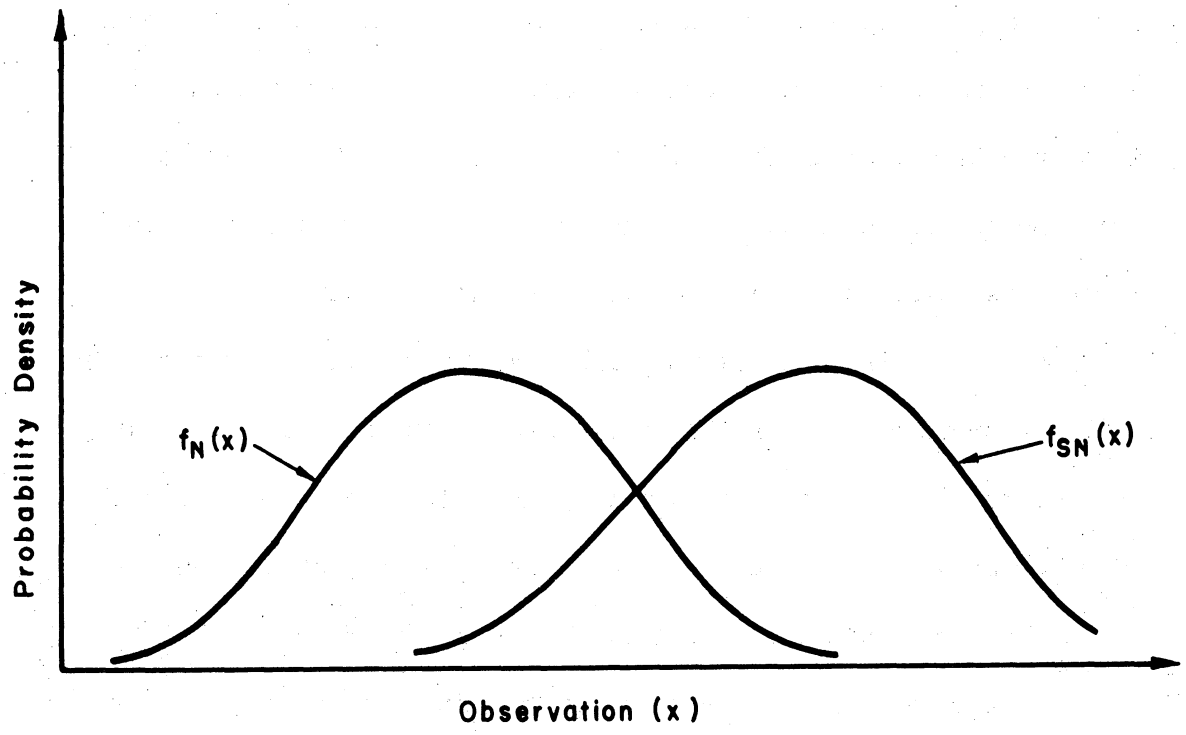

FIG. 2 THE PROBABILITY DENSITY FUNCTIONS OF NOISE AND SIGNAL PLUS NOISE

off point - to require a higher total then before for stating that a 3 occurred.

Similarly, your behavior will change if the values and cost associated with the various decision outcomes are changed. If it costs you five dollars every time you state that a 3 showed when in fact it did not, and if you win five dollars every time you state that a 0 showed when in fact it did (the other value and the other cost in the game remaining at one dollar), you will raise your cutoff to a point somewhere above the point where the two distributions cross. Or if, instead, the premium is placed on being correct when a 3 occurred, rather than when a 0 occurred as in the immediately preceding example, you will assume a cutoff somewhere below the point where the two distributions cross.

Again, your behavior will change if the amount of overlap of the two distributions is changed. You will assume a different cutoff than you did in the game as first described if the three sides of the third die showing spots now show 4 spots rather than 3.

This game is simply an example of the type of situation for which the theory of statistical decision was developed. It is intended only to recall the frame of reference of this theory. Statistical decision theory - or the special case of it of relevance here, i.e., the theory of testing statistical hypotheses - specifies the optimal behavior in a situation where one must choose between two alternative statistical hypotheses on the basis of an observed event. In particular, it specifies the optimal cutoff, along the continuum on which the observed events are arranged, as a function of (a) the a priori probabilities of the two hypotheses, (b) the values and costs associated with the various decision outcomes, and (c) the amount of overlap of the distri-

butions that constitute the hypotheses.

According to the mathematical theory of signal detectability, the problem of detecting signals that are weak relative to the background of interference is like the one faced by the player of our dice game. In short, the detection problem is a problem in statistical decision; it requires testing statistical hypotheses. In the theory of signal detectability, this analogy is developed in terms of an idealized observer. It is our thesis that this conception of the detection process may apply to the human observer as well. The next several pages present an analysis of the detection process that will make the bases for this reasoning apparent.[3]

The Fundamental Detection Problem

In the fundamental detection problem, an observation is made of events occurring in a fixed interval of time and a decision is made, based on this observation, whether the interval contained only the background interference or a signal as well. The interference, which is random, we shall refer to as noise and denote as N; the other alternative we shall term signal plus noise, SN. In the fundamental problem, only these two alternatives exist - noise is always present, whereas a signal may or may not be present during a specified observation interval. Actually, the observer, who has advance knowledge of the ensemble of signals to be presented, says either "yes, a signal was present," or

---

[3] It is to be expected that a theory recognized as having a potential application in psychophysics, although developed in another context, will be similar in many respects to previous conceptions in psychophysics. Although we shall not, in general, discuss explicitly these similarities, the strong relationship between many of the ideas presented in the following and Thurstone's earlier work on the scaling of judgments should be noted. (Refs. 5-6) The present theory also has much in common with the recent work of Smith and Wilson and Munson and Karlin. (Refs. 7-8) Of course, for a new theory to arouse interest, it must also differ in some significant aspects from previous theories - these differences will become apparent as we proceed.

"no, no signal was present" following each observation. In the experiments reported below, the signal consisted of a small spot of light flashed briefly in a known location on a uniformly illuminated background. It is important to note that the signal is always observed in a background of noise; some, as in the present case, may be introduced by the experimenter or by the external situation, but some is inherent in the sensory processes.

The Representation of Sensory Information

We shall, in the following, use the term observation to refer to the sensory datum on which the decision is based. We assume that this observation may be represented as varying continuously along a single dimension. Although there is no need to be concrete, it may be helpful to think of the observation as some measure of neural activity, perhaps as the number of impulses arriving at a given point in the cortex within a given time. We assume further that any observation may arise, with specific probabilities, either from noise alone or from signal plus noise. We may portray these assumptions graphically, for a signal of a given amplitude, as in Figure 2. The observation is labelled x and plotted on the abscissa. The lefthand distribution, labelled $f_N(x)$, represents the probability density that x will result given the occurrence of noise alone. The righthand distribution, $f_{SN}(x)$, is the probability density function of x given the occurrence of signal plus noise. (Probability density functions are used, rather than probability functions, since x is assumed to be continuous.) Since the observations will tend to be of greater magnitude when a signal is presented, the mean of the SN distribution will be greater than the mean of the N distribution. In general, the greater the amplitude of the signal, the greater will be the separation of these means.

The Observation as a Value of Likelihood Ratio

It will be well to question at this point our assumption that the observation may be represented along a single axis. Can we, without serious violation, regard the observation as unidimensional, in spite of the fact that the response of the visual system probably has many dimensions? The answer to this question will involve some concepts that are basic to the theory.

One reasonable answer is that when the signal and the interference are alike in character, only the magnitude of the total response of the receiving system is available as an indicator of signal existence. Consequently, no matter how complex the sensory information is in fact, the observations may be represented in theory as having a single dimension. Although this answer is quite acceptable when concerned only with the visual case, we prefer to advance a different answer, one that is applicable also to audition experiments, where, for example, the signal may be a segment of a sinusoid presented in a background of white noise.

So let us assume that the response of the sensory system does have several dimensions, and procede to represent it as a point in an m-dimensional space. Call this point $\underline{y}$. For every such point in this space there is some probability density that it resulted from noise alone [$f_N(y)$], and similarly some probability density that it was due to signal plus noise [$f_{SN}(y)$]. Therefore, there exists a likelihood ratio for each point in the space, $\ell(y) = \frac{f_{SN}(y)}{f_N(y)}$, expressing a likelihood that the point y arose from SN relative to the likelihood that it arose from N. Since any point in the space, i.e., any sensory datum, may be thus represented as a real, non-zero number, these points may be considered to lie along a single axis. We may then, if we choose,

identify the observation x with $\ell(y)$; the decision axis becomes likelihood ratio.[4]

Having established that we may identify the observation x with $\ell(y)$, let us note that we may equally well identify x with any monotonic transformation of $\ell(y)$. It can be shown that we lose nothing by distorting the linear continuum as long as order is maintained. As a matter of fact we may gain if, in particular, we identify x with some transformation of $\ell(y)$ that results in Gaussian density functions of x. We have assumed the existence of such a transformation in the representation of the density functions, $f_{SN}(x)$ and $f_{N}(x)$, in Figure 2. We shall see shortly that the assumption of normality simplifies the problem greatly. We shall also see that this assumption is subject to experimental test. A further assumption incorporated into the picture of Figure 2, one made quite tentatively, is that the two density functions are of equal variance. This is equivalent to the assumption that the SN function is a simple translation of the N function, or that adding a signal to the noise merely adds a constant to the N function. The results of a test of this assumption are also described below.

To summarize these last few paragraphs, we have assumed that an observation may be characterized by a value of likelihood ratio, $\ell(y)$, i.e., the likelihood that the response of the sensory system (y) arose from SN relative to the likelihood that it arose from N. This permits

---

[4] Thus the assumption of a unidimensional decision axis is independent of the character of the signal and noise. Rather, it depends upon the fact that just two decision alternatives are considered. More generally it can be shown that the number of dimensions required to represent the observation is M-1, where M is the number of decision alternatives considered by the observer.

us to view the observations as lying along a single axis. We then assumed the existence of a particular transformation of $\ell(y)$ such that on the resulting variable, x, the density functions are normal. We regard the observer as basing his decisions on the variable x.

The Definition of the Criterion

If the representation depicted in Figure 2 is realistic, then the problem posed for an observer attempting to detect signals in noise is indeed similar to the one faced by the player of our dice game. On the basis of an observation, one that varies only in magnitude, he must decide between two alternative hypotheses. He must decide from which hypothesis the observation resulted; he must state that the observation is a member of the one distribution or the other. As did the player of the dice game, the observer must establish a policy which defines the circumstances under which the observation will be regarded as resulting from each of the two possible events. He establishes a criterion, a cutoff x_c on the continuum of observations, to which he can relate any given observation x_i. If he finds for the i^{th} observation, x_i, that $x_i > x_c$, he says "yes"; if $x_i < x_c$, he says "no." Since the observer is assumed to be capable of locating a criterion at any point along the continuum of observations, it is of interest to examine the various factors that, according to the theory, will influence his choice of a particular criterion. To do so we shall require some additional notation.

In the language of statistical decision theory the observer chooses a subset of all of the observations, namely the critical region A, such that an observation in this subset leads him to accept the hypothesis SN, to say that a signal was present. All other observations are in the complementary subset B; these lead to rejection of the hypothesis

SN, or, equivalently, since the two hypotheses are mutually exclusive and exhaustive, to the acceptance of the hypothesis N. The critical region A, with reference to Figure 2, consists of the values of x to the right of some criterion value x_c.

As in the case of the dice game, a decision will have one of four outcomes: the observer may say yes or no and may in either case be correct or incorrect. The decision outcome, in other words, may be a hit (SN·A, the joint occurrence of the hypothesis SN and an observation in the region A), a miss (SN·B), a correct rejection (N·B), or a false alarm (N·A). If the a priori probability of signal occurrence and the parameters of the distributions of Figure 2 are fixed, the choice of a criterion value, x_c, completely determines the probability of each of these outcomes.

Clearly, the four probabilities are interdependent. For example, an increase in the probability of a hit, P(SN·A), can be achieved only by accepting an increase in the probability of a false alarm, P(N·A), and decreases in the other probabilities, P(SN·B) and P(N·B). Thus a given criterion yields a particular balance among the probabilities of the four possible outcomes; conversely, the balance desired by an observer in any instance will determine the optimal location of his criterion. Now the observer may desire the balance that maximizes the expected value of a decision in a situation where four possible outcomes of a decision have individual values, as did the player of the dice game. The observer, however, may desire a balance that maximizes some other quantity, i.e., a balance that is optimum according to some other definition of optimum, in which case a different criterion will be appropriate. He may, for example, want to maximize P(SN·A) while satisfying a restriction on P(N·A), as we typically do when as experi-

menters we assume a .05 or .01 level of confidence. Alternatively, he may want to maximize the number of correct decisions. Again, he may prefer a criterion that will maximize the reduction in uncertainty in the Shannon sense. (Ref. 9)

In statistical decision theory, and in the theory of signal detectability, the optimal criterion under each of these definitions of optimum is specified in terms of likelihood ratio. That is to say, it can be shown that, if we define the observation in terms of likelihood ratio, $\ell(x) = \frac{f_{SN}(x)}{f_N(x)}$, then the optimal criterion can always be specified by some value β of $\ell(x)$. In other words, the critical region A that corresponds to the criterion contains all observations with likelihood ratio greater than or equal to β, and none of those with likelihood ratio less than β.

It will be well to illustrate this manner of specifying the optimal criterion. We shall do so for just one of the definitions of optimum proposed above, namely, the maximization of the total expected value of a decision in a situation where the four possible outcomes of a decision have individual values associated with them. This, it will be recalled, is the definition of optimum we assumed in the dice game. For this purpose we shall need the concept of conditional probability as opposed to the probability of joint occurrence introduced above. It should be noted that conditional probabilities will have a place in our discussion beyond their use in this illustration; the ones we shall introduce are, as a matter of fact, the fundamental quantities in evaluating the observer's performance.

There are two conditional probabilities of principal interest. These are the conditional probabilities of the observer saying 'yes': $P_{SN}(A)$, the probability of a 'yes' decision conditional upon, or given,

the occurrence of a signal, and $P_N(A)$, the probability of a 'yes' decision given the occurrence of noise alone. These two are sufficient for the other two are simply their complements: $P_{SN}(B) = 1 - P_{SN}(A)$ and $P_N(B) = 1 - P_N(A)$. The conditional and joint probabilities are related as follows:

$$P_{SN}(A) = \frac{P(SN \cdot A)}{P(SN)} \quad \text{and} \quad P_N(A) = \frac{P(N \cdot A)}{P(N)} \tag{1}$$

where $P(SN)$ is the a priori probability of signal occurrence and where $P(N) = 1 - P(SN)$ is the a priori probability of occurrence of noise alone. Equation 1 makes apparent the convenience of using conditional rather than joint probabilities - conditional probabilities are independent of the a priori probability of occurrence of the signal and of noise alone. With reference to Figure 2, we may define $P_{SN}(A)$, or the conditional probability of a hit, as the integral of $f_{SN}(x)$ over the critical region A, and $P_N(A)$, the conditional probability of a false alarm, as the integral of $f_N(x)$ over A. That is, $P_N(A)$ and $P_{SN}(A)$ represent respectively the areas under the two curves of Figure 2 to the right of some criterion value of x.

To pursue our illustration of how an optimal criterion may be specified by a critical value of likelihood ratio β, let us note that the expected value of a decision (denoted EV) is defined in statistical-decision theory as the sum, over the potential outcomes of a decision, of the products of probability of outcome and the desirability of outcome. Thus, using the notation V for positive individual values and K for costs or negative individual values, we have the following equation:

$$EV = V_{SN \cdot A}\, P(SN \cdot A) + V_{N \cdot B}\, P(N \cdot B) - K_{SN \cdot B}\, P(SN \cdot B) - K_{N \cdot A}\, P(N \cdot A). \tag{2}$$

Now if a priori and conditional probabilities are substituted for the joint probabilities in Eq. 2 following Eq. 1, for example, $P(SN)\ P_{SN}(A)$ for $P(SN \cdot A)$, then collecting terms yields the result that maximizing EV is equivalent to maximizing

$$P_{SN}(A) - \beta\, P_{N}(A), \tag{3}$$

where

$$\beta = \frac{P(N)}{P(SN)} \frac{(V_{N \cdot B} + K_{N \cdot A})}{(V_{SN \cdot A} + K_{SN \cdot B})} \tag{4}$$

From Eq. 3 it may be seen that the value β simply weights the hits and false alarms, and from Eq. 4 we see that β is determined by the a priori probabilities of occurrence of signal and of noise alone and by the values associated with the individual decision outcomes. It should be noted that Eq. 3 applies generally, that is, to all definitions of optimum. Eq. 4 shows the determinants of β in only the special case of the expected-value definition of optimum.

Return for a moment to Figure 2, and recall that β is a critical value of $\ell(x) = \frac{f_{SN}(x)}{f_{N}(x)}$. It should now be clear that the optimal cutoff, x_c, along the x axis is at the point on this axis where the ratio of the ordinate value of $f_{SN}(x)$ to the ordinate value of $f_{N}(x)$ is a certain number, namely β. In the symmetrical case, where the two a priori probabilities are equal and the four individual values are equal, $\beta = 1$ and the optimal value of x_c is the point where $f_{SN}(x) = f_{N}(x)$, where the two curves cross. If the four values are equal but $P(SN) = 5/6$ and $P(N) = 1/6$, another case described in connection with the dice game, then $\beta = 1/5$ and the optimal value of x_c is shifted to the left, to the point where the height of the ordinate of the function $f_{SN}(x)$ is 1/5 the height of the ordinate of the function $f_{N}(x)$. This

shift may be seen intuitively to be in the proper direction - a higher value of P(SN) should lead to a greater willingness to accept the hypothesis SN, i.e., a more lenient cutoff. To consider one more example from the dice game, if P(SN) = P(N) = 0.5, if $V_{N \cdot B}$ and $K_{N \cdot A}$ are set at five dollars and $K_{SN \cdot A}$ and $K_{SN \cdot B}$ are equal to one dollar, then $\beta = 5$ and the optimal value of x_c shifts to the right, to the point where the value of $f_{SN}(x)$ is five times the value of $f_N(x)$. Again intuitively, if it is more important to be correct when the hypothesis N is true, a high, or strict, criterion should be adopted.

In any case, β specifies the optimal weighting of hits relative to false alarms: x_c should always be located at the point on the x axis corresponding to β. As we pointed out in discussing the dice game, just where this value of x_c will be with reference to the x axis depends not only upon the a priori probabilities and the values but also upon the overlap of the two density functions, in short, upon the signal strength. We shall define a measure of signal strength within the next few pages. For now, it is important to note that for any detection goal to which the observer may subscribe, and for any set of parameters that may characterize a detection situation (such as a priori probabilities and values associated with decision outcomes), the optimal criterion may be specified in terms of a single number, β, a critical value of likelihood ratio.[5]

---

[5] We have reached a point in the discussion where we can justify the statement made earlier that the decision axis may be equally well regarded as likelihood ratio or as any monotonic transformation of likelihood ratio. Any distortion of the linear continuum of likelihood ratio, that maintains order, is equivalent to likelihood ratio in terms of determining a criterion. The decisions made are the same whether the criterion is set at likelihood ratio equal to β or at the value that corresponds to β of some new variable. To illustrate, if a criterion leads to a 'yes' response whenever $\ell(y) > 2$, if $x = [\ \ell(y)]^2$ the decisions will be the same if the observer says 'yes' whenever $x > 4$.

The Receiver Operating Characteristic

Whatever criterion the observer actually uses, even if it is not one of the optimal ones, can also be described by a single number, by some value of likelihood ratio. Let us procede to a consideration of how the observer's performance may be evaluated with respect to the location of his criterion, and, at the same time, we shall see how his performance may be evaluated with respect to his sensory capabilities.

As we have noted, the fundamental quantities in the evaluation of performance are $P_N(A)$ and $P_{SN}(A)$, these quantities representing respectively the areas under the two curves of Figure 2 to the right of some criterion value of x. If we set up a graph of $P_{SN}(A)$ versus $P_N(A)$ and trace it on the curve resulting as we move the decision criterion along the decision axis of Figure 2, we sketch one of the arcs shown in Figure 3. Ignore, for the moment, all but one of these arcs. If the decision criterion is set way at the left in Figure 2, we obtain a point in the upper right hand corner of Figure 3: both $P_{SN}(A)$ and $P_N(A)$ are unity. If the criterion is set at the right end of the decision axis in Figure 2, the point at the other extreme of Figure 3, $P_{SN}(A) = P_N(A) = 0$, is obtained. In between these extremes lie the criterion values of more practical interest. It should be noted that the exact form of the curve shown in Figure 3 is not the only form which might result, but it is the form which will result if the observer chooses a criterion in terms of likelihood ratio, and the probability density functions are normal and of equal variance.

Such a plot was first suggested by Peterson, Birdsall and Fox. (Ref. 2) They called the curve the receiver operating characteristic (ROC). The optimal "operating level" may be seen from Eq. 3 to be at the point of the ROC curve where its slope is β, where β is the value

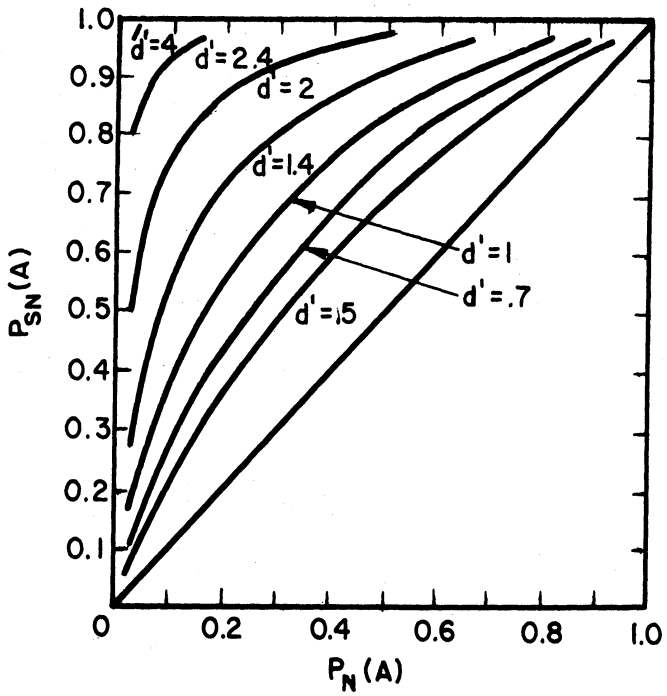

FIG. 3 THE RECEIVER-OPERATING-CHARACTERISTIC (ROC) CURVES

These curves show $P_{SN}(A)$ vs. $P_N(A)$ with d' as the parameter. They are based on the assumptions that the probability density functions, $f_N(x)$ and $f_{SN}(x)$, are normal and of equal variance.

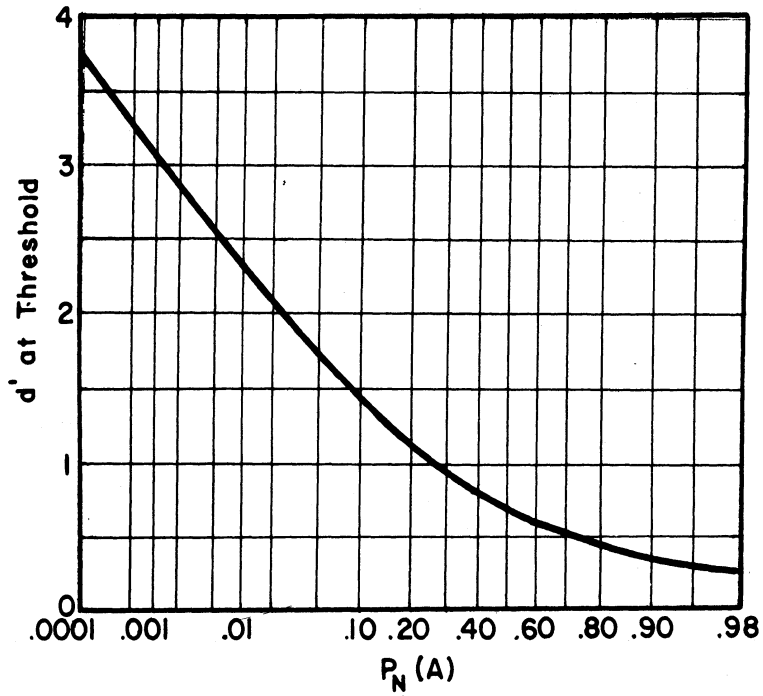

FIG. 4 THE RELATIONSHIP BETWEEN d' AND $P_N(A)$ AT THRESHOLD

of likelihood ratio corresponding to the optimal criterion. That is, the expression $P_{SN}(A) - P_N(A)$ defines a utility line of slope β, and the point of tangency of this line to the ROC curve is the optimal operating level. Thus the theory specifies the appropriate hit probability and false alarm probability for any definition of optimum and any set of parameters characterizing the detection situation. It is now apparent how the observer's choice of a criterion in a given experiment may be indexed. The observer yields a pair of values, $P_N(A)$ and $P_{SN}(A)$; i.e., a point on an ROC curve. The slope of the curve at this point corresponds to the value of likelihood ratio at which he has located his criterion. Thus we work backward from the ROC curve to infer the criterion that is employed by the observer.

There is, of course, a family of ROC curves, as shown in Figure 3, a given curve corresponding to a given separation between the means of the density functions $f_N(x)$ and $f_{SN}(x)$. The parameter of these curves has been called d', where d' is defined as the difference between the means of the two density functions expressed in terms of their standard deviation, that is,

$$d' = \frac{M_{f_{SN}}(x) - M_{f_N}(x)}{\sigma_{f_N}(x)} \qquad (5)$$

Since the separation between the means of the two density functions is a function of signal amplitude, d' is an index of the detectability of a given signal for a given observer.

Recalling our assumptions that the density functions, $f_N(x)$ and $f_{SN}(x)$, are normal and of equal variance, we may see from Eq. 5 that the quantity denoted d' is simply the familiar normal deviate, or x/σ measure. From the pair of values, $P_N(A)$ and $P_{SN}(A)$, that are obtained

experimentally, one may procede to a published table of areas under the normal curve to determine a value of d'. The quantity, $P_N(A)$, is used to determine the distance of the criterion from the mean of $f_N(x)$ in x/σ units; likewise $P_{SN}(A)$ indicates the distance of the criterion from the mean of $f_{SN}(x)$ in x/σ units. If the criterion lies between the means, the two x/σ values are added without regard to sign to yield d'; if the criterion does not lie between the means, the lesser of the x/σ values is subtracted from the larger, again without regard to sign, to yield d'. A simple computational procedure is achieved with the use of graph paper having a probability scale and a normal-deviate scale on both axes. A specially constructed slide rule is a still further convenience.

We see now that the pair of values, $P_N(A)$ and $P_{SN}(A)$, determined in an experiment (using the obtained proportions as estimates of the probabilities) yields measures of two independent aspects of the observer's performance. The variable d' is a measure of the observer's sensory capabilities, or of the effective signal strength. This may be thought of as the object of interest in classical psychophysics. The criterion β that is employed by the observer, which determines the $P_N(A)$ and $P_{SN}(A)$ for some fixed d', reflects at least in part the effect of variables which have been variously called the set, attitude, and motives of the observer. It is the ability to distinguish between these two aspects of detection performance that comprises one of the main advantages of the theory proposed here. These two aspects of behavior are confounded in an experiment whose dependent variable is the intensity of the signal that is required for a threshold response.

The Relationship of d' to Signal Energy

We have seen that the optimal value of the criterion, β, can be computed. It should be noted that, in certain instances, an optimal value of d', i.e., the sensitivity of the mathematically ideal device, can also be computed. If, for example, the exact waveform and starting time of the signal are determinable, then $d'_{optimum} = \sqrt{\frac{2E}{N_o}}$, where E is the signal energy and N_o is the noise power in a one-cycle band. (Ref. 2) Unfortunately for our present purposes, it is not entirely clear how one might determine an optimal d' for visual signals. In the case of auditory signals, however, the procedure is quite straightforward. This means that in audition experiments it is possible to compare an empirically determined index of detectability with ideal detectability, just as observed and optimal indices of decision criteria may be compared. The ratio of the two detectability indices provides a measure of the observer's efficiency. This measure has been a useful one in several contexts.

The Use of Ideal Descriptions as Models

It might be worthwhile to describe at this point some of the reasons for the emphasis placed here on optimal measures, and, indeed, the reasons for the general enterprise of considering a theory of ideal behavior as a model for studies of real behavior.[6] In view of the deviations from any ideal which are bound to characterize real organisms, it might appear at first glance that any deductions based on ideal premises could have no more than academic interest. We do not think that this is the case. In any study, it is desirable to specify rigorously the factors pertinent to the study. Ideal condi-

[6] The discussion immediately following is, in part, a paraphrase of one in Horton. (Ref. 10)

tions generally involve few variables and permit these to be described in simple terms. Having identified the performance to be expected under ideal conditions, it is possible to extend the model to include the additional variables associated with real organisms. The ideal performance, in other words, constitutes a convenient base from which to explore the complex operation of a real organism.

In certain cases, as in the problem at hand, values characteristic of ideal conditions may actually approximate very closely those characteristics of the organism under study. The problem then becomes one of changing the ideal model in some particular so that it is slightly less than ideal. This is usually accomplished by depriving the ideal device of some particular function. This method of attack has been found to generate useful hypotheses for further studies. Thus, whereas it is not expected that the human observer and the ideal detection device will behave identically, the emphasis in early studies is on similarities. If the differences are small, one may rule out entire classes of alternative models, and regard the model in question as a useful tool in further studies. Proceeding on this assumption, one may then, in later studies, emphasize the differences, the form and extent of the differences suggesting how the ideal model may be modified in the direction of realism.

Alternative Conceptions of the Detection Process

The earliest studies that were undertaken to test the applicability of the decision model to human observers were quite naturally oriented toward determining its value relative to existing psychophysical theory. As a result, some of the data presented below are meaningful only with respect to differences in the predictions based upon different theories. We shall, therefore, briefly consider alternative

theories of the detection process.

Although it is difficult to specify with precision the alternative theories of detection, it is clear that they all involve the concept of the threshold in an important way. The development of the threshold concept is fairly obscure. It is differently conceived by different people, and few popular usages of the concept benefit from explicit statement. One respect, however, in which the meaning of the threshold concept is entirely clear is its assertion of a lower limit on sensitivity. As we have just seen, the decision model does not include such a boundary. The decision model specifies no lower bound on the location of the criterion along the continuous axis of sensory inputs. Further, it implies that any displacement of the mean $f_{SN}(x)$ from the mean of $f_N(x)$, no matter how small, will result in a greater value of $P_{SN}(A)$ than $P_N(A)$, irrespective of the location of the criterion.

For purposes of experimental comparison of decision theory and threshold theory, we shall consider a special version of threshold theory. (Ref. 11) Although it is a special version, we believe it retains the essence of the threshold concept. In this version, the threshold is described in the terms of decision theory. It is regarded as a cutoff on the continuum of observations (see Figure 2) with a fixed location, with values of x above the cutoff always evoking a positive response, and with discrimination impossible among values of x below the cutoff. This assertion of a fixed cutoff and a stimulus effect that varies randomly, it will be noted, is equivalent to the more common description in terms of a randomly varying cutoff and a fixed stimulus effect. There are several reasons for assuming that the hypothetical threshold cutoff is located quite high relative to the density function $f_N(x)$, say at approximately $+3\sigma$ from the mean of

$f_N(x)$. We shall compare our data with the predictions of such a "high-threshold" theory, and shall indicate their relationship to predictions from a theory assuming a lower threshold. We shall, in particular, ask how low a threshold cutoff would have to be to be consistent with the reported data. It may be noted that if a high threshold exists, the observer will be incapable of ordering values of x likely to result from noise alone, and hence will be incapable of estimating $\ell(x)$ and of varying his criterion over a significant range.

If a threshold exists that is rarely exceeded by noise alone, this fact will be immediately apparent from the ROC curves (see Figure 3) that are obtained experimentally. It can be shown that the ROC curves in this case are straight lines from points on the left-hand vertical axis [$P_{SN}(A)$] to the upper right-hand corner of the plot. These straight-line curves represent the implication of a high-threshold theory that an increase in $P_N(A)$ must be effected by responding 'yes' to a random selection of observations that fail to reach the threshold, rather than by a judicious selection of observations, that is, a lower criterion level. If we follow the usual procedure of regarding the stimulus threshold as the signal intensity yielding a value of $P_{SN}(A)$ = 0.5 for $P_N(A)$ = .00, then an appreciation of the relationship between d' and $P_N(A)$ at threshold may be gained by visualizing a straight line in Figure 3 from this point to the upper right-hand corner. If we note which of the ROC curves drawn in Figure 3 are intersected by the visualized line, we see that the threshold decreases with increasing $P_N(A)$. For example, a response procedure resulting in a $P_N(A)$ = .02 requires a signal of d' = 2.0 to reach the threshold, whereas a response procedure yielding a $P_N(A)$ = .98 requires a signal of d' $<$0.5 to reach the threshold. A graph showing what threshold would be calculated as

a function of $P_N(A)$ is plotted in Figure 4. The calculated threshold is a strictly monotonic function of $P_N(A)$ ranging from infinity to zero.

Thus, the fundamental difference between the threshold theory we are considering and decision theory lies in their treatment of false alarm responses. According to the threshold theory, these responses represent guesses determined by non-sensory factors; that is, $P_N(A)$ is independent of the cutoff which is assumed to have a fixed location. Decision theory assumes, on the other hand, that $P_N(A)$ varies with the temporary position of a cutoff under the observer's control; that false alarm responses arise for valid sensory reasons, and that therefore a simple correction will not eliminate their effect on $P_{SN}(A)$. A similar implication of Figure 4 that should be noted is that reliable estimates of $P_{SN}(A)$ or of the stimulus threshold are not guaranteed by simply training the observer to maintain a low, constant value of $P_N(A)$. Since extreme probabilities cannot be estimated with reliability, the criterion may vary from session to session with the variation having no direct reflection in the data. Certainly, false alarm rates of .01, .001 and .0001 are not discriminable in an experimentally feasible number of observations; the differences in the calculated values of the threshold associated with these different values of $P_N(A)$ may be seen from Figure 4 to be sizeable. The experiments reported in Part II were designed, in large measure, to clarify the relationship that exists between $P_N(A)$ and $P_{SN}(A)$, to show whether or not the observer is capable of controlling the location of his criterion for a 'yes' response.

II. SOME EXPERIMENTS

Five experiments are reported in the following. Three of them pose for the observer what we have called the fundamental detection problem, the problem that occupied our attention throughout the theoretical discussion. The other two experiments present slightly different tasks. They test certain implications of decision theory that we have not yet treated explicitly. These implications, however, will be seen to follow very directly from the theory and to contribute significantly to an evaluation of it.

The first two experiments described below test the observer's ability to assume that criterion which maximizes the expected value of a decision. The a priori probability of a signal occurrence and the individual values associated with the four possible decision outcomes are varied systematically, in order to determine the range over which the observer can vary his criterion and the form of the resultant ROC curve. A third experiment tests the observer's ability to maximize the proportion of hits while satisfying a restriction on the proportion of false alarms. This experiment is largely concerned with the degree of precision with which the observer can locate a criterion.

The remaining two experiments differ in that the tasks they present to the observer do not require him to establish a criterion, that is, they do not require a 'yes' or 'no' response. In one of these the observer is asked to report after each observation interval his subjective probability that the signal existed during the interval. This response is a familiar one; it is essentially a rating or a judgment of confidence. The report of "a posteriori probability of signal existence," as it is termed in detection theory, may be regarded as reflecting the

likelihood ratio of the observation. This case is of interest since an estimate of likelihood ratio preserves more of the information contained in the observation than does a report merely that the likelihood ratio fell above or below a critical value. We shall see that it is also possible to construct the ROC curve from this type of response.

The other experiment not requiring a criterion employs what has been termed the temporal forced-choice method of response. On each trial a signal is presented in exactly one of n temporal intervals, and the observer states in which interval he believes the signal occurred. The optimal procedure for the observer to follow in this case is to make an observation in each interval and to choose the interval having the greatest value of likelihood ratio associated with it. Since decision theory specifies how the proportion of correct responses obtained with the forced-choice method is related to the detectability index d', the internal consistency of the theory may be evaluated. That is to say, if the observer follows the optimal procedure, then the estimate of the detectability of a signal of a given strength that is based on forced-choice data will be comparable to that based on yes-no data. The forced-choice method may also be used to make a strong test of a fundamental assumption of decision theory, namely that sensory information is continuous, or that sensory information does not exhibit a threshold discontinuity. For an experiment requiring the observer to rank the n intervals according to their likelihood of containing the signal, the continuity and discontinuity assumptions lead to very different predictions concerning the probability that an interval ranked other than first will be the correct interval.

All of the experiments reported in the following employed a circular signal with a diameter of 30 minutes of visual angle and a duration of 1 100 of a second. The signal was presented on a large, uniformly illuminated background having a luminance of ten foot-lamberts. Details of the apparatus have been presented elsewhere. (Ref. 12)

Maximizing the Expected-Value of a Decision - An Experimental Analysis

A direct test of the decision model is achieved in an experiment in which the a priori probability of signal occurrence or the values of the decision outcomes, or both, are varied from one group of observations to another - in short, in which β (Eqs. 3 and 4) assumes different values. The observer, in order to maximize his expected value, or his payoff, must vary his willingness to make a 'yes' response, in accordance with the change in β. Variations in this respect will be indicated by the proportion of false alarms, $P_N(A)$. The point of interest is how $P_{SN}(A)$, the proportion of hits, varies with changes in $P_N(A)$, that is, in the form of the observer's ROC curve. If the experimental values of $P_N(A)$ reflect the location of the observer's criterion, if the observer responds on the basis of the likelihood ratio of the observation, and if the density functions (Figure 2) are normal and of equal variance, the ROC curve of Figure 3 will result. If, on the other hand, the location of the criterion is fixed, i.e., is not under the observer's control, then the resulting ROC curve will be a straight line, as we have indicated above. We shall examine some empirical ROC curves with this distinction in mind.

This experiment can be made to yield another, and, in one sense, a stronger test of these two hypotheses, by employing several values of signal strength. For in this case stimulus thresholds can be calculated, and correlational techniques can be used to determine whether

the calculated threshold is dependent upon $P_N(A)$ as predicted by decision theory or independent of $P_N(A)$ as predicted by what we have termed the high-threshold theory. The first of the two expected-value experiments performed employed four values of signal strength.

The First Expected-Value Experiment. Three observers, after considerable practice, served in 16 two-hour sessions. In each session, signals at four levels of intensity (0.44, 0.69, 0.92 and 1.20 foot-lamberts) were presented along with a "blank" or "no-signal" presentation. The order of the presentation was random within a restriction placed upon the total number of occurrences of each signal intensity and the blank in a given session. Each of the signal intensities occurred equally often within a session. The proportion of trials on which a signal (of any intensity) was presented, P(SN), was either 0.80 or 0.40 in the various sessions. In all, there were 300 presentations in each session - six blocks of 50 presentations, separated by rest periods. Thus each estimate of $P_N(A)$ is based on either 60 or 180 observations, and each estimate of $P_{SN}(A)$ is based on 30 or 60 observations, depending on P(SN).

In the first four sessions, no values were associated with the various decision outcomes. For the first and fourth sessions the observers were informed that P(SN) = 0.80 and, for the second and third sessions, that P(SN) = 0.40. The average value of $P_N(A)$ obtained in the sessions with P(SN) = 0.80 was 0.43, and, in the sessions with P(SN) = 0.40, it was 0.15 - indicating that the observer's willingness to make a 'yes' response is significantly affected by changes in P(SN) alone. In the remaining twelve sessions, these two values of P(SN) were used in conjunction with a variety of values placed on the decision outcomes. In the fifth session, for example, the observers were

told that P(SN) = 0.80 and were, in addition, given the following payoff matrix:

	No	Yes
Signal	$-1_{(K_{SN \cdot B})}$	$+1_{(V_{SN \cdot A})}$
No Signal	$+2_{(V_{N \cdot B})}$	$+2_{(K_{N \cdot A})}$

A variety of simple matrices was used, including [-1, +1, +3, -3] and [-1, +1, +4, -4] with P(SN) = 0.80; and [-1, +1, +2, -2], [-1, +1, +1, -1], [-2, +2, +1, -1] and [-3, +3, +1, -1] with P(SN) = 0.40. By reference to Eq. 4, it may be seen that these matrices and values of P(SN) define values of β ranging from 0.25 to 3.00. The observers were actually paid in accordance with these payoff matrices, in addition to their regular wage. The values were equated with fractions of cents, these fractions being adjusted so that the expected earnings per session remained relatively constant, at approximately 1.00.

The obtained values of $P_N(A)$ varied in accordance with changes in the a priori probability of signal occurrence. Just how closely the obtained values of $P_N(A)$ approached those specified as optimal by the theory, we shall discuss shortly. For now, we may note that the range of values of $P_N(A)$ obtained from the three observers is shown in Figures 5, 6 and 7. These figures also show four values of $P_{SN}(A)$ corresponding to each value of $P_N(A)$; the four values of $P_{SN}(A)$, one for each signal strength, are indicated in the figures by different symbols. We have, then, in each of Figures 5, 6 and 7, four ROC curves.

Although entire ROC curves are not precisely defined by the data of the first experiment, these data are adequate for the purpose of distinguishing between the predictions of decision theory and the predictions of a high-threshold theory. It is clear, for example, that the straight lines fitted to the data do not intersect the upper right-hand corner of the graph, as required by the concept of a high threshold.

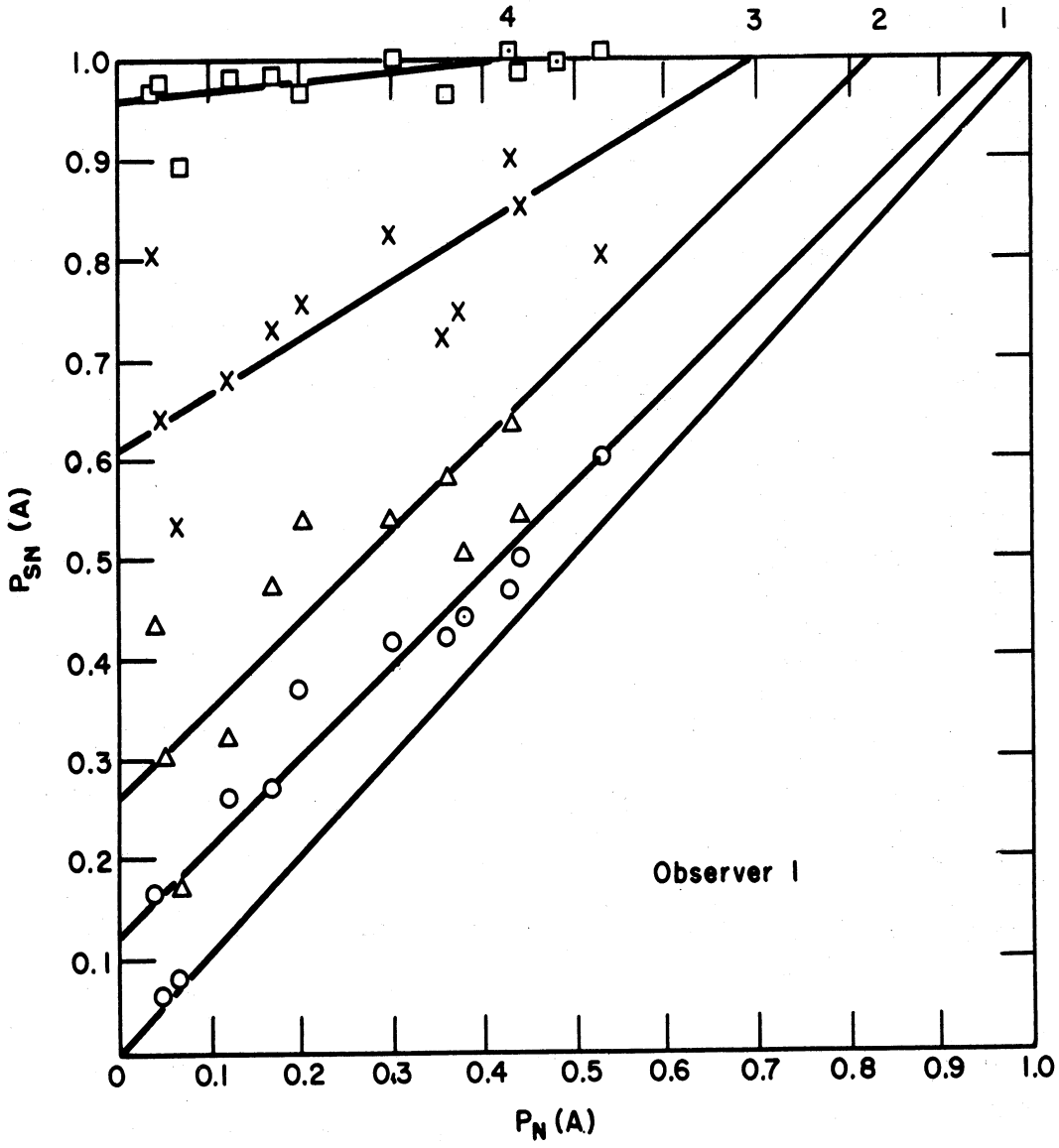

FIG. 5 EMPIRICAL ROC CURVE IN THE FIRST EXPECTED-VALUE EXPERIMENT

It is also apparent that the data are better fitted by theoretical ROC curves of the form predicted by decision theory, as shown in Figure 3.

Another analysis of the data is of interest in distinguishing the two theories we are considering. As we have indicated earlier in this paper and developed in detail elsewhere,(Ref. 3) the concept of a high-threshold leads to the prediction that the stimulus threshold is inde-

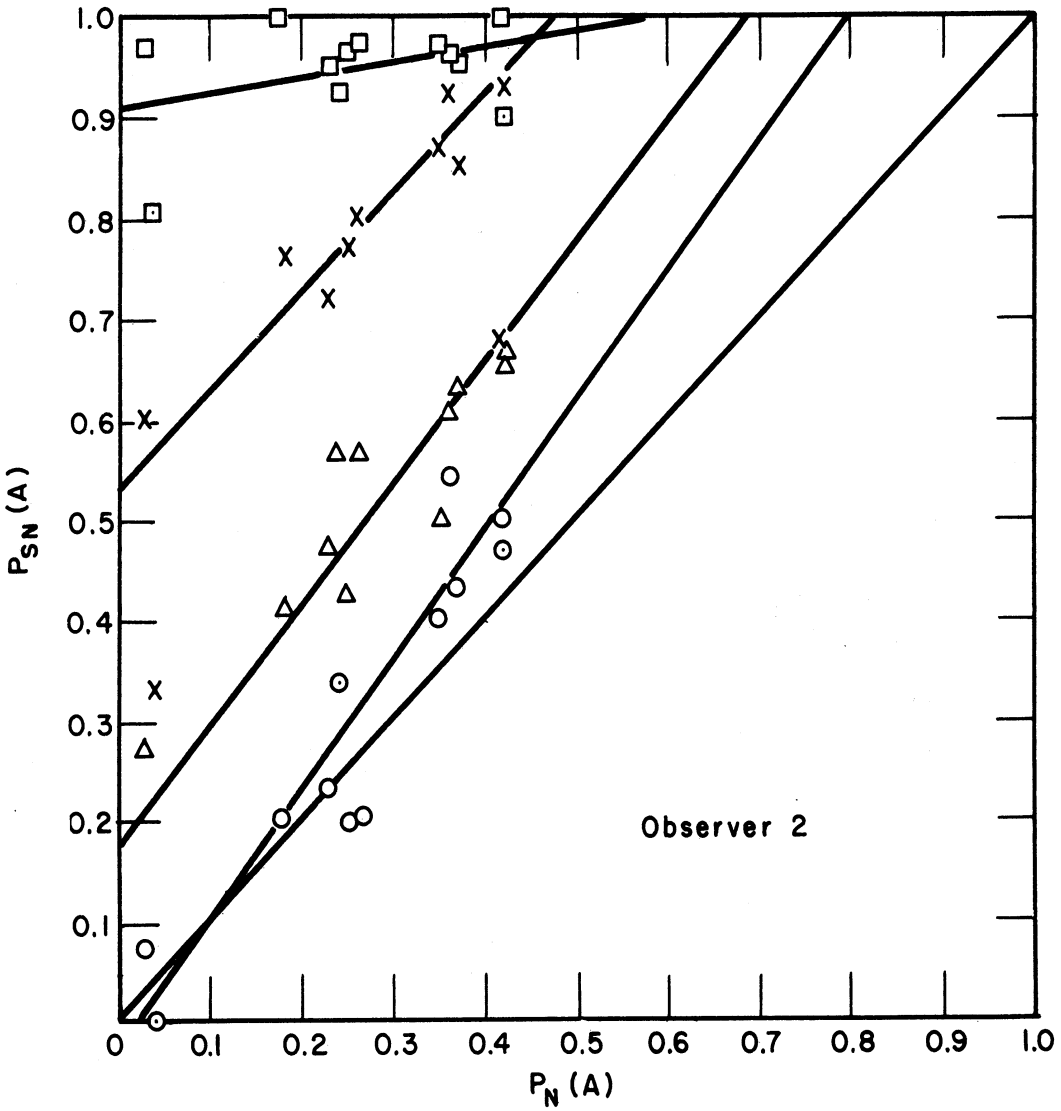

FIG. 6 EMPIRICAL ROC CURVE IN THE FIRST EXPECTED-VALUE EXPERIMENT

pendent of $P_N(A)$, whereas decision theory predicts a negative correlation between the stimulus threshold and $P_N(A)$. Within the framework of the high-threshold model that we have described, the stimulus threshold is defined as the stimulus intensity that yields a $P_{SN}(A) = 0.50$ for $P_N(A) = 0.0$. This stimulus intensity may be determined by interpolation from psychophysical functions [$P_{SN}(A)$ vs. signal intensity]

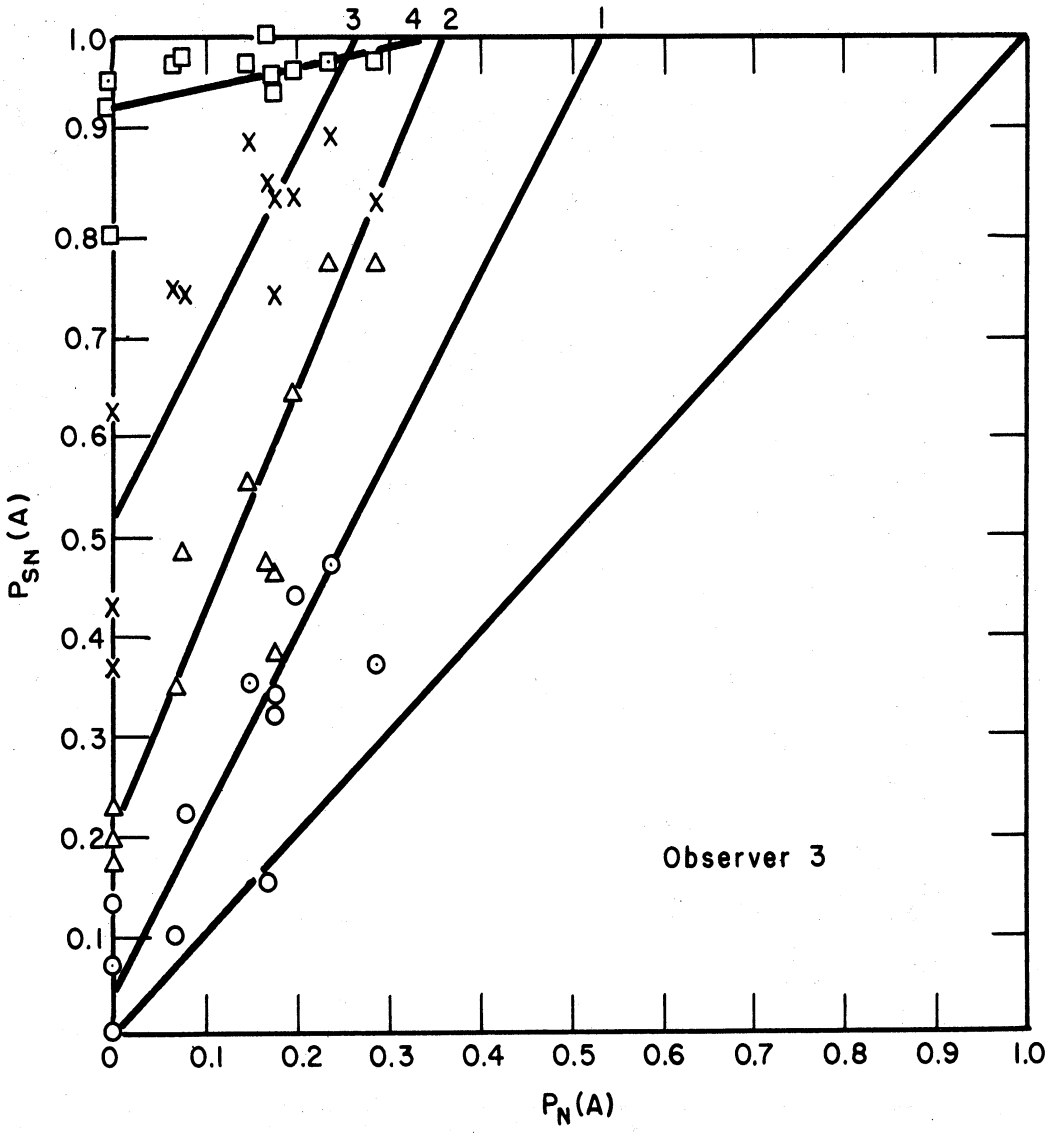

FIG. 7 EMPIRICAL ROC CURVE IN THE FIRST EXPECTED-VALUE EXPERIMENT

that are normalized so that $P_N(A) = 0.0$. The normalization is effected by the equation

$$P_{SN}(A)_{corrected} = \frac{P_{SN}(A) - P_N(A)}{1 - P_N(A)}, \qquad (6)$$

commonly known as the "correction for chance success." The intent of the correction is to remove what has been regarded as the spurious element of $P_{SN}(A)$ that is contributed by an observer's tendency to make a 'yes' response in the absence of any sensory indication of a signal, i.e., to make a 'yes' response following an observation that fails to reach the fixed criterion level. The validity of this correction procedure can be shown to be implied by the assumption of what we have termed a high treshold. The decision model, as we have indicated, differs in that it regards sensory information as thoroughly probabilistic, without discontinuity - it asserts that the presence and absence of some sensory indication of a signal are not separable categories. According to the decision model, the observer does not achieve more 'yes' responses by responding positively to a random selection of observations that fall short of the fixed criterion level, but by lowering his criterion. In this case, the chance correction is inappropriate; the stimulus threshold will not remain invariant with changes in $P_N(A)$.

The relationship of the stimulus threshold to $P_N(A)$ in this first experiment is illustrated in Figures 8 and 9. The portion of data comprising each of the curves in these figures was selected to be relatively homogeneous with respect to $P_N(A)$. The curves are average curves for the three observers. Figure 8 shows $P_N(A)$ and $P_{SN}(A)$ as a function of signal intensity, ΔI. Figure 9 shows the <u>corrected</u> value

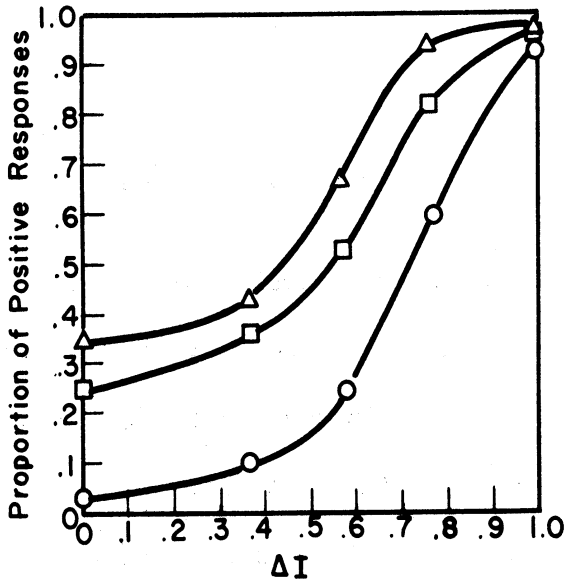

FIG. 8 THE RELATIONSHIP BETWEEN THE STIMULUS THRESHOLD AND $P_N(A)$

Proportion of positive responses to four positive values of signal intensity, $P_{SN}(A)$, and to the blank or zero-intensity presentation, $P_N(A)$, at three values of $P_N(A)$.

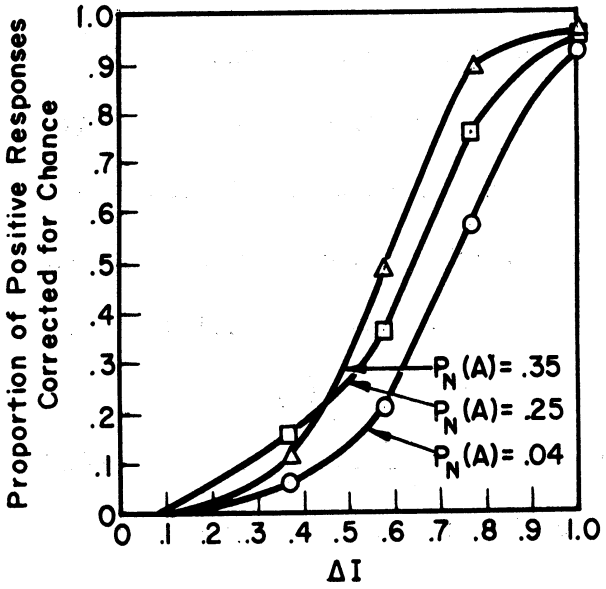

FIG. 9 THE RELATIONSHIP BETWEEN THE STIMULUS THRESHOLD AND $P_N(A)$

The three curves corrected for chance success, by Eq. 6.

of $P_{SN}(A)$ plotted against signal intensity. It may be seen in Figure 9 that the stimulus threshold (the value of ΔI corresponding to a corrected $P_{SN}(A)$ of 0.50) is dependent upon $P_N(A)$ in the direction predicted by decision theory.[7]

Figures 8 and 9 portray the relationship in question in a form to which many of us are accustomed. We can, of course, achieve a stronger test by computing the coefficients of correlation between $P_N(A)$ and the calculated threshold. We have made this computation, and have in the process avoided the averaging of data obtained from different observers and different experimental sessions. The product-moment coefficients for the three observers are -.37 ($p = .245$), -60 ($p = .039$) and -.81 ($p = .001$) respectively. For the three observers combined, $p = .0008$. The implication of these correlations is the same as the implication of the straight lines fitted to the data of Figures 5, 6 and 7, namely, that a dependence exists between the conditional probability that an observation arising fron SN will exceed the criterion and the conditional probability that an observation arising from N will exceed the criterion. Stated otherwise, the correlations imply that the observer's decision function is likelihood ratio or some monotonic function of it and that his criterion varies with changes in the optimal criterion, β.

The Second Expected-Value Experiment. A second expected-value experiment was conducted to obtain a more precise definition of the

---

[7] ΔI is plotted in Figures 8 and 9 in terms of the transmission values of the filters that were placed selectively in the signal beam to yield different signal intensities. These values (.365, .575, .765, 1.000) are converted to the signal values in terms of foot-lamberts that we have presented above, by multiplying them by 1.20, the value of the signal in foot-lamberts without selective filtering.

ROC curve than that provided by the experiment just described. In the second experiment greater definition was achieved by increasing the number of observations on which the estimates of $P_{SN}(A)$ and $P_N(A)$ were based, and by increasing the range of values of $P_N(A)$.

In this experiment only one signal intensity (0.78 foot-lamberts) was employed. Each of 13 experimental sessions included 200 presentations of the signal and 200 presentations of noise alone, i.e., 200 observation intervals in which no signal was presented. Thus, P(SN) remained constant at 0.50 throughout this experiment. Changes in the optimal criterion, β, and, hence, in the obtained values of $P_N(A)$, were effected entirely by changes in the values associated with the decision outcomes. These values were manipulated to yield β's (Eq. 4) varying from 0.16 to 8.00. A different set of observers served in this experiment.

The results are portrayed in Figures 10 to 13. It may be seen that the experimentally determined points are fitted quite well by the type of ROC curve that is predicted by decision theory. It is equally apparent, excepting Observer 1, that the points do not lie along a straight line intersecting the point $P_N(A) = P_{SN}(A) = 1.00$, as predicted by the high-threshold model.

One other feature of these figures is worthy of note. It will be recalled that in our presentation of decision theory, we tentatively assumed that the density functions of noise and signal plus noise, $f_N(x)$ and $f_{SN}(x)$, are of equal variance. Although we did not, in order to preserve the continuity of the discussion, we might have acknowledged at that point that the assumption of equal variance is not necessarily the best one. In particular, one might rather expect the variance of $f_{SN}(x)$ to be proportional to its mean. At any rate,

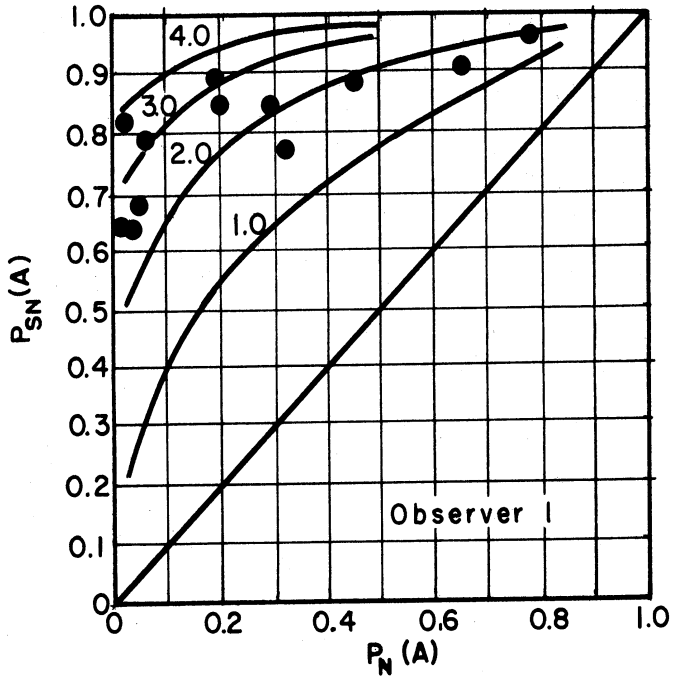

FIG. 10 EMPIRICAL ROC CURVE IN THE SECOND EXPECTED-VALUE EXPERIMENT

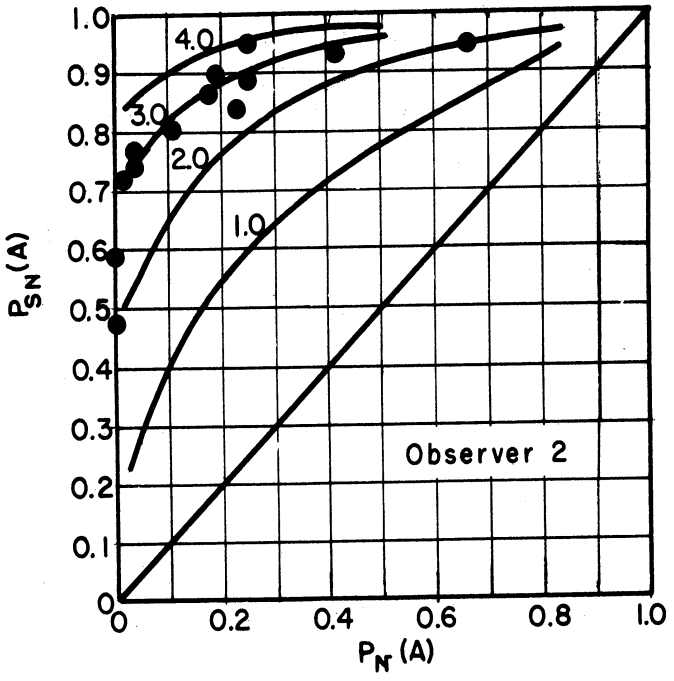

FIG. 11 EMPIRICAL ROC CURVE IN THE SECOND EXPECTED-VALUE EXPERIMENT

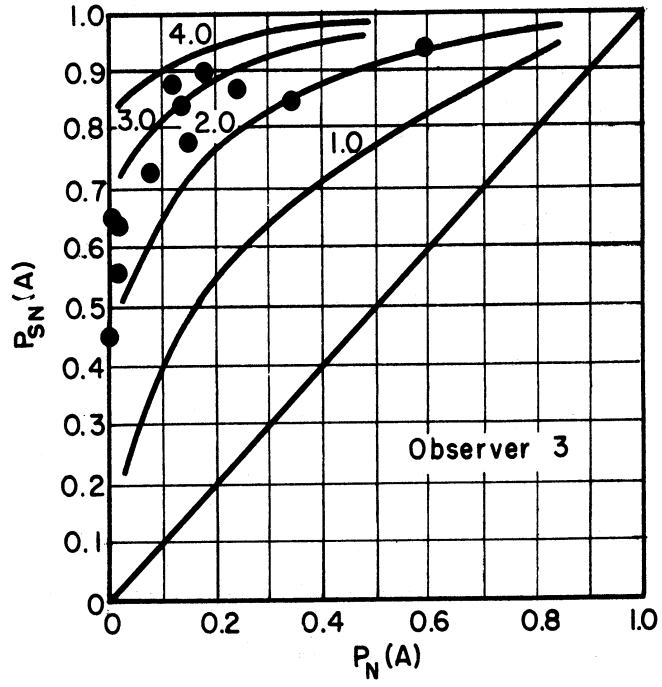

FIG. 12 EMPIRICAL ROC CURVE IN THE SECOND EXPECTED-VALUE EXPERIMENT

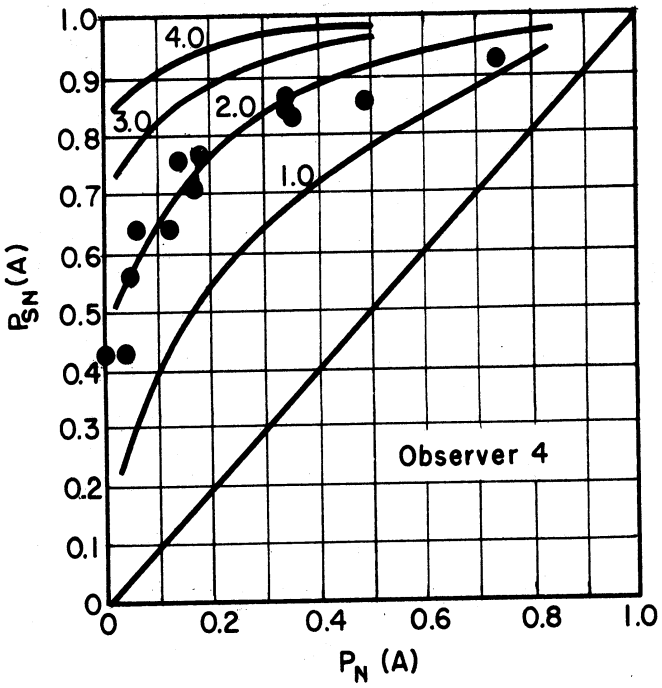

FIG. 13 EMPIRICAL ROC CURVE IN THE SECOND EXPECTED-VALUE EXPERIMENT

the assumption made about the variances represents a degree of freedom of the theory that we have not emphasized previously. We have, however, made use of this degree of freedom in the construction of the theoretical ROC curves of Figures 10 to 13. It may be seen that these curves are not symmetrical about the diagonal, as are the curves of Figure 3 that are predicated on equal variance. The curves of Figures 10 to 13 are based on the assumption that the ratio of the increment of the mean of $f_{SN}(x)$ to the increment of its standard deviation is equal to 4, $\frac{\Delta m}{\Delta \sigma} = 4$. A close look at these figures suggests that ROC curves calculated from a still greater ratio would provide a still better fit. Since some other data presented in the following bear directly on this question of a dependence between variance and signal strength, we shall postpone further discussion of it. We shall also consider later whether the exact form of the empirical ROC curves supports the assumption of normality of the density functions, $f_N(x)$ and $f_{SN}(x)$. For now, the main point of interest is that decision theory predicts the curvilinear form of the ROC curves that are yielded by the observers.

The Forced-Choice Experiments

We have indicated above that an extension of the decision model may be made to predict performance in a forced-choice test. On each trial of a typical forced-choice test, the signal is presented in one of n temporal intervals, and the observer selects the interval he believes to have contained the signal. It will be intuitively clear that, to behave optimally, the observer must estimate the likelihood ratio, $\ell(x)$, associated with each interval and choose the interval having the greatest associated likelihood ratio. Equivalently, he may rank the intervals according to the observation x made in each and choose that interval yielding the greatest value of x.

If the observer behaves optimally, then the probability that a correct answer will result, P(c), for a given value of d', is expressed by

$$P(c) = \int_{-\infty}^{+\infty} [F(x)]^{n-1} \; g(x)dx \; , \qquad (7)$$

where F(x) is the area of the noise function to the left of x, g(x) is the ordinate of the signal-plus-noise function, and n is the number of intervals used in the test. This is simply the probability that one drawing from the distribution due to signal plus noise is greater than the greatest of n-1 drawings from the distribution due to noise alone.

It is intuitively clear that if the signal produces a large shift in the noise function, i.e., if d' is large, then the probability that the greatest value of x will be obtained in the interval that contains the signal is also large, and conversely - indeed, P(c) is a monotonic function of d'. It may perhaps be most easily seen that Eq. 7 is a function of d' by noting that, under the assumption of equal variance, the signal plus noise function is simply the noise function shifted by d', that is, g(x) = f(x-d'). Thus d' may be defined in a forced-choice experiment by determining a value of P(c) for some signal intensity and then using Eq. 7 to determine d'. A plot of P(c) versus d', for the case of equal variance, is shown in Figure 14.

Estimates of Signal Detectability Obtained from Different Procedures. According to detection theory, the estimates of d' for a signal and background of given intensities should be the same irrespective of the physhophysical procedure used to collect the data. Thus we may check the internal consistency of the theory by comparing estimates of d' based on yes-no and on forced-choice data. The results of such a comparison have been reported in a previous paper.

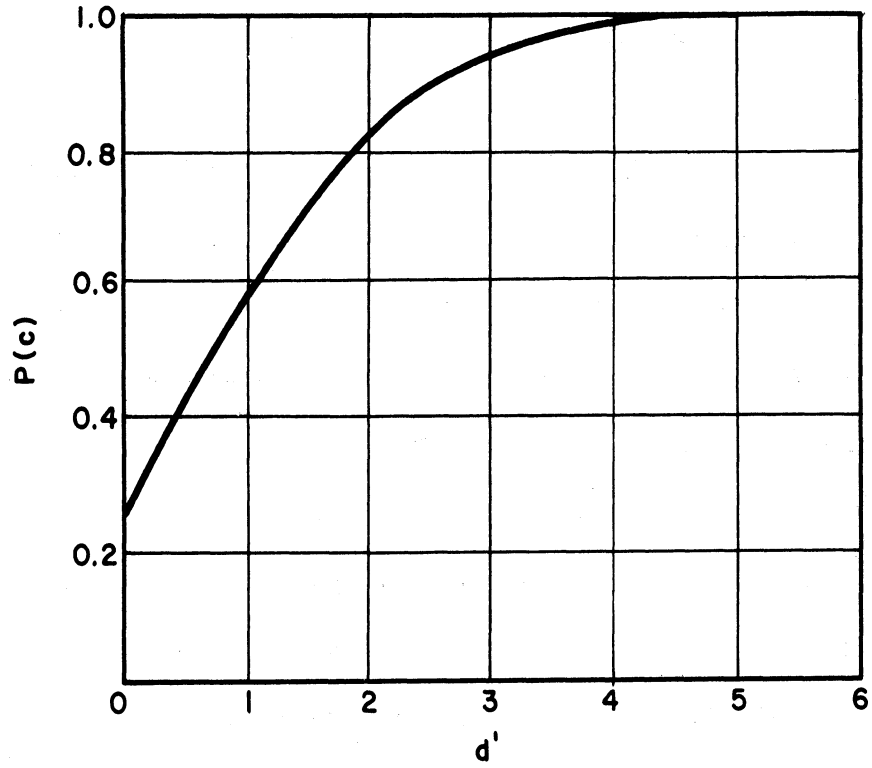

FIG. 14 THE PROBABILITY OF A CORRECT CHOICE IN A FOUR-ALTERNATIVE, FORCED-CHOICE EXPERIMENT AS A FUNCTION OF d'

(Ref. 3) It was shown there that estimates of d' based on the data of the first expected-value experiment that we have presented above and on forced-choice tests conducted in conjunction with it, are highly consistent with each other. Comparable estimates of d' have also been obtained in auditory experiments - from yes-no and forced-choice procedures, and from forced-choice procedures with from two to eight alternatives. (Ref. 13) Hence, decision theory provides a unification of the data obtained with different procedures; it enables one to predict the performance in one situation from data collected from another.

It is a commonplace that calculated values of the stimulus threshold are not independent of the psychophysical procedure that is employed. (Ref. 14) Of particular relevance to our present concern is the finding that thresholds obtained with the forced-choice procedure

are lower than those obtained with the yes-no procedure. (Ref. 15) This finding is accounted for, in terms of decision theory, by the fact that the calculated threshold varies monotonically with the false-alarm rate (see Figure 4) - with high thresholds corresponding to low false-alarm rates such as were obtained in these experiments. The dependence of the stimulus threshold upon the false-alarm rate, however the threshold is calculated, precludes the existence of a simple relationship between thresholds obtained with the yes-no procedure and those obtained with other response procedures. It is also the case that the normalization of the psychophysical function provided by the correction for chance or the normalization achieved by defining the threshold as the stimulus intensity yielding a proportion of correct responses halfway between chance performance and perfect performance, does not serve to relate forced-choice thresholds obtained with different numbers of alternatives.

Theoretical and Experimental Analysis of Second Choices. As we have indicated, a variation of the forced-choice procedure - in which the observer indicates his second choice as well as his first - provides a powerful test of a basic difference between the decision model and the high-threshold model. If the observer is capable of discriminating among values of the observations x that fail to reach what we have termed the threshold, i.e., a criterion fixed at approximately $+3\sigma$ from the mean of the noise function, then the proportion of second choices that are correct will be considerably higher than if he is not.[8]

[8] This experiment was suggested to us by R. Z. Norman, formerly a member of the Electronic Defense Group, now at Princeton University. The general rationale of this experiment, and the results of its application to the perception of words exposed for short durations, have been presented by P. D. Bricker and A. Chapanis, (Ref. 16) and by D. H. Howes, (Ref. 17).

According to the high-threshold model, only very infrequently will more than one of the n observations of a forced-choice trial exceed the threshold. Since the observations which do not exceed the threshold are assumed by the model to be indiscriminable, the second choice will be made among the n-1 alternatives on a chance basis. Thus, for a four-alternative experiment as described in the following, the high-threshold model predicts that, when the first choice is incorrect, the probability that the second choice will be correct is 0.33. This predicted value, it may be noted, is independent of the signal strength.

Decision theory, on the other hand, implies that the observer is capable of ordering the four alternatives according to their likelihood of containing the signal. If this is the case, the proportion of correct second choices will be greater than 0.33. Should one of the samples of the noise function be the greatest of the four, leading to an incorrect first choice, the probability that the observation from the signal plus noise distribution will be the second greatest is larger than the probabilities that either of the observations of the noise distribution will be the second greatest. Again, it is intuitively clear that this probability is a function of d' or of signal strength - that is, the probability that the observation of signal plus noise value will be greater than two of the observations of noise increases with increases in d' or signal strength. Specifically, the probability of a correct second choice in a four-alternative, forced-choice test, for a given value of d', is given by the expression

$$\frac{3\int_{-\infty}^{+\infty}[F(x)]^2\,[1 - F(x)]\,g(x)dx}{1 - \int_{-\infty}^{+\infty}[F(x)]^3\,g(x)dx} \qquad (8)$$

where the symbols have the same meaning as in Eq. 7. This relationship is plotted in Figure 15 under the assumptions that the density functions of noise and signal plus noise are Gaussian and of equal variance. (The function predicted by decision theory for the proportion of correct first choices in a three-alternative situation is included in Figure 15 to show that this function is not the same as the predicted function of the probability of a correct second choice, given an incorrect first choice, for the four-alternative situation.)

To distinguish between the two predictions, data were collected from four observers; two of them had served previously in the second expected-value experiment, whereas the other two had received only routine forced-choice training. Each of the observers served in three experimental sessions. Each session included 150 trials in which both a first and second choice were required. The resulting twelve proportions of correct second choices are plotted against d' in Figure 15. The values of d' were determined by using the proportion of correct first choices as estimates of the probability of a correct choice, P(c), and reading the corresponding values of d' from the middle curve of Figure 15, which is the same curve shown in Figure 14. Although just one value of signal intensity was used (0.78 foot-lamberts as in the second expected-value experiment), the values of d' differed sufficiently from one observer to another to provide an indication of the agreement of the data with the two predicted functions. Additional variation in the estimates of d' resulted from the fact that, for two observers, a constant distance from the signal was not maintained in all three of the experimental sessions.

A systematic deviation of the data from a proportion of 0.33 clearly exists. Considering the data of the four observers combined,

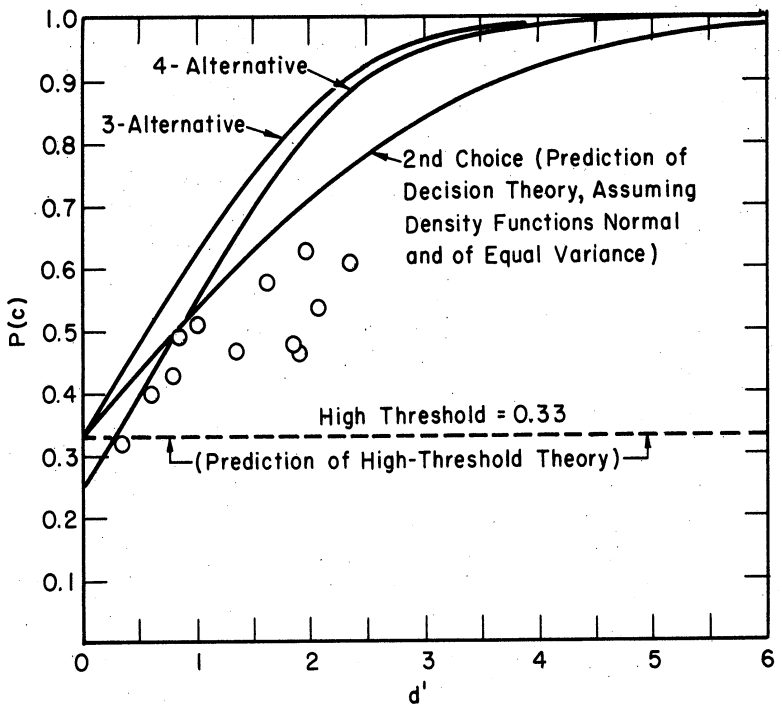

FIG. 15 THE RESULTS OF THE SECOND CHOICE EXPERIMENT

The proportions of correct second choices are plotted against d'.

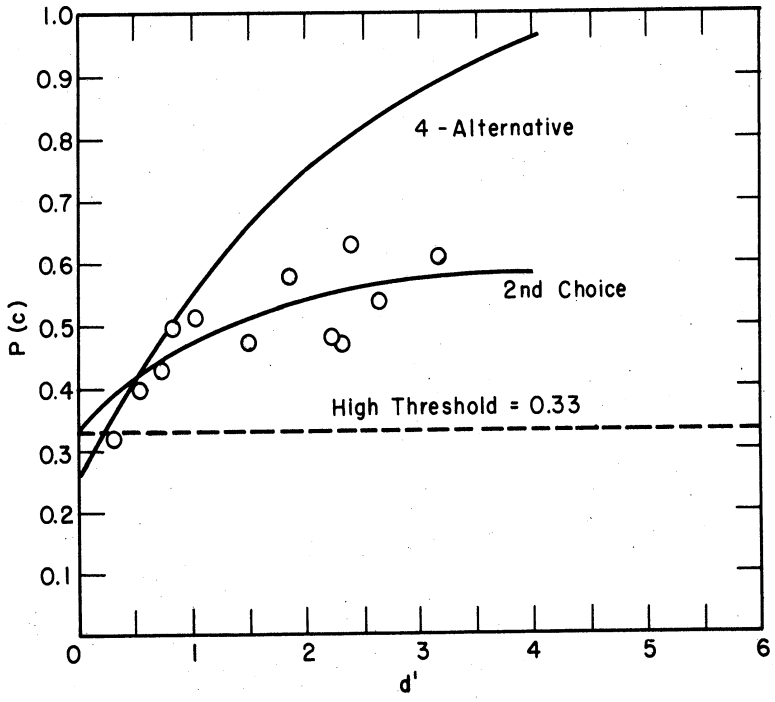

FIG. 16 THE RESULTS OF THE SECOND CHOICE EXPERIMENT CALCULATED UNDER ANOTHER ASSUMPTION

The predictions from decision theory for first and second choices are plotted under the assumption that $\frac{\Delta M}{\Delta \sigma} = 4$.

the proportion of correct second choices is 0.46. Further, a correlation between the proportion of correct second choices and d' is evident.

Two control conditions aid in interpreting these data. The first of these allowed for the possibility that requiring the observer to make a second choice might depress his first-choice performance. During the experiment, blocks of 50 trials in which only a first choice was required were alternated with blocks of 50 trials in which both a first and second choice were required. Pooling the data from the four observers, the proportions of correct first choices for the two conditions are 0.650 and 0.651, a difference that is obviously not significant. A preliminary experiment in which data were obtained from a single observer for five values of signal intensity also serves as a control. 150 observations were made at each value of signal intensity. The relative frequencies of correct second choices for the lowest four values of signal intensity were, in increasing order of signal intensity, 26/117 (0.22), 33/95 (0.35), 30/75 (0.40), and 20/30 (0.67). For the highest value of signal intensity, none of five second choices were correct. In this experiment, then, the proportion of correct signal choices is seen to be correlated with a physical measure of signal intensity as well as with the theoretical measure d' - this eliminates the possibility that the correlation found with a constant value of signal intensity, involving d' as one of the variables (Figure 15), is an artifact of theoretical manipulation.

It may be seen from Figure 15 that the second-choice data also deviate systematically from the predicted function derived from decision theory. This discrepancy, as will be seen, results from the inadequacy of the assumption, i.e., of equal variance of the noise and signal plus noise density functions, upon which the predicted

functions in Figure 15 are based. It was pointed out above that the data obtained in the second expected-value experiment (see Figures 10-13 and accompanying text) indicate that a better assumption would be that the ratio of the increment in the mean of the signal plus noise function to the increment in its standard deviation is equal to 4. Figure 16 shows the second-choice data and the predicted four-alternative and second-choice curves derived from the theory under this assumption that $\frac{\Delta m}{\Delta \sigma} = 4$. In view of the variance associated with each of the points (each first-choice d' was estimated on the basis of 300 observations and each second-choice proportion on less than 100 observations), the agreement of the data and the predicted function shown in Figure 16 is quite good.

The conclusion to be drawn from these results of the second-choice experiment, though perhaps more obvious here, is the same as that drawn from the yes-no, or expected value, experiments: the sensory informatio or the decision axis, is continuous over a greater range than allowed for by the high-threshold model. If a threshold cutoff, below which there is no discrimination among observations, exists at all, it is located in such a position that it is exceeded by much of the noise distribution.

A Note on the Variance Assumption

Before considering the two experiments remaining to be described, we should pause briefly to take up the problem of the relative sizes of the variances of the noise and signal plus noise distributions. We have seen, as indicated in the theoretical discussion, that an assumption concerning these variances may be tested by experiment. We have found that two sets of data, from yes-no and forced-choice experiments, support the assumption that the variance of the signal plus noise dis-

tribution increases with its mean. In particular, the assumption that $\frac{\Delta m}{\Delta \sigma} = 4$ is seen to fit the data reasonably well and noticeably better than the assumption of equal variance. We should like to point out two aspects of this topic in the following paragraphs: first, that these data do not settle the issue, that is, the assumption of $\frac{\Delta m}{\Delta \sigma} = 4$ is probably not generally applicable, and, second, that we have good reason to suspect in advance of experimentation, in the visual case, that the variance of the signal plus noise distribution is greater than that of the noise distribution.

It will be apparent that if the variance of these sampling distributions is a function of sample size, then their variances will differ as a function of the duration and the area of the signal. The assumption of $\frac{\Delta m}{\Delta \sigma} = 4$ will probably not fit the results of experiments with different physical parameters. Further, as we have indicated, we have not explored the extent of agreement between other specific assumptions and our present data. It appears likely that more precise data will be required to determine the relative adequacy of different assumptions about the increase in variance with signal strength.

Peterson, Birdsall and Fox, after developing the general theory of signal detectability, spelled out the specific forms it takes in a variety of different detection problems. (Ref 2) By way of illustration, we may mention the problems in which the signal is known exactly, the signal is known exactly except for phase, and the signal is a sample of white Gaussian noise. A principal difference among these problems lies in the shape of the expected ROC curve. For our present purposes, we may regard these problems as differing with respect to how much variance is contributed by the signal itself. For the first case mentioned, the signal contributes no variance - the signal plus

noise distribution is simply a translation of the noise distribution, the two have equal variances. In the other two cases, the signal itself has a variability which increases with its strength.

Clearly, if we are to select one of the specific models incorporated within the theory of signal detectability to apply to a visual detection problem, we would not select the one that assumes that the signal is known exactly - the visual signal does not contain phase information. Thus, the second model is more likely to be applicable than the first. Actually, the third model, which assumes that the signal is a sample of noise, is the best representation of a visual signal. The point of fundamental interest here is that either of the last two models leads to predicted results quite similar to those that are predicted under the assumption that $\frac{\Delta m}{\Delta \sigma} = 4$. Further discussion of this subject would lead us too far off the path; we would like at this point simply to note that a specific form of the theory of signal detectability, which on a priori grounds is most likely to be applicable to vision experiments, predicts results very similar to those obtained. It is worthwhile to remark in this connection that the results of auditory experiments using pure tones as signals are in close agreement with the signal-known-exactly model, with the assumption of equal variance.

An Analysis of the Rating Scale

We have concluded from the experiments just described that the observer bases his decision on likelihood ratio, or, at the least, on some monotonic function of the likelihood ratio. We might expect then, in the language of decision theory, that he will be able to report the a posteriori probability of signal existence, i.e., that he will be able to state, following an observation interval, the proba-

bility that a signal existed during the interval. In more familiar terms, we are expecting that the observer will be capable of reporting a subjective probability, or of employing a rating scale. Experimental verification of this hypothesis is required, of course, for a reasonable doubt remains whether the observer will be able to maintain the multiple criteria essential to the use of a rating scale. If, for example, six categories of a posteriori probability are used, or a six-point rating scale, the observer must establish five criteria instead of just one as in the yes-no procedure - this may be considerably more difficult.

The ability to make a probability or rating response is of interest, in part, because such a response is highly efficient - in principle, it retains all of the information contained in the observation. In contrast, breaking up the observation continuum into yes and no sections is a process that loses information. From a procedure forcing a binary response, one learns from the observer only that the observation fell above or below a critical value and not how far above or below. In some practical detection problems, the finer-grain information gained from a probability response can be utilized to advantage: the observer may record a posteriori probability so that yes or no decisions concerning the action to be taken can be made at a later time, or by someone else who may be more responsible or who may possess more information about the values and costs of the decision outcomes.

More to the point in terms of our present interests, an experimental test of the ability to make a rating response contributes to the evaluation of decision theory, and also to distinguishing between the adequacy of decision theory and the high-threshold theory. Since the data obtained with a rating procedure may be used to construct ROC curves, this experiment attacks the same problem as the experiments

described above, i.e., whether the observer can discriminate among observations likely to result from noise alone. It is also the case that the rating procedure generates ROC curves, of a given reliability, with a considerable economy of time compared to the yes-no procedure. Therefore, it is of interest, with respect to future applications of decision theory, to determine whether the observer can perform as well, as indexed by d', with the rating procedure as with the yes-no procedure.

The task posed for the observers in this experiment was to place each observation in one of six categories of a posteriori probability. Four categories of equal size (0.2) were used in the range between 0.2 and 1.0; the other two categories were 0.0 to .04 and .05 to .19. The boundaries of the categories were chosen in conference with the observers; they believed that they would be able to operate reasonably within this particular scheme. Actually, the specific sizes of the categories used are not important for most purposes; we can as well think of a six-point rating scale and assume only the property of order.

The four observers in this experiment were the same observers who served in the second expected-value experiment. Further, the same signal intensity (0.78 foot-lamberts) and the same a priori probabilities $P(SN) = P(N) = 0.50$ that were employed in that experiment were employed in this one. The observers made a total of 1200 observations in three experimental sessions.

The Results. The raw data for each observer consist of the number of observations of signal plus noise and the number of observations of noise alone that were placed in each of the six categories of a posteriori probability, and the total number of observations placed in each category. Before proceeding with more complex analyses, we

shall first make a rough determination of the validity of the observers' use of the categories, i.e., of whether we are, in fact, dealing with a scale. This may be achieved by computing the proportion of the total number of observations placed in each category that were actually observations of a signal. If the categories were used properly, this proportion will increase with increases in the probabilities that define the categories.

The results of this analysis are shown in Figure 17. Five curves are plotted there, one for each of the four observers and one showing the average result. We may note, as an aside, that the very neat average curve conceals an interesting individual difference, namely, that Observer 4 is considerably more cautious than the others. A look at the raw data reveals that he uses the lowest category twice to four times as often as the other observers; as a matter of fact, he placed 60% of his observations in that category. We may look for this difference to reappear in other analyses of the data of this experiment. The major point here, however, is that three of the four individual curves are monotonic increasing, whereas the fourth shows only one reversal. This result indicates the feasibility of using a scaling procedure - it indicates that requiring an observer to maintain five criteria simultaneously in a detection problem is not unreasonable. The result is consistent with an ability to order completely the observations, those arising from noise alone as well as those arising from signal plus noise.

The ROC Curves Obtained from the Rating Data. ROC curves can be generated from data obtained with the rating procedure since these data can be compressed to those of the binary-decision procedure with any of several criterion levels. That is to say, we can calculate the pair of

values, $P_N(A)$ and $P_{SN}(A)$, ignoring all but one of the (five) criteria, or category boundaries, employed by the observer. We successively calculate five pairs of these values, each time singling but a different criterion and thus trace out an ROC curve. In particular, we first compute the conditional probabilities that observations arising from noise alone and from signal plus noise will be placed in the top two categories,

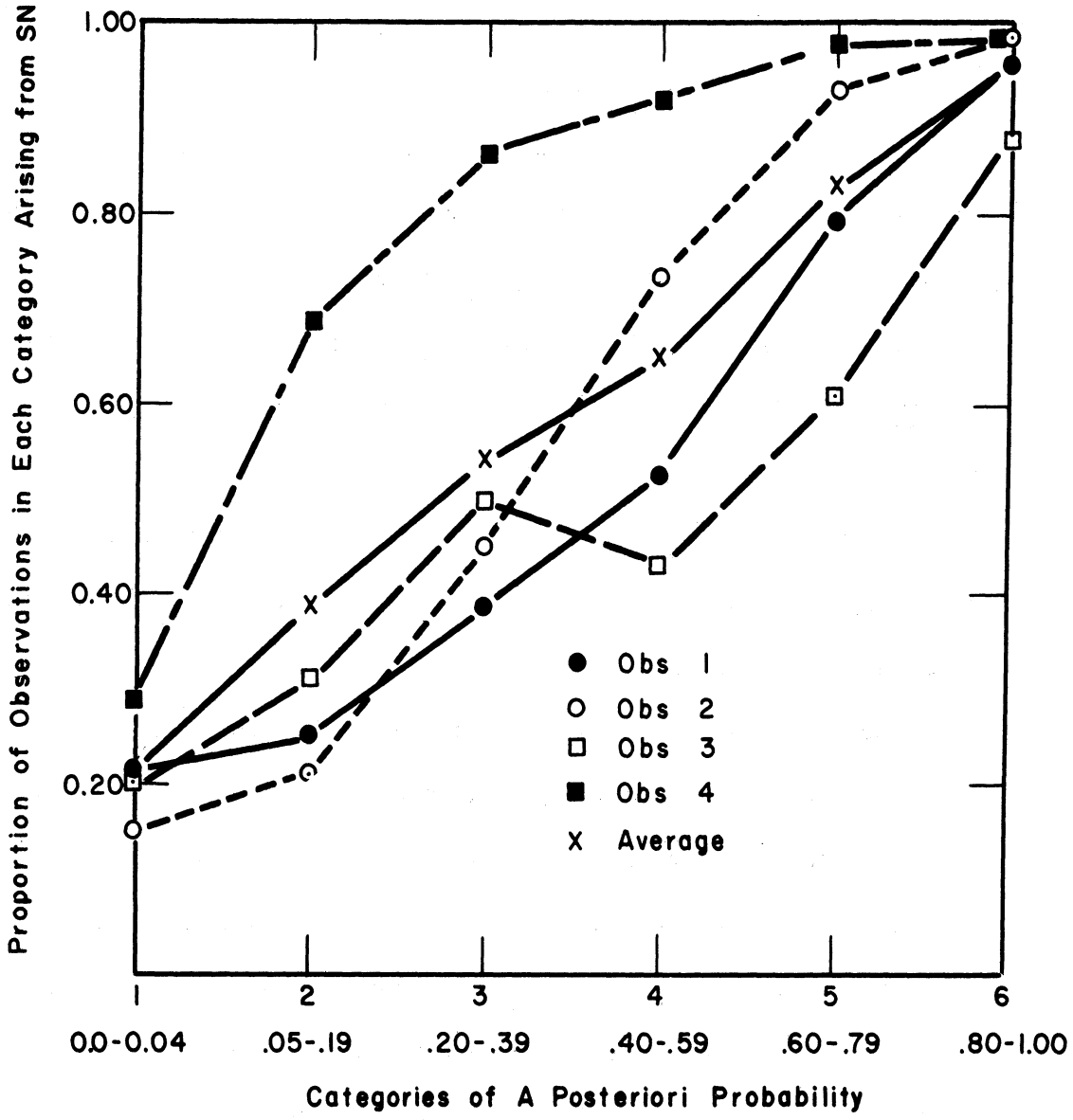

FIG. 17 THE RESULTS OF THE RATING EXPERIMENT

and so forth. We assume, in these calculations, that observations placed in a particular category would fall above the criteria that define a lower category.

The ROC curves so obtained are shown in the upper left-hand portions of Figures 18 to 21. We may note that the data are well described by the type of ROC curve predicted from decision theory. As is the case with the empirical ROC data from yes-no experiments, they cannot be fitted well by a straight line intersecting the point $P_{SN}(A) = P_N(A) = 1.0$, the prediction made from the high-threshold theory. This result indicates that the observers can discriminate among observations likely to result from noise alone and are capable of maintaining the multiple criteria required for the rating response.

A Comparison of ROC Curves Obtained from Ratings and Binary Decisions. It is intuitively clear that an estimate of d' of given reliability can be achieved with fewer observations by the rating procedure than by the yes-no procedure. This proposition is supported by a comparison of the yes-no data shown in Figures 10-13 with the rating data shown in Figures 18-21. The rating data, which show considerably less variation, are based on 1200 observations whereas the yes-no data are based on approximately 5000 observations.

The economy provided by the rating procedure makes it desirable to determine whether the two procedures are equivalent means of generating the ROC curve. Unfortunately, to answer the question immediately, there are some clear differences between the ROC curves we have obtained with the two procedures. These differences are best illustrated by plotting the data on normal coordinates, i.e., on probability scales transformed so that the normal deviates are linearly spaced. As we indicated in the theoretical discussion, these scales are convenient

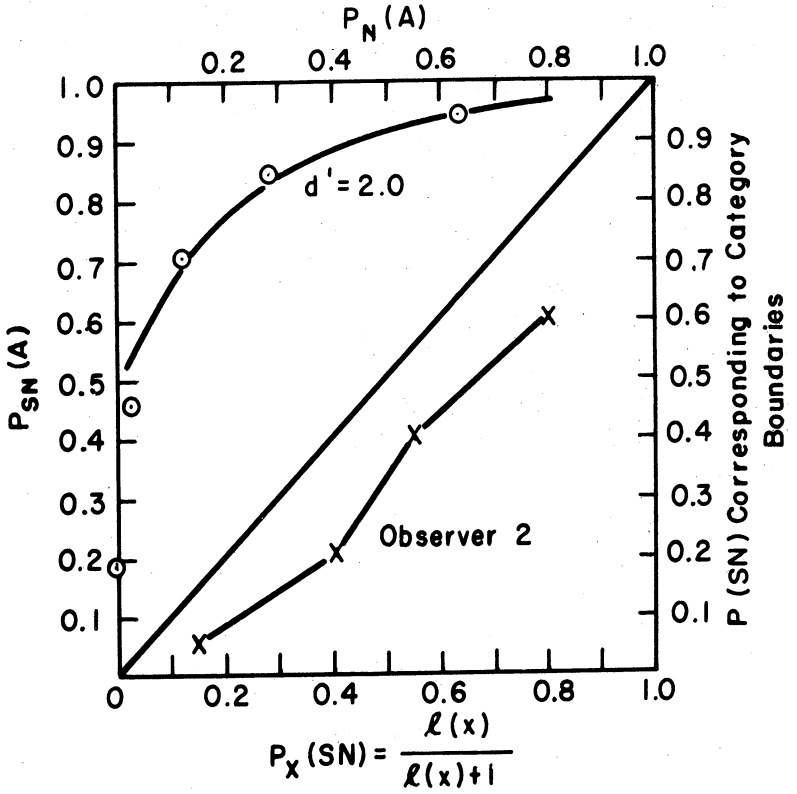

FIG. 18 EMPIRICAL ROC CURVE IN THE RATING EXPERIMENT – TWO-ALTERNATIVE PRESENTATIONS

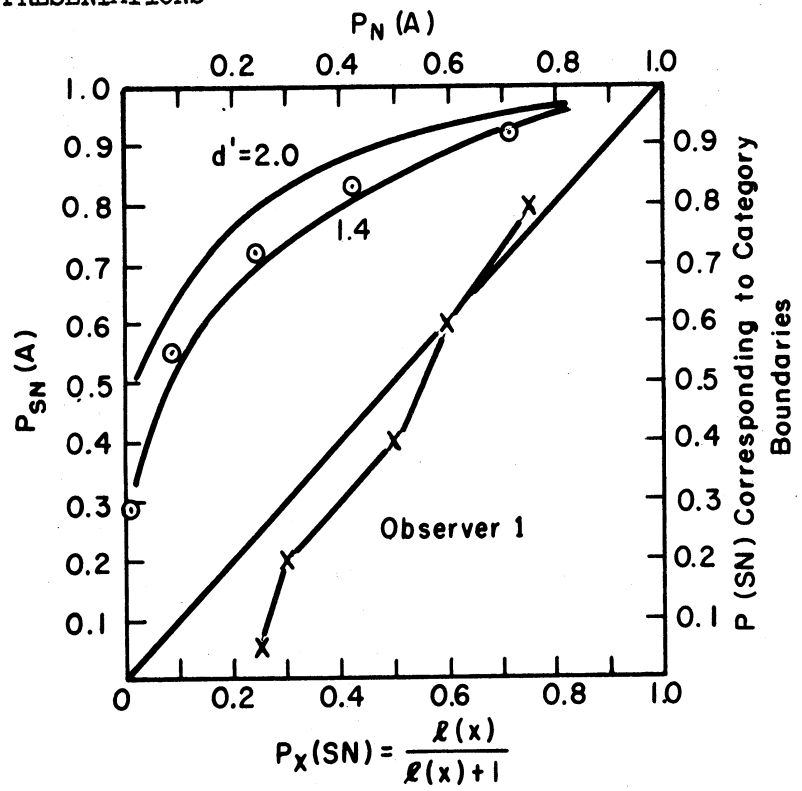

FIG. 19 EMPIRICAL ROC CURVE IN THE RATING EXPERIMENT – TWO-ALTERNATIVE PRESENTATIONS

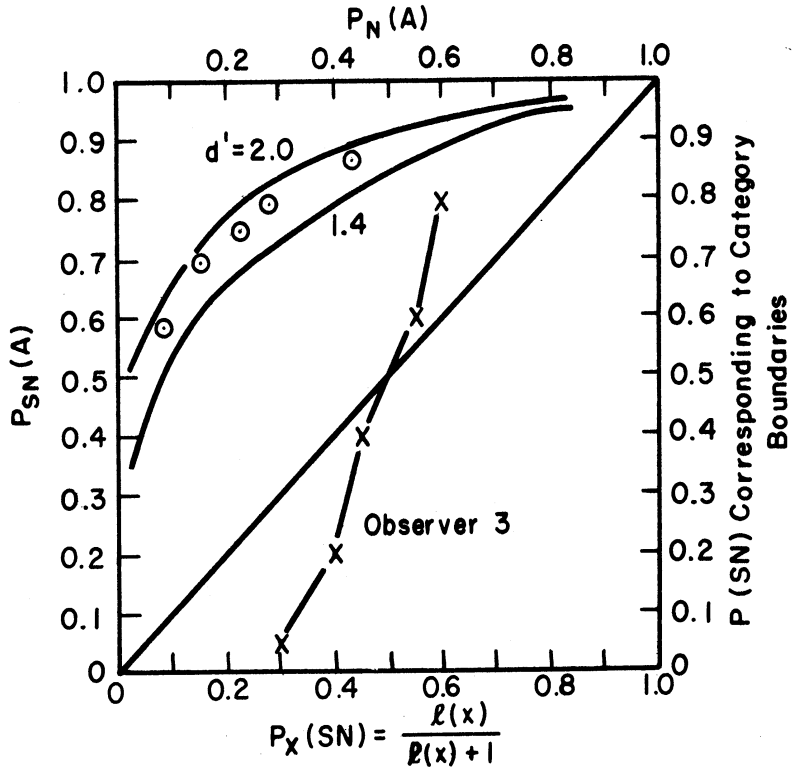

FIG. 20 EMPIRICAL ROC CURVE IN THE RATING EXPERIMENT – TWO-ALTERNATIVE PRESENTATIONS

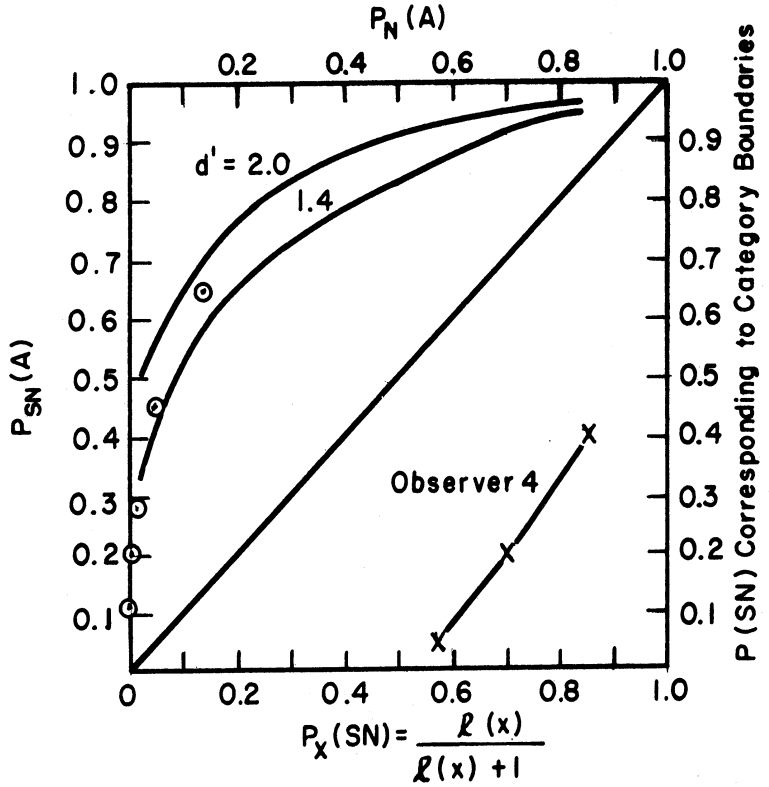

FIG. 21 EMPIRICAL ROC CURVE IN THE RATING EXPERIMENT – TWO-ALTERNATIVE PRESENTATIONS

since on them the ROC curve specified by decision theory becomes a straight line. We have not mentioned it before, but the slope of this line represents the relative variances of the density functions, $f_N(x)$ and $f_{SN}(x)$, that underlie the ROC curve. In particular, it can be shown that the reciprocal of the slope (with respect to the normal-deviate scales) is equal to the ratio $\frac{\sigma SN}{\sigma N}$.

The empirical ROC curves obtained with the rating and yes-no procedures are shown on normal coordinates in Figures 22 to 25. It is immediately evident from these figures that a lower detectability resulted from the rating procedure for all four observers. We may see from the alternative presentations of these data in Figures 10-13 and 18-21 that the values of d' range from 2.0 to 3.0 for the yes-no data and from 1.5 to 2.0 for the rating data.[9] It is further apparent in Figures 22-25 that, consistent with the difference in d', the rating curve has a greater slope than the yes-no curve. This difference is small - the greater variance of $f_{SN}(x)$ under the yes-no procedure did not show clearly in the plots on linear probability axes - but it is regular. We may also note again, as this way of plotting the data makes very clear, that the rating data show considerably less scatter than the yes-no data.

The values and costs associated with the decision outcomes in this situation make us hesitant, on the basis of the data we obtained, to reject the hypothesis that the rating and yes-no procedures are equivalent means of generating ROC curves. It is possible, of course, that

---

[9] Values of d' can, of course, be computed from the normal-deviate scales of the plots in Figures 22-25. A problem arises, however, if the slope of the line fitted to the data is not one. A solution to this problem is proposed by Clarke, Birdsall and Tanner (Ref. 18).

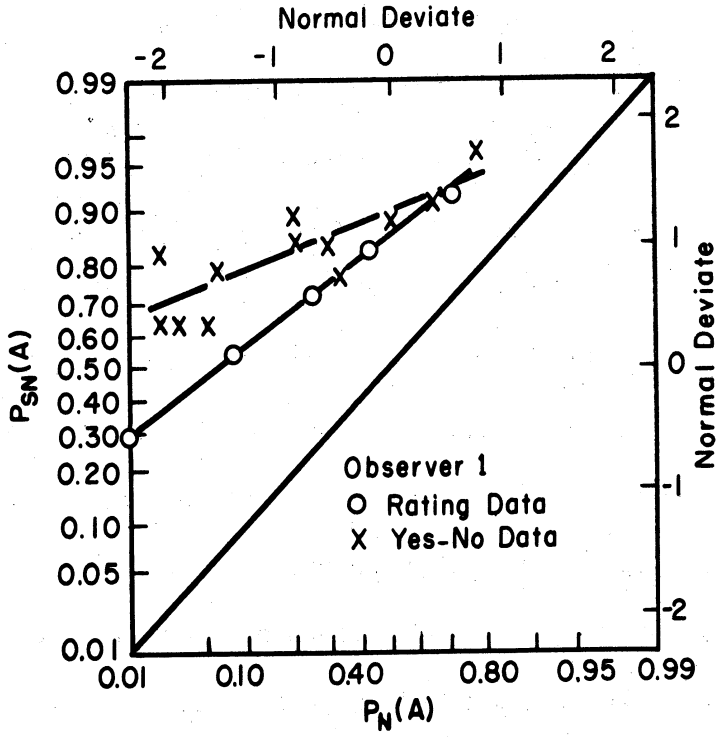

FIG. 22 A COMPARISON OF THE ROC CURVES OBTAINED FROM RATINGS AND BINARY DECISIONS

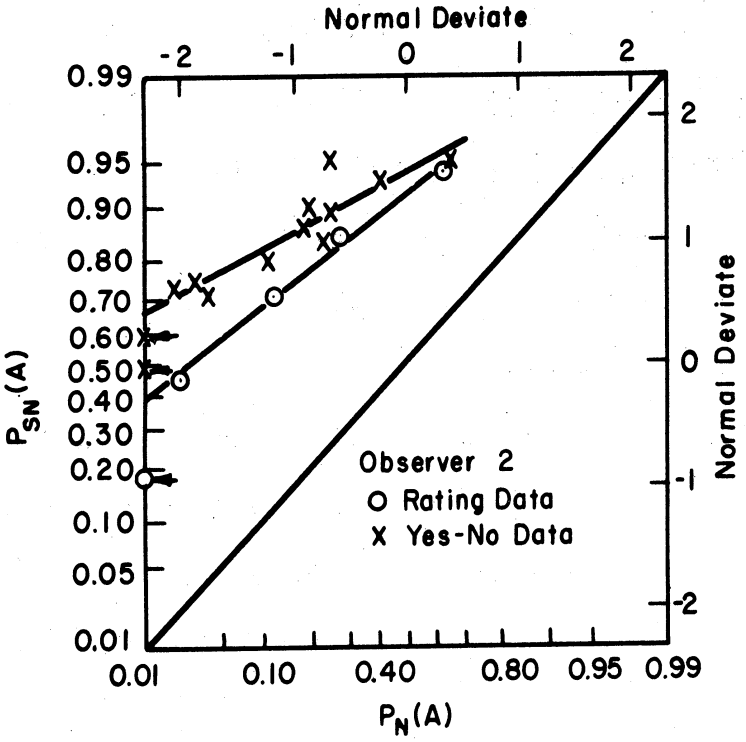

FIG. 23 A COMPARISON OF THE ROC CURVES OBTAINED FROM RATINGS AND BINARY DECISIONS

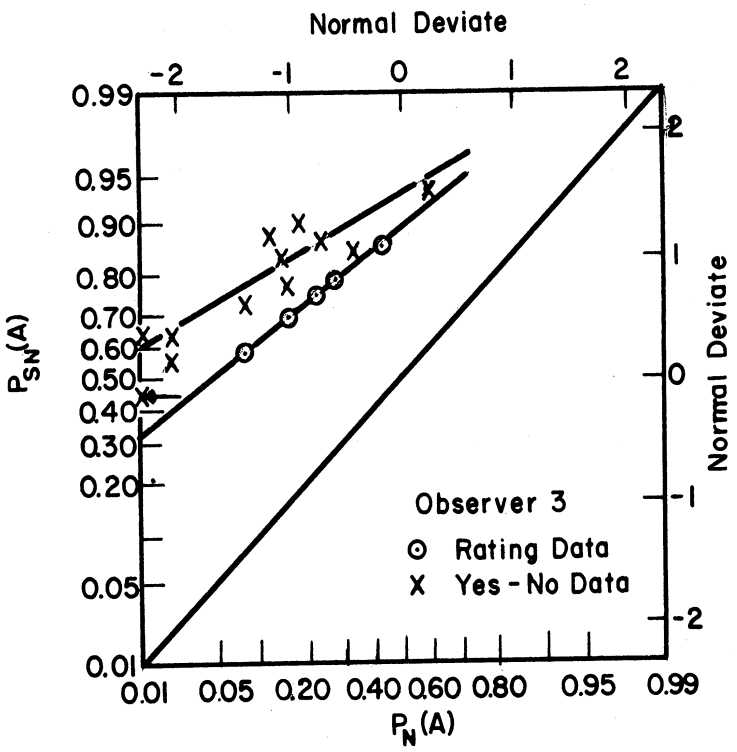

FIG. 24 A COMPARISON OF THE ROC CURVES OBTAINED FROM RATINGS AND BINARY DECISIONS

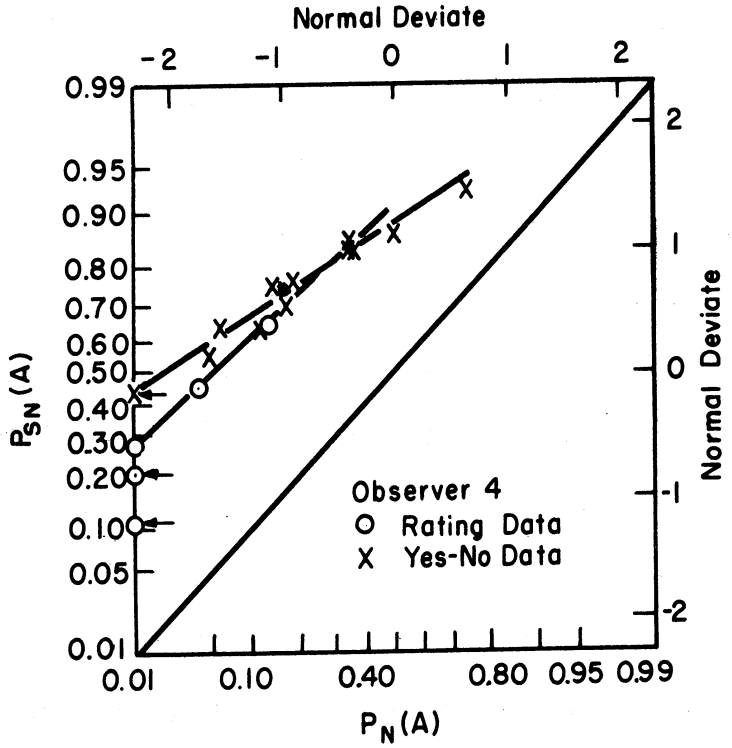

FIG. 25 A COMPARISON OF THE ROC CURVES OBTAINED FROM RATINGS AND BINARY DECISIONS

some undetected difference existed between the experimental conditions in the two experiments; one was conducted after the other was completed. Such a difference might easily account for the relatively small discrepancies observed. Again, it has recently been shown in an auditory experiment that the two procedures result in essentially the same ROC curve, both with respect to d' and to slope (Ref. 19). Still, we can not discount the present results on the basis of the auditory experiment, for we have noted several differences between visual and auditory data that are likely to be real - one perhaps relevant to this issue is that the ROC curves obtained with pure tones have slopes that are uniformly near one. We should perhaps be content, at this point, with the admittedly weak conclusion that no data exist to support the hypothesis that the two procedures are equivalent in the case of visual stimuli.

A Test of the Normality of the Density Functions. At this juncture, it is convenient to turn briefly, but explicitly, to a topic first considered in the theoretical discussion. It was stated there that we would assume the density functions on the observer's decision axis to be Gaussian in form, but that the assumption was subject to experimental test. A test of this assumption is provided by plotting the empirical ROC curves on normal coordinates. Having now introduced plots of the data in this form in Figures 22-25, we may use them for this purpose. If the observer's density functions are normal, then the empirical points of an ROC curve plotted on normal coordinates will be fitted best by a straight line. Clearly, a straight line provides the most adequate description of the data in these figures. Thus the assumption of normality, an important one for the sake of simplicity of analysis, is consistent with the data.

The Approach to Optimal Behavior

In the presentation of experimental results thus far, we have concentrated on the continuity of the observer's decision axis, and on his ability to adopt various criteria along this axis. A remaining question is how closely the criteria he adopts correspond to those specified by decision theory as the optimal criteria. To answer this question we shall consider some further analyses of experimental results already described and the results of an additional experiment.

It should be recalled that decision theory specifies as the optimal decision function either likelihood ratio, $\ell(x)$, or some monotonic function of likelihood ratio, call it $\ell(x)'$. That is to say, any transformation of the decision axis is acceptable as long as order is maintained. If the decision function is $\ell(x)$, then the optimal criterion is the value of $\ell(x)$ equal to β (Eq. 3). If the decision function is $\ell(x)'$, then the optimal criterion is the value of this function that corresponds to β, call it β'. The monotonic relationship means that $\ell(x)' > \beta' \leftrightarrow \ell(x) > \beta$. Thus to establish the applicability of decision theory, it is necessary to demonstrate only that the observer's criteria are monotonically related to β. If sampling error is taken into account, it is sufficient to demonstrate a significant correlation between the observer's criteria and β. It is of interest, however, to determine just how closely the observer's criteria do approach the optimal criteria as specified by β. In examining this question we shall make use of the fact that, in order to index the observer's criterion, it is not strictly necessary to compute a value of likelihood ratio from the proportions of hits and false alarms - it is more convenient, and for purposes of interpretation, more direct, to take simply the proportion of false alarms as the index.

The Criteria Employed in the Expected-Value Experiments. In the first expected-value experiment, the observers were told only the a priori probabilities of signal and noise and the values of the various decision outcomes that were in effect during each experimental session. They were not told that any combination of these factors can be expressed by a single number (β) which, in conjunction with a value of d', specifies the optimal criterion or the optimal false-alarm rate. The rank-order correlations between β and the obtained proportions of false alarms that were computed from the data of this first study were .70, .46 and .71 for the three observers respectively. A correlation of .68 is significant at the .01 level of confidence. This result indicates that the observer did not merely vary his criterion from one session to another, but that his criterion varied appropriately with changes in β.

In the second expected-value experiment, the observers were told the optimal proportion of false alarms for each session as well as the a priori probabilities and decision values. This information was available to the experimenter since values of d' had previously been determined by the forced-choice procedure during a training period. Thus, in the second study, we were asking how closely the observer would approach the optimal false-alarm rate given knowledge of it. The rank-order correlations between the false-alarm rates announced as optimal and the false-alarm rates yielded by the four observers were .94, .97, .86 and .98. Again, a coefficient of .68 is significant at the .01 level of confidence. Data obtained later in an auditory experiment showed coefficients of this magnitude - as a matter of fact, the rank-order coefficient based on five pairs of measures for each of two observers in the auditory experiment was 1.0 - when the

observers were not informed of the optimal false-alarm rate (Ref. 20).

Satisfying a Restriction on the Proportion of False Alarms. A more direct attack on the question of the observer's ability to reproduce a given false-alarm rate is provided by an experimental procedure not previously described in detail, one involving a different definition of optimal behavior. Under this definition of optimal behavior, no values and costs are assigned the various decision outcomes; instead, a restriction is placed on the proportion of false alarms permitted. The optimal behavior is to maximize the proportion of hits while satisfying the restriction on false alarms. This, it will be recognized, is the procedure most popular among experimenters for testing statistical hypotheses.

An experiment using this procedure was conducted with a different set of four observers. The a priori probability of signal occurrence was 0.72 throughout the experiment. There were, then, fourteen presentations of noise alone in a block of fifty presentations. There were four different experimental conditions, each extending over eighteen blocks of fifty presentations. In each of these conditions, the observers were instructed to adopt a criterion that would result in 'yes' responses to n or to n+1 of the fourteen presentations of noise alone in a block of fifty presentations. For the four conditions of the experiment, n was equal to 0, 3, 6 and 9, respectively. Thus the acceptable range for the proportion of false alarms was 0.0 - .07, .21 - .28, .43 - .50. or .64 - .71. The primary data consist of four values of false-alarm rate for each observer; each value is based upon 252 presentations of noise alone.

The data are shown in Figure 26. The false-alarm rates obtained are plotted against the restricted ranges of false-alarm rate. The

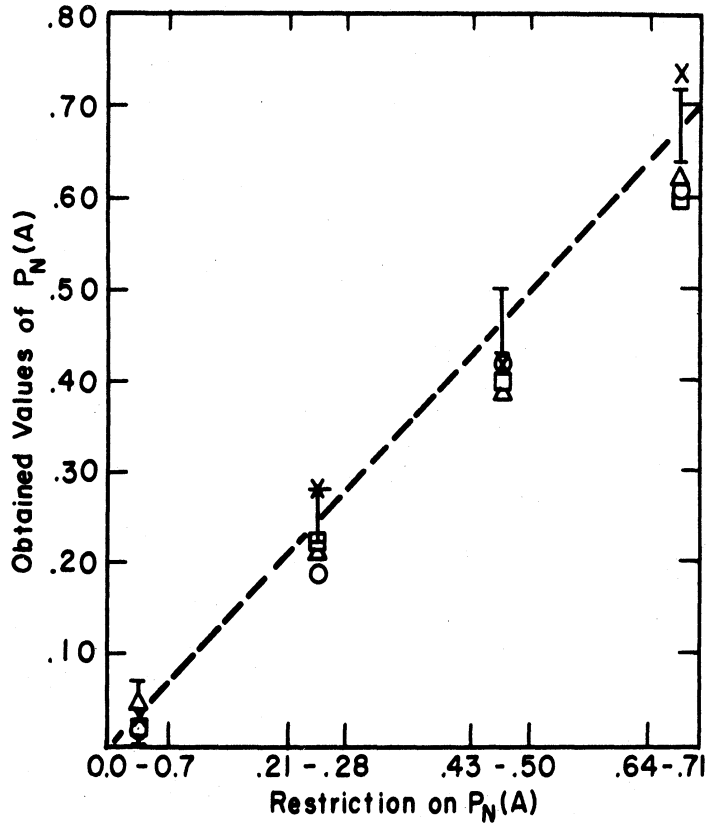

FIG. 26 THE REPRODUCTION OF A GIVEN FALSE-ALARM RATE

four observers are represented by different symbols; the vertical bars designate the acceptable range. It may be seen that the largest deviation from the range stipulated is .04. This result suggests that the observer is able to adjust his criterion with considerable precision.

Two other pieces of information are needed, however, to interpret the data shown in Figure 26. For, of course, if the observer were given information about the correctness of his response after each response, these data could be obtained even if the observer were unable to vary his criterion. The observer could then approximate any false-alarm by rate by saying 'yes' until the desired number of false alarms was achieved and then saying 'no' on the remaining presentations. That procedure would entail a severe depression of d'. Actually, the obser-

vers were given information about correctness only after each block of fifty presentations, and the values of d' were not depressed. Thus the false-alarm rates that were obtained may legitimately be regarded as reflecting the observer's criteria.

The Criteria Employed in the Rating-Scale Experiment. We may also investigate how closely the multiple criteria adopted by the observers in the rating-scale experiment approach the optimal criteria. Stated otherwise, we may examine the relationship that existed between the subjective and objective probabilities of signal occurrence in that experiment. It may be noted in advance that an alternative presentation of the results, in Figure 17, gives an indication of the extent of agreement we may expect.

As stated earlier, the a posteriori probability of signal existence is a monotonic function of likelihood ratio. In particular, the optimal relationship between the two is

$$P_X(SN) = \frac{\ell(x)\ P(SN)}{\ell(x)\ P(SN) + P(N)} \tag{9}$$

where $P_X(SN)$ denotes the probability that the signal existed given the observation x, i.e., the a posteriori probability; where $\ell(x)$ is the likelihood ratio; and where P(SN) and P(N) are the a priori probabilities (Ref. 2). For our experiment, with P(SN) = P(N) = .50, this equation reduces to

$$P_X(SN) = \frac{\ell(x)}{\ell(x) + 1} \tag{10}$$

As described above, a point on the ROC curve can be obtained for each of the boundaries of the six categories employed by the observer, i.e., for the five criteria he employed. Since, as we have also pointed

out, the criterion value of $\ell(x)$ corresponds to the slope of the ROC curve at the point in question, this criterion value of $\ell(x)$ can be determined. Thus $P_X(SN) = \frac{\ell(x)}{\ell(x) + 1}$ can be computed for each of the criteria employed by the observer. Assuming now that the observer's decision function is likelihood ratio, then if he is behaving according to the optimal relationship between $P_X(SN)$ and $\ell(x)$, the values of $\frac{\ell(x)}{\ell(x) + 1}$ computed from his data will correspond directly to probability values that were announced as defining the categories. In short, we know the values of $P_X(SN)$ that were announced as marking off the categories; by pursuing a route through the empirical ROC curve and $\ell(x)$ we can calculate the values of $P_X(SN)$ that bound the categories the observer actually used - therefore, we can assess how well the two sets of criterion values of $P_X(SN)$, the objective and subjective probabilities, agree.

The lower right-hand portions of Figures 18-21 show the probability values that were announced as defining the categories, plotted against the probability values that characterize the criteria actually employed by the observers, i.e., against $P_X(SN) = \frac{\ell(x)}{\ell(x) + 1}$ as determined from the data. [Some points are missing since $\ell(x)$ is indeterminate at very low values of $P_N(A)$.] It is apparent from these plots that Observers 1, 2 and 3 are operating with a decision function similar to likelihood ratio and approximately according to the optimal relationship between $P_X(SN)$ and $\ell(x)$. The pattern exhibited by Observers 1 and 3, that of over-estimating small deviations from a probability of 0.5, will be familiar to those acquainted with the literature on subjective probability. Observer 4, as we noted earlier, is quite different from the others. His tendency, also evidenced but to a far

lesser extent by Observer 2, is to consistently underestimate the a posteriori probability, i.e., to set all of his criteria too high.

To summarize our discussion of how nearly the criteria adopted by the observers in these several experiments correspond to the optimal criteria, we may say that the observer, for want of a better term, behaves in "an optimal fashion." He is responsive to changes in both the a priori probability of signal occurrence and the values of the decision outcomes; the criteria he adopts are highly correlated with the optimal criteria. Subjective transformations of the real probability scale and of the "real" value scale do, of course, exist and differ somewhat from one observer to another. Undoubtedly, values also play a role in those experiments in which no values are explicitly assigned by the experimenter. Nevertheless, we have seen that the observer can adopt successively as many as ten different criteria, on the basis of different combinations of probabilities and values presented to him, that are almost perfectly ordered. He can maintain simultaneously at least five criteria that are a reasonable facsimile of the optimal criteria. If he is told the optimal false-alarm rate, he can approximate it with a very small error. It is probably fair to conclude that the human observer can be trusted, more than was generally expected, to adopt an appropriate criterion.

III. SUMMARY, CONCLUSIONS, AND REVIEW OF IMPLICATIONS

We imagine the process of signal detection be choice between two Gaussian variables. One, having a mean equal to zero, is associated with noise alone; the other, having a mean equal to d', is associated with signal plus noise. In the most common detection problem the

observer decides, on the basis of an observation that is a sample of one of these populations, which of the two alternatives existed during the observation interval. The particular decision that is made depends upon whether or not the observation exceeds a criterion value; the criterion, in turn, depends upon the observer's detection goal and upon the information he has about relevant parameters of the detection situation. The accuracy of the decision that is made is a function of the variable d' which is monotonically related to the signal strength.

This description of the detection process is an almost direct translation of the theory of statistical decision. The main thrust of this conception, and of the experiments that support it, is that more than sensory information is involved in detection. Conveniently, a large share of the non-sensory factors are integrated into a single variable, the criterion. There remains a measure of sensitivity (d') that is purer than any previously available, a measure largely unaffected by other than physical variables. This separation of the factors that influence his sensitivity is the major contribution of the psychophysical application of statistical decision theory.[10]

---

[10] It is interesting to note that the present account is not the first to model psychophysical theory after developments in the theory of statistical decision - as a matter of fact, Fechner was influenced by Bernoulli's suggestion that expectations might be expressed in terms of satisfaction units. As Boring relates the story, Bernoulli's interest in games of chance led him to formulate the concept of "mental fortune;" he believed changes in mental fortune to vary with the ratio of the change in physical fortune to the total fortune. This mathematical relationship between metnal and physical terms was the sort of relationship that Fechner sought to establish with his psychophysics (Ref. 21). It should also be observed that Fechner anticipated the decision model under discussion in a much more direct way. His concept of "negative sensations," largely dismissed by subsequent workers in the field, denies the existence of such a cut in the continuum of observations that the magnitudes of observations below the cut are indiscriminable.

We have indicated several times in the preceding that another conception of the detection process - one involving what we termed a high threshold - is inconsistent with the data reported. It should be noted, however, that these data, to the extent analyzed in this paper, do not preclude the existence of a lower threshold. The analyses presented do not indicate explicitly how far down into the noise the observations are being ordered, i.e., how low a threshold must be relative to the noise distribution in order to be compatible with the data. As it happens, further analyses of the yes-no and forced-choice results show them to be consistent with a threshold slightly above the mean of the noise distribution. Of course, a determination of the level at which a threshold may possibly exist is neither critical nor useful. A threshold well within the noise distribution is not a workable concept. Such a concept, since it is inconsistent with the correction for chance, complicates rather than facilitates the mathematical treatment of the data. Moreover, a threshold that low is, for practical purposes, not measurable. The forced-choice experiment is a case in point; the observer converys less information that he is capable of conveying if only a first choice is required. That the second choice contains a significant amount of information has been demonstrated; auditory experiments have shown that the fourth choice conveys information (Ref. 20). Thus it is difficult to determine when enough information has been extracted to yield a valid estimate of a low threshold. In addition, the existence of such a threshold is of little consequence for the application of the decision model, e.g., yes-no data resulting from a supra-threshold criterion depend upon the criterion but are completely independent of the threshold value.

One of the major reasons for our concern with the threshold concept is that this concept supports several common psychophysical procedures that are invalidated by the results we have described. The correction for chance success has already been mentioned as a technique that stems from a high-threshold theory and one that is inconsistent with the data. This correction is frequently applied to data collected with the method of constant stimuli. It is used implicitly whenever the threshold is defined as the stimulus intensity that yields a probability of correct response halfway between chance and perfect performance. The method of adjustment and the standard method of serial exploration are also inappropriate given the mechanism of detection described above. When the method of serial exploration is used with the signal always present, or with insufficient "catch trials" to estimate the probability of a false alarm, the raw data will not permit separating the variation in the observer's criterion from variation in his sensitivity. Changes in an observer's criterion from one session to another can be estimated only if it is assumed that his sensitivity has not changed, and conversely. The same applies to data collected with the method of adjustment. To be sure, unrecognized variations in the criterion are not important in many psychophysical measurements for they may be expected to contribute relatively little variation to the computed value of the threshold. Fairly large changes in the criterion will affect the threshold value by less than 3 db in the case of vision, and by no more than 6 db in the case of audition. This degree of reliability is acceptable in clinical audiometry, for example, in which the method of limits is usually employed. Neither would it distort appreciably curves of the course of dark adaptation. In many experiments, however,

i.e., in experiments concerned with substantive as well as with theoretical problems, a reliability of less than one db is required, and in these cases a knowledge of the criterion used by the observer is essential.

By way of illustrating the problems in which the threshold concept and its associated procedures may have led to improper conclusions, we may single out one of current interest, the problem of "subliminal perception." In most of the studies of this phenomenon, the evidence for it consists of the finding that subjects who first report seeing no stimulus can then identify the stimulus with greater than chance accuracy when forced to make a choice.[11] We have mentioned above as a typical result in psychophysical work that the forced-choice procedure yields lower threshold values than does the yes-no procedure. We also suggested that this result may be accounted for by the fact that with the yes-no procedure the calculated value of the threshold varies directly with the observer's criterion and that a strict criterion is usually employed by observers under this procedure. That a strict criterion is usually used with the yes-no procedure is not surprising in view of the fact that observers are often instructed to avoid making false-alarm responses. It is also likely that the stigma associated with "hallucinating" promotes the use of a strict criterion in the absence of an explicit caution against false alarms. Thus it may be expected that on many occasions when an observer does not choose to report the existence of the stimulus, he nevertheless possesses some

---

[11] This procedure was used explicitly in the earlier studies of subliminal perception; several of these studies are reviewed by J. G. Miller (Ref. 22). With minor variations, this procedure also underlies many of the more recent studies; see, e.g., Bricker and Chapanis (Ref. 16).

information about it. It may be, therefore, that "subliminal perception" exists only when a high criterion is incorrectly identified as a limen.[12]

Having presented a theory of detection behavior in Part I and some detection experiments in Part II, and having just discussed the relationship of this work to "psychophysics," it remains to articulate with the title and the introductory paragraph of this paper, to consider the relationship of the work to the study of "perception."

In principle, the general scheme we have outlined may apply to perception as well as to detection. It seems reasonable to suppose that perception is also a choice among Gaussian variables. Consistent with the existence of many alternatives in the case of perception, we may imagine many critical regions to exist in the observation space. This space will have more dimensions than are involved in detection - as we have previously indicated, one less dimension than the number of alternatives considered. We may presume, in perception as in detection, that the boundaries of the critical regions are defined in terms of likelihood ratio, and are determined by the a priori probabilities of the alternatives and the relative values of the decision outcomes.

It may also be contended that what we have been referring to as a detection process is itself a perceptual process. Certainly, if perceptual processes are to be distinguished from sensory processes on the grounds that the former must be accounted for in terms of events presumed to occur at higher centers whereas the latter can be

---

[12] This analysis of the problem of subliminal perception has been elaborated by I. Goldiamond (Ref. 23).

accounted for in terms of events occurring within the receptor systems, then the processes with which we have been concerned qualify as perceptual processes. Since, in detecting signals, the observer's detection goal and the information he possesses about probabilities and values play a major role, we must assume either that signal detection is a perceptual process or that the foregoing distinction between sensory and perceptual processes is of little value. Thus the thesis of the present paper is, in one of its aspects, another stage in the history of the notion that the process of perceiving is not merely one of passively reflecting events in the environment, but one to which the perceiver himself makes a substantial contribution. Various writers have suggested that our perceptions are based upon unconscious inferences, that sensory events are interpreted in terms of unconscious assumptions about their probable significance, that our responses to stimuli reflect the influence of our needs and expectancies, that we utilize cues in selectively placing sensory events in categories of identity, and so forth. The present view differs from these in regarding the observer as relating this sense data to information he has previously acquired and to his goals in a manner specified by statistical decision theory. The approach from decision theory has the advantage that it specifies the perceiver's contribution to perception at other than the conversational level; it provides quantitative relationships between the non-sensory factors and both the independent and dependent variables.

We submit, then, that the present paper, although confined to detection experiments, is aptly named. We may view detection and perception as made of the same cloth. Of course, signal detection is a relatively simple perceptual process, but it is exactly its

simplicity that makes the detection setting most appropriate to a preliminary examination of the value of statistical-decision theory for the study of perception. Because detection experiments permit precise control over the variables specified by the theory as pertinent to the perceptual process, they provide the rigor desirable in the initial tests of a theory. Once these tests are passed, the theory may be extended and applied to more complex problems. Problems that have been studied recently within the framework of decision theory include the recognition of one of two signals, combined detection and recognition, problems in which a single decision is based on a series of observations, problems in which the observer decides sequentially whether to make another observation before making a final decision and the recognition of speech (Refs. 24-32).

REFERENCES

1. See, for example, A. Wald, Statistical Decision Functions, 1950.

2. W. W. Peterson, T. G. Birdsall, and W. C. Fox, The theory of signal detectability, Trans. of the IRE, PGIT-4, 1954, 171-212. See also D. Van Meter and D. Middleton, Modern statistical approaches to reception in communication theory, Trans. of the IRE, PGIT-4, 1954, 119-145.

3. W. P. Tanner, Jr., and J. A. Swets, A decision-making theory of visual detection, Psychol. Rev., 61, 1954, 401-409.

4. W. P. Tanner, Jr., and T. G. Birdsall, Definitions of d'and η as psychophysical measures, J. Acoust. Soc. Amer., 30, 1958, 922-928.

5. L. L. Thurstone, A law of comparative judgment, Psychol. Rev., 34, 1927, 273-286.

6. _____, Psychophysical analysis, Amer. J. Psychol., 38, 1927, 368-389.

7. M. Smith and Edna Wilson, A model of the auditory threshold and its application to the problem of the multiple observer, Psychol. Monogr., 67, No. 9, 1953.

8. W. A. Munson and J. E. Karlin, The measurement of the human channel transmission characteristics, J. Acoust. Soc. Amer., 26, 1956, 542-553.

9. C. E. Shannon, The mathematical theory of communication, Bell System Technical J., 27, 1948, 379-423.

10. Horton, Fundamentals of Sonar, 1957.

11. H. R. Blackwell, Unpublished manuscript.

12. H. R. Blackwell, B. S. Pritchard, and T. G. Ohmart, Automatic apparatus for stimulus presentation and recording in visual threshold experiments, J. Opt. Soc. Am., 44, 1954, 322-26.

13. J. A. Swets, Indices of signal detectability obtained with various psychophysical procedures, J. Acoust. Soc. Amer., 31, 1959, 511-513.

14. See, for example, C. E. Osgood, Method and Theory in Experimental Psychology, 1952.

15. H. R. Blackwell, Psychophysical Thresholds: Experimental Studies of Methods of Measurement, Engineering Research Institute Bulletin No. 36, The University of Michigan, 1953.

16. P. D. Bricker and A. Chapanis, Do incorrectly perceived tachistoscopic stimuli convey some information?, Psychol. Rev., 60, 1953, 181-188.

17. D. H. Howes, A statistical theory of the phenomenon of subception, Psychol. Rev., 61, 1954, 98-110.

18. F. R. Clarke, T. G. Birdsall, and W. P. Tanner, Jr., Two types of ROC curves and definitions of parameters, J. Acoust. Soc. Amer., 31, 1959, 629-630.

19. J. P. Egan, A. I. Schulman, and G. Z. Greenberg, Operating characteristics determined by binary decisions and by ratings, J. Acoust. Soc. Amer., 31, 1959, 768-773.

20. W. P. Tanner, Jr., J. A. Swets, and D. M. Green, Some general properties of the hearing mechanism, Tech. Rept. No. 30, Electronic Defense Group, The University of Michigan, Ann Arbor, 1955.

21. E. G. Boring, A History of Experimental Psychology, 1950 (2nd edition), 285.

22. J. G. Miller, Unconsciousness, 1942.

23. I. Goldiamond, Indicators of perception: I. Subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical indicator methodology, Psychol. Bull., 55, 1958, 373-411.

24. W. P. Tanner, Jr., A theory of recognition, J. Acoust. Soc. Amer., 28, 1956, 882-888.

25. J. A. Swets and T. G. Birdsall, The human us of information: III. Decision-making in signal detection and recognition situations involving multiple alternatives, Trans. of the IRE, IT-2, 1956, 138-165.

26. J. A. Swets, Elizabeth F. Shipley, Mary J. McKey, and D. M. Green, Multiple observations of signals in noise, J. Acoust. Soc. Amer., 31, 1959, 514-521.

27. J. A. Swets and D. M. Green, Sequential observations by human observers, to be published in J. Acoust. Soc. Amer.

28. J. P. Egan, F. R. Clarke, and E. C. Carterette, On the transmission and confirmation of messages in noise, J. Acoust. Soc. Amer., 28, 1956, 536-550.

29. J. P. Egan and F. R. Clarke, Source and receiver behavior in the use of a criterion, J. Acoust. Soc. Amer., 28, 1956, 1267-1269.

30. J. P. Egan, Monitoring task in speech communication, J. Acoust. Soc. Amer., 28, 1957, 482-489.

31. I. Pollack and L. R. Decker, Confidence ratings, message reception, and the receiver operating characteristic, J. Acoust. Soc. Amer., 30, 1958, 286-292.

32. L. Decker and I. Pollack, Confidence ratings and message reception for filtered speech, J. Acoust. Soc. Amer., 30, 1958, 432-434.

The Theory of Signal Detectability[1]

Frank R. Clarke, The University of Michigan

The detectability of a signal may be thought of as a property of the signal and the background in which the signal is presented. The performance of an observer in the detection of a signal is affected both by the detectability of the signal and by external variables related to the criteria of the observer. Thus many types of performance measures may not be good measures of the detectability of a signal simply because they are affected by criterial variables which are independent of the detectability of the signal as such.

In this chapter we shall examine the concept of a mathematically ideal observer. In so doing we shall isolate all of the variables which are of importance to an ideal observer in the detection of signals in a background of noise. The manner in which these variables affect the performance of the ideal observer will be stated explicitly, and a performance measure which reflects the detectability of the signal independent of criterial variables will be developed.

The theory of signal detectability views the problem of signal detection as one of testing statistical hypotheses. If the actual observation may arise either from signal plus noise or from noise alone, it is necessary to define a region in some decision space such that an observation falling within this region will lead to the acceptance of the hypothesis signal plus noise; while an observation falling in the complement of this

---

[1] The material in this chapter is based largely on papers by Peterson, Birdsall, and Fox. (Refs. 1-4) The purpose of this chapter is to present some of their ideas in a fairly simple manner. The resulting lack of mathematical rigor and any possible errors must be attributed to the author and not to Peterson, Birdsall, and Fox.

region will lead to an acceptance of the hypothesis noise alone. In the first section of this chapter we shall define this region. It will be shown that for a large class of possible criteria, the ideal observer should work on a unidimensional decision axis, namely, likelihood ratio. The actual observation may be translated into a likelihood ratio. If this value is greater than some well-specified critical value, the observer should accept the hypothesis signal plus noise. Otherwise, the ideal observer should accept the hypothesis noise alone. This is a general result. However, in order to calculate likelihood ratio the parameters of the signal and of the noise must be specified. In order to study the behavior of the ideal observer, it is convenient to assume a Fourier series band-limited white Gaussian noise and a signal which is contained entirely within the bandwidth of the noise.

In the second section of this chapter some of the properties of a Fourier series band-limited voltage waveform will be investigated. It will be seen that such a waveform may be specified exactly by a finite number of amplitude samples drawn from the waveform. This convenient fact will be relied upon in the subsequent section where we shall calculate the actual distributions of likelihood ratio in the case where a signal specified exactly is to be detected in a background of white Gaussian noise. We shall see that the separation of the two distributions, that for signal plus noise and that for noise alone, will depend only on the energy of the signal and the noise power per unit bandwidth.

In Section IV we will investigate some of the implications of the development up to this point. The manner in which some particular variables affect the performance of the ideal observer will be examined. There will be specified a clear separation between those variables which

affect the detectability of the signal as such and those other variables which affect some performance measures without affecting the detectability of the signal in any way.

The remaining two sections of this chapter will contain derivations of likelihood ratio distributions in a simple recognition problem and in the case where the signal to be detected is itself a sample of white Gaussian noise.

I. LIKELIHOOD RATIO AS A DECISION AXIS

Consider a sample which is drawn from one of two possible distributions. On the basis of this sample, which we shall term an observation, the observer must attempt to identify the distribution from which the sample was drawn. For notational convenience, though with no intention of restricting the generality of the problem, let us say that one of the distributions is associated with noise alone while the other is associated with signal plus noise. Let us denote a decision region, A, such that if the observation falls within this region the observer accepts the hypothesis that the observation was drawn from the signal plus noise distribution. If the observation falls in the complement of this region, CA, the observer accepts the hypothesis that the observation was drawn from the noise alone distribution. There are, then, four possible outcomes to this experiment: 1) The observation is drawn from the signal plus noise distribution and the observer accepts the hypothesis signal plus noise. This outcome will occur with some probability which we shall denote $P(SN \cdot A)$. 2) The observation is drawn from the signal plus noise distribution and the observer accepts the hypothesis noise alone. Here the associated probability will be denoted by $P(SN \cdot CA)$. 3) With probability, $P(N \cdot A)$, the

observation will be drawn from the noise alone distribution and the observer will accept the hypothesis signal plus noise. 4) With probability, $P(N \cdot CA)$, the observation will be drawn from the noise alone distribution and the observer will accept this hypothesis. As these are the only possible outcomes of the experiment we have:

$$P(SN \cdot A) + P(SN \cdot CA) + P(N \cdot A) + P(N \cdot CA) = 1 \qquad (1)$$

We shall be particularly interested in conditional probabilities and it is clear that the possible outcomes of the experiment may also be defined in terms of the following three equations:

$$P(SN) + P(N) = 1.0 \qquad (2)$$

$$P_{SN}(A) + P_{SN}(CA) = 1.0 \qquad (3)$$

$$P_{N}(A) + P_{N}(CA) = 1.0 \qquad (4)$$

The behavior of the observer may be defined in terms of the probabilities in Eqs. (3) and (4), for the probabilities in Eq. (2) are not under his control. Furthermore, as both Eq. (3) and (4) sum to one, the behavior of the observer may be specified by only two probabilities. For this purpose we shall use $P_{SN}(A)$ and $P_{N}(A)$.

Now we wish to determine how the region A should be chosen. We wish to choose a region which will, in some sense, lead to optimal decisions on the part of the observer. For example, we may require that the observer maximize the expected value of the experiment, or that he maximize rate of information transmission, or we may require that the observer minimize the total error in the experiment, etc. In any given experiment, an observer which maximizes one of these measures will not necessarily maximize any of the other measures. Yet in all cases, the decision region can be shown to be a unidimensional decision axis with the region A defined as all points on this axis which are above some critical cut-off value.

The decision axis is likelihood ratio.

The Weighted-Combination Criterion:

We shall be able to show that many possible criteria may be expressed as a weighted-combination criterion. Thus we shall first demonstrate that the decision space in this case is a simple unidimensional axis, i.e., likelihood ratio.

The weighted-combination criterion requires that we choose an optimum region, A_{opt}, such that:

$$P_{SN}(A) - \omega P_N(A) = a \max. \tag{5}$$

The observation, for the moment let us call it $\underline{x}$, will fall in some space. The probability, $P_{SN}(A)$, will then be given by an integration over all points falling within some region of acceptance, A, in the case where $\underline{x}$ arose from signal plus noise. The quantity, $P_N(A)$, may be defined in a similar manner. Consequently, we require a critical region, A_{opt}, such that:

$$\int_A f_{SN}(x)\, dx - \omega \int_A f_N(x)\, dx = a \max; \tag{6}$$

which may be written,

$$\int_A [f_{SN}(x) - \omega f_N(x)]\, dx = a \max. \tag{7}$$

It is clear that we wish to choose the region, A, such that it will include all $\underline{x}$ for which $f_{SN}(x) - \omega f_N(x)$ is positive and exclude all $\underline{x}$ such that this quantity is negative. What is done when this quantity is zero is irrelevant. Thus, for the weighted-combination criterion, we should choose a region, A, for accepting the hypothesis signal plus noise such that:

$$f_{SN}(x) \geq \omega f_N(x) \tag{8}$$

or

$$\ell(x) = \frac{f_{SN}(x)}{f_N(x)} \geq \omega \tag{9}$$

Thus, regardless of the dimensionality of the distributions, $f_{SN}(x)$ and $f_N(x)$, from which the observation $\underline{x}$ is drawn, we see that we can maximize the weighted-combination criterion merely by calculating the likelihood ratio for the given observation $\underline{x}$ and comparing this value with some critical value of likelihood ratio; in this case, ω. If $\ell(x) \geq \omega$, we accept the hypothesis signal plus noise. If $\ell(x) < \omega$, we accept the hypothesis noise alone. In mapping from some multidimensional space down to a single decision axis, we have lost much information regarding the observation $\underline{x}$, but none of this information is relevant to the task of estimating which distribution gave rise to the observation.

We shall now show that other commonly used criteria are special cases of the weighted-combination criterion. Consequently, the above result is a fairly general one.

Siegert's Ideal Observer

The task of the observer is to respond in such a way as to minimize total error. This is, of course, equivalent to maximizing proportion of correct responses. Thus, we desire to choose a region, A, such that:

$$P(SN \cdot A) + P(N \cdot CA) = a \max; \tag{10}$$

which we may rewrite as,

$$P(SN)\, P_{SN}(A) + P(N) - P(N)\, P_N(A) = a \max. \tag{11}$$

But the observer's behavior has no effect on P(N); consequently, satisfying Eq. (11) is equivalent to requiring that:

$$P_{SN}(A) - \frac{P(N)}{P(SN)} P_N(A) = a \max. \tag{12}$$

This expression is seen to be the weighted-combination criterion where $\omega = P(N)/P(SN)$. Thus, in the case where the observer's task is to minimize total error, he should choose a region of acceptance such that $\ell(x) \geq P(N)/P(SN)$.

The Neyman-Pearson Criterion

Under this criterion the observer is to hold $P_N(A) \leq k$ and, at the same time, to maximize $P_{SN}(A)$. When dealing with continuous functions, $P_N(A)$ will be taken as equal to $\underline{k}$ in order to make the region of acceptance as large as possible and consequently to maximize $P_{SN}(A)$. Then there must exist a $\omega = W$, such that $\int_A f_N(x)\, dx = k$, where A consists of all $\underline{x}$ such that $\ell(x) > W$. As we saw above, this will maximize the quantity, $\int_A f_{SN}(x)\, dx - W \int_A f_N(x)$; and, as the right-hand member of this expression is a constant in this case, it is clear we have satisfied the condition, $\int_A f_{SN}(x)\, dx$, a maximum. Thus, the Neyman-Pearson criterion is seen to be a likelihood ratio criterion where A consists of all points of $\underline{x}$ for which $\ell(x) \geq W$ and W is chosen such that $\int_A f_N(x)\, dx = k$.

Expected Value Criterion

Here the observer's task is to maximize the expected value of the experiment: where $V_{SN \cdot A}$ is the value of a correct decision; $V_{N \cdot CA}$ is the value of a correct rejection; $K_{SN \cdot CA}$ is the cost of a miss; and $K_{N \cdot A}$ is the cost of a false alarm. The expected value, E(V), of the experiment is given by the following equation:

$$E(V) = P(SN \cdot A)\, V_{SN \cdot A} + P(N \cdot CA)\, V_{N \cdot CA} - P(SN \cdot CA)\, K_{SN \cdot CA} - P(N \cdot A)\, K_{N \cdot A} \quad (13)$$

Now we wish this expression to be a maximum and we may rewrite it as:

$$E(V) = P(SN)\, [V_{SN \cdot A} + K_{SN \cdot CA}]\, P_{SN}(A) - P(N)\, [V_{N \cdot CA} + K_{N \cdot A}]\, P_N(A)$$
$$- P(SN)\, K_{SN \cdot CA} + P(N)\, V_{N \cdot CA} = \text{a max.} \quad (14)$$

As the last two terms in Eq. (14) are constants and not affected by the observer's behavior, the problem of maximizing Eq. (14) is equivalent to requiring that:

$$P_{SN}(A) - \frac{P(N)\ [V_{N\cdot CA} + K_{N\cdot A}]}{P(SN)\ [V_{SN\cdot A} + K_{SN\cdot CA}]}\ P_N(A) = \text{a max.} \qquad (15)$$

But this is seen to be the weighted-combination criterion. Thus in this case, the observer should accept the hypothesis signal plus noise whenever:

$$\ell(x) \geq \frac{P(N)\ [V_{N\cdot CA} + K_{N\cdot A}]}{P(SN)\ [V_{SN\cdot A} + K_{SN\cdot CA}]}$$

Other Criteria

It may be shown that still other types of criteria reduce to the weighted-combination criterion. However, the above are the most common types of criteria utilized and further examples will not be considered.

A Posteriori Probability

Though this is not a criterion as such, it may be desirable in some instances to require the observer to report a posteriori probability. That is, given an observation x, we may require that the observer state the probability that the observation x arose from signal plus noise. The following equations show that $\ell(x)$ may be converted to a posteriori probability if the observer has knowledge of a priori probabilities. By definition:

$$P_x(SN) = \frac{P_{SN}(x)\ P(SN)}{P_{SN}(x)\ P(SN) + P_N(x)\ P(N)} \qquad (16)$$

and dividing both numerator and denominator by $P_N(x)$, we obtain:

$$P_x(SN) = \frac{\ell(x)\ P(SN)}{\ell(x)\ P(SN) + P(N)} \qquad (17)$$

Summary

We have shown in this section that for a large class of criteria

the observer may order his observations in terms of likelihood ratio with no loss of information relevant to the decision process. In discriminating between two hypotheses, the decision axis is unidimensional regardless of the dimensionality of the inputs giving rise to the observation $\underline{x}$.

II. THE FOURIER SERIES BAND-LIMITED WAVEFORM

We wish to investigate the behavior of an ideal observer in the task of detecting signals in a background of noise. We have seen that such an ideal observer should utilize likelihood ratio as a decision axis. However, in order to actually calculate likelihood ratio, we must have knowledge of the distributions, $f_{SN}(x)$ and $f_N(x)$, which are involved. These distributions are, of course, dependent upon the parameters of the signal and of the noise. It is particularly convenient to work with a white Gaussian noise which has a Fourier series band-limit. We shall also regard the signal as Fourier series band-limited. This type of a signal and of noise leads to a fairly uncomplicated calculation of likelihood ratio. The degree to which this type of signal and of noise can betaken as representative of the signals and noises actually used in laboratory experiments is discussed in the following chapter.

Let us consider some voltage waveform, $x(t)$, over some interval of time, 0 to T. When we state that this voltage waveform is Fourier series band-limited, we simply mean that it can be represented exactly by a finite series of the form:

$$x(t) = \frac{a_o}{2} + \sum_{n=1}^{N} \left(a_n \cos \frac{2\pi nt}{T} + b_n \sin \frac{2\pi nt}{T} \right) \qquad (18)$$

Consequently, the continuous waveform, $x(t)$, can be precisely specified by $2N + 1$ numbers, or, identically, by a single point in a $2N + 1$ dimensional space.

The Fourier series is generally introduced in the study of periodic waveforms. However, one need not have a periodic function in order to perform a Fourier analysis. We are considering a voltage waveform, $x(t)$, on the interval 0 to T. The observation occurs during this interval and the waveform which occurs during this interval may be analyzed in terms of a Fourier series without regard for what occurs outside of this interval. Thus, in substituting into the Fourier series, $\underline{t}$ is only allowed to take values from 0 to T and nothing is said about the function outside of the interval. The function is not defined outside of the interval and the observer must base his decision solely on the waveform occuring within the interval 0 to T.

Now the observer is presented with a voltage waveform and not with a Fourier series as such. It can be shown that if W is any low-pass bandwidth large enough to include all of the frequency components of the waveform, then $x(t)$ may be completely specified by $2WT$ amplitude values of the waveform. It should be intuitively clear that, if a waveform can be completely specified by some particular set of $2N + 1$ numbers, it should be possible to specify that waveform by some other set of $2N + 1$ numbers, or $2WT$ numbers where W is chosen such that $2WT \geq 2N + 1$. It can be shown that almost any set of $2WT$ amplitude samples of the waveform, $x(t)$, on the open interval, 0 to T, will completely specify that waveform. Thus, the waveform may be represented by a point in a $2WT$ dimensional space.

The Sampling Theorem

If the waveform, x(t), can be specified by 2WT equally spaced samples on the open interval, 0 to T, then it is possible to compute the energy of the waveform by utilizing these amplitude samples. Though a rigorous derivation is possible (Ref. 1), we shall merely state the result:

$$\text{Energy} = \int_0^T [x(t)]^2 \, dt = \frac{1}{2W} \sum_{i=1}^{2WT} (x_i)^2 \qquad (19)$$

Now x_i is the amplitude of the voltage waveform, x(t), at the i^{th} sampling point. It is assumed that the voltage is applied across a one ohm resistor.

III. THE DETECTABILITY OF A SIGNAL SPECIFIED EXACTLY

We shall derive an expression for the detectability of a signal specified exactly. It should be recalled that we are now studying an ideal observer - a mathematical concept.

An observation interval, 0 to T, is defined. During this observation interval, a voltage waveform, x(t), will be presented. With probability, P(SN), x(t) = s(t) + n(t), where s(t) is the signal to be detected and n(t) is noise. With probability, P(N) = 1 - P(SN), x(t) = n(t). The waveform, s(t), is Fourier series band-limited. The waveform, n(t), arises from a Fourier series band-limited white Gaussian noise. The waveform of the signal is precisely specified and known by the observer. On the basis of the observation, x(t), the observer must state whether or not a signal was present in the interval, i.e., whether x(t) arose from noise alone or from signal plus noise.

We have seen above that for a large class of criteria, the observer should base his decision on likelihood ratio. Thus, we are interested

in calculating likelihood ratio in this case and determining how likelihood ratio is distributed given that $x(t) = s(t) + n(t)$ and how it is distributed given $x(t) = n(t)$.

It is convenient to work with amplitude samples of the voltage waveform, $x(t)$. Consider the i^{th} sampling point which will have an amplitude, x_i, associated with it. If the experiment were repeated many times with $x(t) = n(t)$, then x_i would be normally distributed (since the noise is Gaussian) with mean zero and variance equal to the noise power, N. If the experiment were repeated many times with $x(t) = s(t) + n(t)$, then x_i would be normally distributed with a mean of s_i and variance N. Here s_i is the amplitude of the signal, $s(t)$, at the i^{th} sampling point. Thus, we may calculate the likelihood ratio at the i^{th} sampling point by the following equation:

$$\ell(x_i) = \frac{f_{SN}(x_i)}{f_N(x_i)} = \frac{\frac{1}{\sqrt{2\pi N}} e^{-\frac{1}{2N}(x_i - s_i)^2}}{\frac{1}{\sqrt{2\pi N}} e^{-\frac{1}{2N}x_i^2}} = e^{-\frac{1}{2N}(s_i^2 - 2x_i s_i)} \quad (20)$$

As there are 2WT independent samples[2] which completely specify the waveform $x(t)$, we may write the likelihood ratio for the entire observation as:

$$\ell(x) = \prod_{i=1}^{2WT} \ell(x_i) = e^{-\frac{1}{2N}\left[\sum_{i=1}^{2WT} s_i^2 - 2\sum_{i=1}^{2WT} x_i s_i\right]} \quad (21)$$

We have stressed likelihood ratio as the proper decision axis for the ideal observer. However, it should be clear that any monotonic transformation of likelihood ratio will serve equally well; for, if the decision rule states that a particular decision should be made if $\ell(x) > W$, then

---

[2] The amplitude samples are independent if W is chosen as the low-pass bandwidth of the noise and if the noise is white over that frequency range. Naturally it is assumed that the signal is contained entirely within this bandwidth for if it were not, the ideal observer would merely need to look at those frequency components outside of the noise bandwidth and detection would be perfect.

the same decision will be made if we require that a particular monotonic transformation of $\ell(x)$ be greater than the same monotonic transformation of W.

Our development will be simplified if we consider $\log_e \ell(x)$. Applying this monotonic transformation to Eq. (21), we have

$$\log \ell(x) = \frac{1}{N} \sum_{i=1}^{2WT} x_i s_i - \frac{1}{2N} \sum_{i=1}^{2WT} s_i^2. \tag{22}$$

Now we may ask how $\log \ell(x)$ is distributed when the observation interval contains signal plus noise and how this quantity is distributed when the observation interval contains noise alone.

When the interval contains signal plus noise, we have $x_i = s_i + n_i$ and Eq. (22) becomes

$$[\log \ell(x)]_{SN} = \frac{1}{2N} \sum_{i=1}^{2WT} s_i^2 + \frac{1}{N} \sum_{i=1}^{2WT} s_i n_i. \tag{23}$$

The first term in this equation is a constant and we see by the sampling theorem (Eq. 19) that it is equal to WE/N where W is the bandwidth of the noise, E is the energy of the signal and N is the noise power. In the second term of Eq. (23), n_i is a normally distributed variable with mean zero and variance N. Hence, the expected value of this term is zero. The variance of the i^{th} component is s_i^2/N and total variance is $\frac{1}{N}\sum_{i=1}^{2WT} s_i^2$ or $2E/N_o$. As n_i is a Gaussian variable, we see that $[\log \ell(x)]_{SN}$ is itself a Gaussian variable with a mean of E/N_o and a standard deviation of $(2E/N_o)^{1/2}$, where N_o is the noise power per unit bandwidth, N/W.

A similar development for the case in which noise alone is presented in the observation interval yields the result that $[\log \ell(x)]_N$ is a Gaussian variable with a mean of $-E/N_o$ and a standard deviation of $(2E/N_o)^{1/2}$.

We may normalize these distributions by adding E/N_o to all values of log $\ell(x)$ and dividing all values by the standard deviation, $\frac{2E}{N_o}$, which is common to both distributions. We then have the distributions shown in Figure 1. The detectability of the signal is given by the separation of the means of the normalized distributions and is dependent only on the energy of the signal and the noise power per unit bandwidth. We shall use this separation between the two distributions as our index of detectability, d', when

$$d' = \frac{2E}{N_o} \tag{24}$$

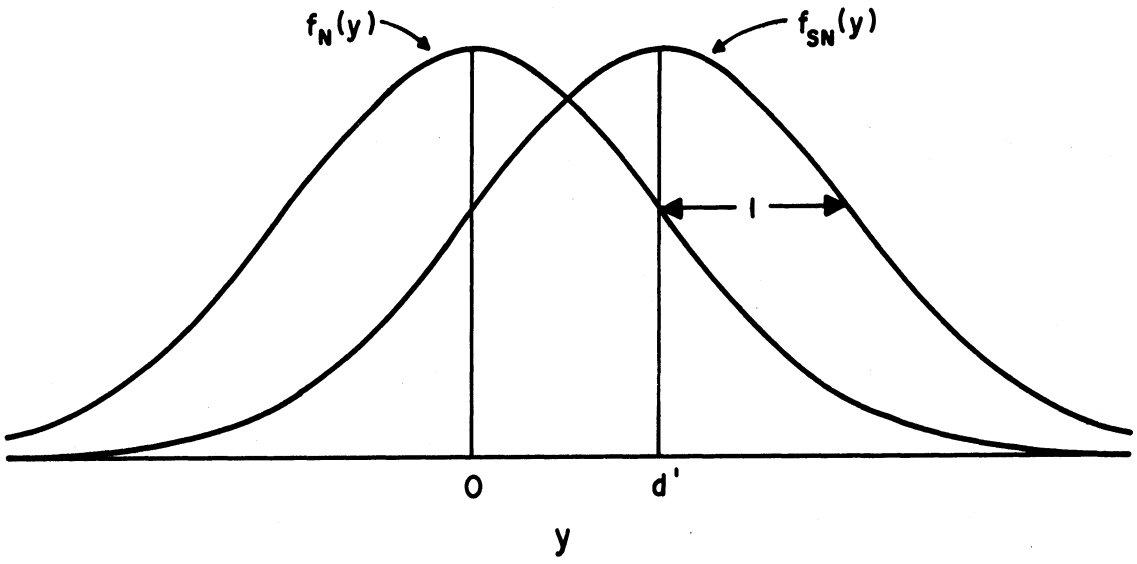

FIG. 1. For convenience we have labeled the decision axis "y", where y is a monotonic transformation of likelihood ratio,

$$y = [\log \ell(x) + E/N_o] \,/\, (2E/N_o)^{1/2}.$$

Given noise alone, y is distributed normally with a mean of zero and unit variance, N(0,1). Given signal plus noise, y is N(d',1).

IV. THE RECEIVER OPERATING CHARACTERISTIC

In this section we will examine the relation between the measure of detectability, d', and some other types of performance measures. In Section I we saw that only two probabilities were necessary to specify the behavior of the ideal observer. These were the probability of a "hit," $P_{SN}(A)$, and the probability of a "false alarm," $P_N(A)$. These two values are affected by both the detectability of the signal and the criterion of the observer.

In order to illustrate the manner in which these variables are related to d', let us refer to the distributions shown in Fig. 1. Now for some cut-off value, Y, on the decision axis y, the probability of a "hit" is given by the following equation:

$$P_{SN}(A) = P_{SN}(y > Y)$$

and the probability of a "false alarm" is given by:

$$P_N(A) = P_N(y > Y).$$

This is illustrated in Fig. 2.

The particular cut-off value of Y which should be adopted by the ideal observer is a function of the a priori probability of a signal and of various criterion variables as shown in Section III. With the detectability of the signal fixed, that is, d' is a constant, both $P_N(A)$ and $P_{SN}(A)$ may vary from zero to unity. It is clear from Figure 2 that, as the cut-off value Y varies from plus infinity to minus infinity, both $P_N(A)$ and $P_{SN}(A)$ will increase monotonically, though at different rates, from zero to one. The relation between these two variables for various values of d' is shown in Figure 3. These curves are called ROC-

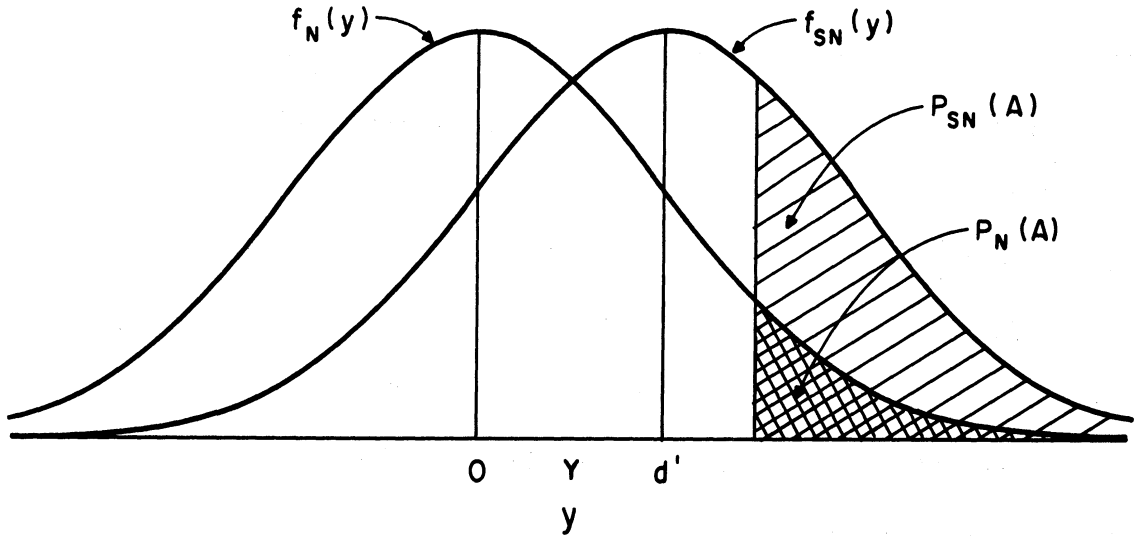

FIG. 2. The probabilities of a "hit," $P_{SN}(A)$, and of a "false alarm," $P_N(A)$, for some particular criterion cut-off, Y.

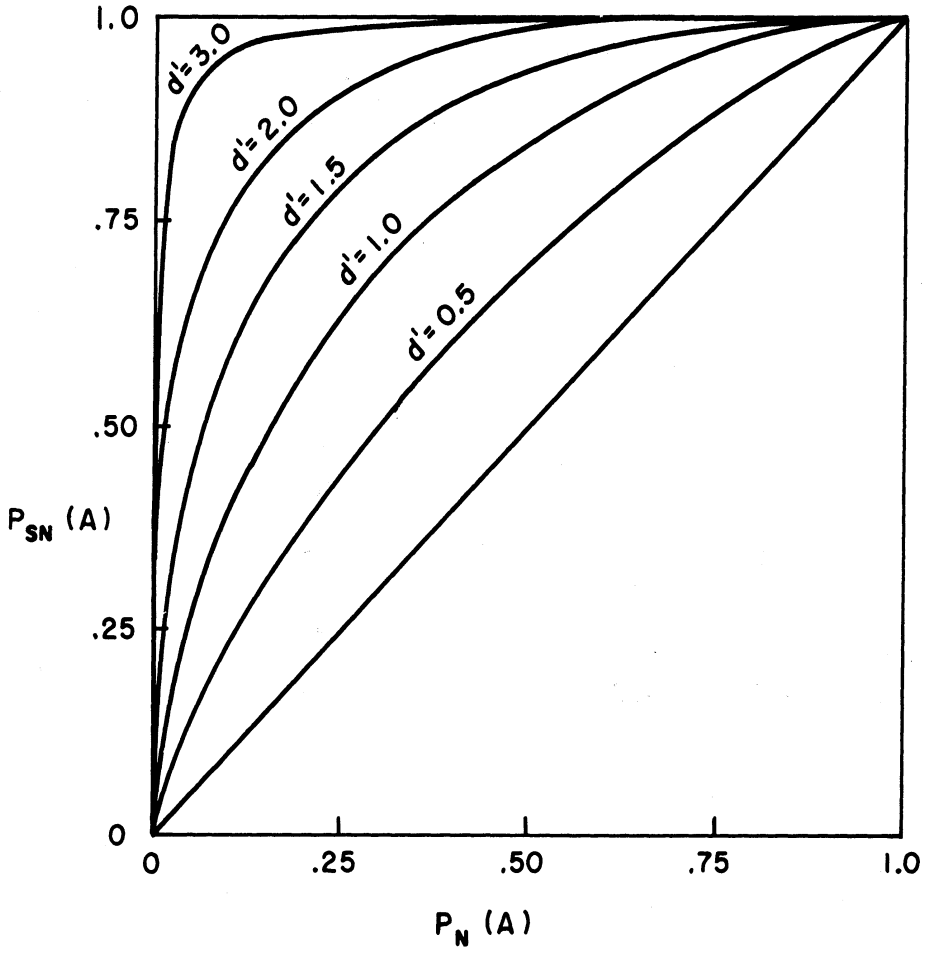

FIG. 3. A family of ROC-curves showing the relation between $P_{SN}(A)$ and $P_N(A)$ for various values of d'.

curves or receiver operating characteristics. Joint consideration of the values of $P_{SN}(A)$ and $P_N(A)$ enables one to recover the index of detectability, d', which does not vary as a function of the observer's criterion.

For the moment, let us consider three other possible measures of performance which are sometimes used as indices of the detectability of a signal. The first of these is simply $P_{SN}(A)$, that is, the observer's hit rate with no account taken of the false-alarm rate. The second is the observer's "hit rate" corrected by his "false-alarm rate" usually by the familiar correction for guessing,

$$P = \frac{P_{SN}(A) - P_N(A)}{1 - P_N(A)}$$

A third possible measure is the observer's proportion of correct responses considered over both the intervals containing a signal and those which contain noise alone. That is,

$$P(C) = P(SN \cdot A) + P(N \cdot CA).$$

The manner in which these three performance measures vary as a function of the observer's false-alarm rate is shown in Figure 4. For the curves shown in Figure 4,,the detectability of the signal is a constant, d' = 1, and the a priori probability of a signal is one-half.

V. SIGNAL RECOGNITION

The following development derives the distributions of likelihood ratio in the case where the observation interval always contains one of two signals plus noise. The task of the observer is to determine which of the two signals was present in the observation interval. The development of the previous section is a special case of the signal recognition problem, namely, that where one of the signals is s(t) = 0 over the

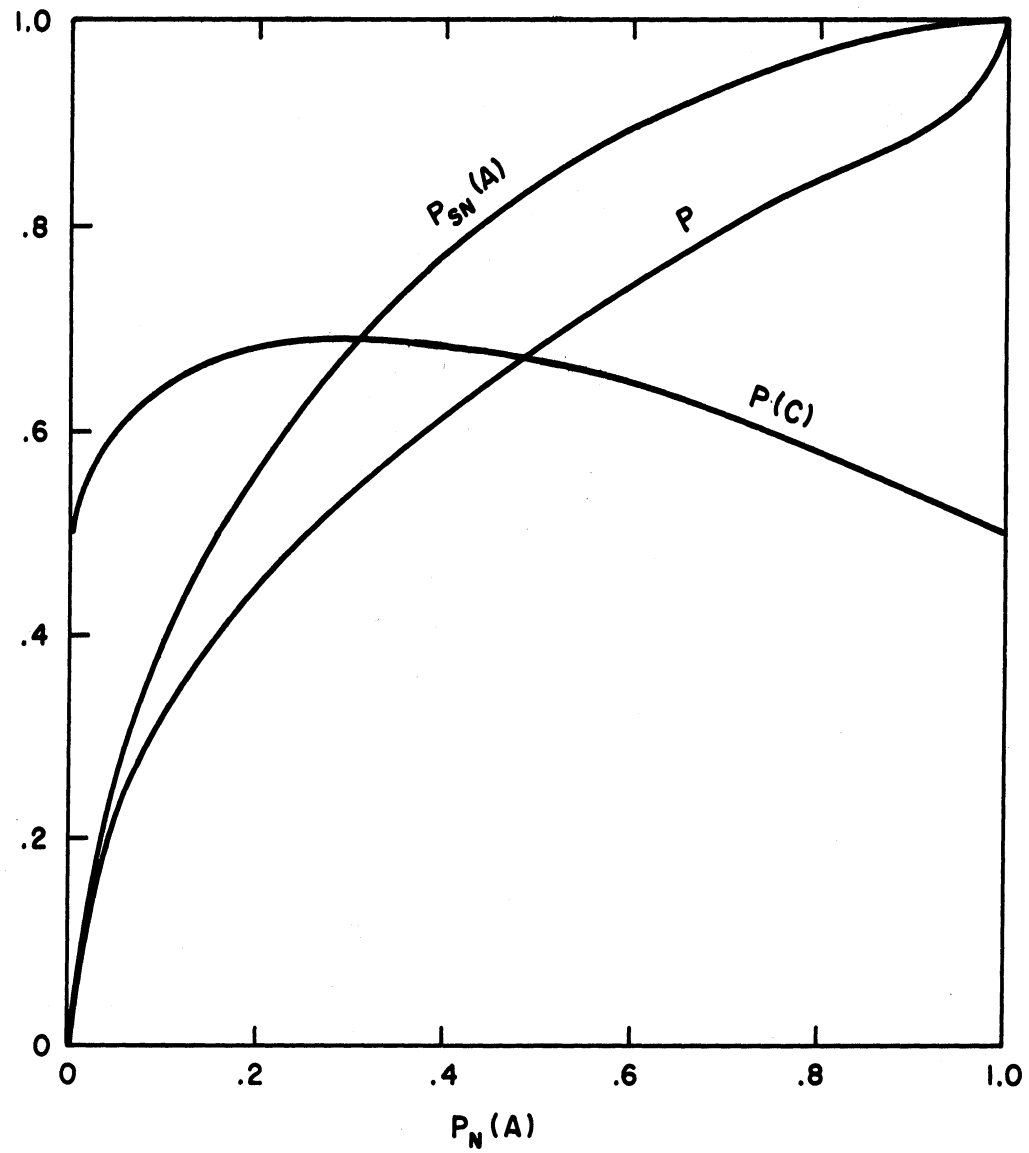

FIG. 4. The three performance measures described in the text are shown in a case where the detectability of the signal is constant, d' = 1. The common variable leading to the variation in these performance measures is the false-alarm rate of the observer. Consequently, these performance measures cannot be taken as indices of signal detectability.

entire observation interval. As this is a tutorial paper, this redundant presentation is favored in order that the reader may easily follow the mathematical development.

In this more general case of signal recognition, the observation interval may contain either $x(t) = s_1(t) + n(t)$ or $x(t) = s_2(t) + n(t)$. We wish to calculate likelihood ratio and determine its distribution under each of these two hypotheses. Here, as in the earlier development, we are assuming a Fourier series band-limited white Gaussian noise and Fourier series band-limited signals contained within the bandwidth of the noise.

We define

$$\ell(x) = \prod_{i=1}^{2WT} \frac{f_{S_1N}(x_i)}{f_{S_2N}(x_i)} = \prod_{i=1}^{2WT} \frac{\frac{1}{\sqrt{2\pi N}} e^{-\frac{1}{2N}(x_i - s_{1i})^2}}{\frac{1}{\sqrt{2\pi N}} e^{-\frac{1}{2N}(x_i - s_{2i})^2}} \tag{25}$$

and by algebraic manipulation, we obtain

$$\log \ell(x) = \frac{1}{2N} \sum_{i=1}^{2WT} s_{2i}^2 - \frac{1}{2N} \sum_{i=1}^{2WT} s_{1i}^2 + \frac{1}{N} \sum_{i=1}^{2WT} x_i(s_{1i} - s_{2i}) \tag{26}$$

Now we wish to see how $\log \ell(x)$ is distributed in the case where $x(t) = s_1(t) + n(t)$. Equation (26) then becomes

$$[\log \ell(x)] = \frac{1}{2N} \Sigma s_{2i}^2 + \frac{1}{2N} \Sigma s_{1i}^2 - \frac{1}{N} \Sigma s_{1i}s_{2i} + \frac{1}{N} \Sigma n_i(s_{1i} - s_{2i}). \tag{27}$$

The first three terms in this equation are constants and n_i is a normally distributed variable, consequently $[\log \ell(x)]_{S_1N}$ is itself a normally distributed variable. Appealing to the sampling theorem we find the mean and the variance of this distribution to be:

$$M_{S_1N}[\log \ell(x)] = \frac{E_1}{N_o} + \frac{E_2}{N_o} - \frac{1}{N} \sum_{i=1}^{2WT} s_{1i}s_{2i} \tag{28}$$

$$\mathrm{Var}_{S_1 N}[\log \ell(x)] = \frac{2E_1}{N_o} + \frac{2E_2}{N_o} - \frac{2}{N} \sum s_{1i} s_{2i}. \tag{29}$$

Similarly, we find

$$M_{S_2 N}[\log \ell(x)] = \frac{-E_1}{N_o} - \frac{E_2}{N_o} + \frac{1}{N} \sum_{i=1}^{2WT} s_{1i} s_{2i} \tag{30}$$

$$\mathrm{Var}_{S_2 N}[\log \ell(x)] = \frac{2E_1}{N_o} + \frac{2E_2}{N_o} - \frac{2}{N} \sum s_{1i} s_{2i} \tag{31}$$

Normalizing both distributions by dividing all values by the common standard deviation, we then have two normal distributions, each with variance one and a separation between the means of the two distributions equal to

$$d'_{12} = \left(\frac{2E_1}{N_o} + \frac{2E_2}{N_o} - \frac{2}{N} \sum_{}^{2WT} s_{1i} s_{2i} \right)^{1/2}. \tag{32}$$

By definition of the Pearson product moment correlation:

$$\rho_{12} = \frac{\Sigma s_1 s_2}{\sqrt{\Sigma s_1^2} \sqrt{\Sigma s_2^2}} ,$$

and by further application of the sampling theorem we may rewrite Eq. (32) as

$$d'_{12} = \left(\frac{2E_1}{N_o} + \frac{2E_2}{N_o} - 2 \rho_{12} \sqrt{\frac{2E_1}{N_o}} \sqrt{\frac{2E_2}{N_o}} \right)^{1/2} . \tag{33}$$

This is a general formula and may be applied in any situation where the observation interval contains one of two signals which are specified exactly. Two special cases are worth noting. The first of these is the case already examined in Section III. It is clear that if $E_2 = 0$, Eq. (33) reduces to Eq. (24) of Section III. Another case of special interest is that in which the two signals are perfectly correlated, that is, an amplitude discrimination situation. In this case, if $E_1 > E_2$, Eq. (33) reduces to

$$d'_{12} = \left(\frac{2E_1}{N_o} \right)^{1/2} - \left(\frac{2E_2}{N_o} \right)^{1/2} = d'_1 - d'_2 \tag{34}$$

VI. SIGNAL: A SAMPLE OF WHITE GAUSSIAN NOISE

A special case of particular interest is that where the signal is itself a sample of white Gaussian noise of noise power, S. Thus, the observation interval may be characterized as containing noise alone, that is, a sample of white Gaussian noise of noise power, N; or signal plus noise, i.e., a sample of white Gaussian noise of noise power, S+N. The task of the observer is to state whether the interval contained noise alone or signal plus noise. As in earlier developments, we shall again assume Fourier series band-limited white Gaussian noise. As before, the ideal observer should base his decision on likelihood ratio and we wish to determine how likelihood ratio, or some monotonic function of likelihood ratio, is distributed in the case of noise alone and in the case of signal plus noise. Following a development similar to that of Section III, we may write:

$$\ell(x) = \prod_{i=1}^{2WT} \frac{f_{SN}(x_i)}{f_N(x_i)} = \frac{\left[\frac{1}{2\pi(N+S)}\right]^{WT} e^{-\frac{1}{2(N+S)}\sum_{i=1}^{2WT}(x_i)^2}}{\left(\frac{1}{2\pi N}\right)^{WT} e^{-\frac{1}{2N}\sum_{i=1}^{2WT}(x_i)^2}} \tag{35}$$

which reduces to

$$\ell(x) = \left(\frac{N}{N+S}\right)^{WT} e^{\frac{S}{2N(N+S)}\sum_{i=1}^{2WT}(x_i)^2} \tag{36}$$

Taking logarithms of both sides and transposing a term, we have:

$$\log \ell(x) - WT \log\left(\frac{N}{N+S}\right) = \frac{S}{2N(N+S)} \sum_{i=1}^{2WT} x_i^2 \tag{37}$$

The expression on the right of the above equation is seen to be a simple monotonic transformation of likelihood ratio; thus, for describing the behavior of the ideal observer, it is only necessary to see how this

quantity is distributed for noise alone and for signal plus noise. If the interval contains noise alone, we may write

$$\frac{S}{2N(N+S)} \sum_{i=1}^{2WT} (x_i)^2 = \frac{S}{2(N+S)} \sum_{i=1}^{2WT} \left(\frac{x_i}{\sqrt{N}}\right)^2 \tag{38}$$

The quantity, $\frac{x_i}{\sqrt{N}}$, is, in this case, normally distributed with mean zero and variance one; and thus $\sum_{i=1}^{2WT}\left(\frac{x_i}{\sqrt{N}}\right)^2$ is distributed as chi square with 2WT degrees of freedom. If 2WT is large, we have a distribution which is approximately normal:

$$\frac{S}{2(N+S)} \Sigma\left(\frac{x_i}{\sqrt{N}}\right)^2 \approx N\left(\frac{WTS}{N+S}, \frac{WTS^2}{(N+S)^2}\right). \tag{39}$$

If signal plus noise is present in the observation interval, we may write

$$\frac{S}{2N(N+S)} \sum_{i=1}^{2WT} (x_i)^2 = \frac{S}{2N} \sum_{i=1}^{2WT} \left(\frac{x_i}{\sqrt{S+N}}\right)^2 . \tag{40}$$

In this case, we then have for large 2WT

$$\frac{S}{2N} \sum_{i=1}^{2WT} \left(\frac{x_i}{\sqrt{N+S}}\right)^2 \approx N\left(\frac{WTS}{N}, \frac{WTS^2}{N^2}\right) \tag{41}$$

Normalizing both distributions by dividing all values by the standard deviation of the noise, substracting the mean of the noise distribution, and letting $S = \lambda N$, we see that the two hypotheses are distributed as follows:

$$\text{Noise alone} \approx N(0,1) \tag{42}$$

$$\text{Signal plus noise} \approx N[\lambda\sqrt{WT},\ (1+\lambda)^2] \tag{43}$$

where W is the bandwidth common to both the signal and the noise and T is the duration of the observation interval.

Thus we see that in this situation, and it is true in all cases where the signal is not specified exactly, the variance of the signal plus noise distribution is greater than the variance of the noise alone distribution.

VII. SUMMARY

We have seen that simple detection experiments may be formalized in terms of testing statistical hypotheses. Under the assumption of Fourier series band-limited white Gaussian noise, we have determined the distribution of the hypotheses signal plus noise and noise alone for a number of special cases. We have seen that the behavior of an ideal observer is a function of various criterial variables. The entire development of this chapter has been with reference to an ideal observer. The manner in which this concept may be utilized in the study of human behavior will be discussed in other chapters.

A Survey of the Theory of Probability

R. K. Ritt, The University of Michigan

The theory of probability has its roots in the study of games of chance. In this primitive setting, it starts off with an attempt to define the "probability" of an event. An event is one of perhaps several outcomes of an experiment. For example, the event 'tails' is a possible outcome of tossing a coin; the event 'two aces' is a possible outcome of drawing two cards from a deck. In these terms the probability of an event might be defined in the following fashion:

Let the experiment be repeated N times. Let S(N) be the number of times, among the N repetitions that the event occurs. Then the probability of the event is defined to be the limit, as $N \to \infty$, of $S(N)/N$. Before discussing the value of such a definition, let us examine some of its consequences. If two events are such that they cannot occur simultaneously, and if $S_1(N)$ and $S_2(N)$ are the number of times they, respectively, occur in N repetitions, then the composite event "at least one of the two events," will occur $S_1(N) + S_2(N)$ times; so that the probability of this composite event is the sum of the probabilities of the two events, since

$$\lim_{N\to\infty} \frac{S_1(N) + S_2(N)}{N} = \lim_{N\to\infty} \frac{S_1(N)}{N} + \lim_{N\to\infty} \frac{S_2(N)}{N} .$$

Further, if an event occurs S(N) times in N repetitions, then the event, "the first event does not occur," itself occurs N - S(N) times, so that the probability of what we shall call the <u>negation</u> of an event is 1 minus the probability of the event itself.

Now let us imagine a more complicated experiment. It shall consist of performing the two separate experiments in succession. Let two events

be considered, and let a new event be described as the successive occurrence of these two events. Let this new experiment be performed N^2 times. If N is large, $\frac{S_1(N^2)}{N^2}$ is approximately the same as $\frac{S_1(N)}{N}$, so the first event will occur approximately $\left[\frac{S_1(N)}{N}\right] N^2 = S_1(N)/N$ times. Regarding now just these experiments, $\frac{S_2\, S_1(N)N}{S_1(N)N}$ is approximately $\frac{S_2(N)}{N}$, so that the second event will occur approximately $\left[\frac{S_2(N)}{N}\right] S_1(N)/N = S_1(N)S_2(N)$ times. Therefore, the number of occurrences of the new event will be approximately $S_1(N)S_2(N)$ times, and its probability will be

$$\lim_{N\to\infty} \frac{S_1(N)S_2(N)}{N^2} = \left[\lim \frac{S_1(N)}{N}\right]\left[\lim \frac{S_2(N)}{N}\right];$$

so that the probability of the new event is the product of the probabilities of the original events.

These heuristic observations lead us to a tentative calculus of probabilities. If the possible outcome of an experiment can be described as being among a set of outcomes, $E_1 \ldots E_2 \ldots$, no two of which can occur simultaneously, the probability of each of these outcomes, $P(E_k)$, must be a number between zero and one. $P(E_1) + P(E_2) + \ldots = 1$. If $E_1 + E_2$ is regarded as the combined outcomes, "either E_1 or E_2", then $P(E_1 + E_2) = P(E_1) + P(E_2)$. If $\bar{E}_k$ is the event, "not E_k", $P(\bar{E}_k) = 1 - P(E_k)$. Finally, if two experiments are performed successively, and E is the outcome of the first and F an outcome of the second, then the probability of the events E and F occurring successively is P(E) P(F).

This is the beginning of the elementary theory of probability, with which most scientific investigators are familiar, at least to the extent of being able to perform calculations, and fearlessly to assign operational meaning to the numbers they obtain.

However, if we examine, without prejudice, the above presentation, we find it devoid of either mathematical or scientific meaning. First of all, the definition of probability given above is not a mathematical one. True, it is phrased in the language of algebra, with symbols and division and limits; but saying something with mathematical symbols is not necessarily mathematics. All that we get out of the above "argument" is the concept of associating, in some vague way, a number with the possible outcome of an experiment. I say "vague" advisedly, since an operational definition which involves performing an unlimited number of experiments is, if not worthless, at least difficult to fit into any epistemological system.

However, it is historically true that the calculus of probabilities introduced above has had a remarkable success in the physical sciences, and, I am told, in card playing, tossing dice, and other of the behavioral and social sciences. What do I mean by success? Simply that by using this calculus, predictive theories have been developed, and these theories have been tested by experiment, and have been found to be fairly accurate descriptions of parts of the phenomenological universe. Thus, in dealing with a large number of electrons, the physicist can predict a defraction pattern; in dealing with a large number of people, a geneticist can predict distributions of hereditary characteristics; a gambler who spends his working day at the card table can adopt rules of decision which guarantee him, if he plays long enough, an optimal gain.

Thus, we cannot discard the theory of probability as outlined above; despite its philosophical shortcomings, it must be preserved. However, if we carefully examine these shortcomings, we may be able to reconstruct the theory on a basis which will leave our scientific consciences untroubled.

Let us now attempt to outline the task before us. In order to do this we must first understand the role played by mathematics in scientific theories. Without belaboring the question "what is a mathematical theory?", let us assume that the scientist, through some undescribed process, has come into possession of a mathematical theory. He understands the theorems, he can perform calculations; what does he do next? Presumably he is interested in some class of phenomena and he wishes to make predictions. In some fashion peculiar to his own genius, he must establish a relationship between the things he observes and the mathematics he wants to use. Then he employs the mathematical theory to make predictions about the things he is going to observe. The measure of his success is irrelevant to the inner consistency, or "correctness," of the mathematics - that is taken for granted. It is also irrelevant to his formulation of the problem in terms of mathematics. There is no "logical" or "correct" way to perform such a formulation; it is entirely a personal matter, dependent only upon the scientist's insight, which may differ considerably from the insight of every other worker in the field. The measure of his success is based entirely on the verification of his predictions. As long as they are correct, his theory is correct.

It is true that many mathematical theories, the theory of probability in particular, have been born and have grown with particular applications in mind. But it is also true, that as the theories were applied to more complicated phenomena, paradoxes and inconsistencies developed, and the mathematical theory had to be reexamined and put on a rational basis before any more progress could be made. A striking example of this has occurred in recent years, in mathematical physics, where the mathematical formalism of quantum electrodynamics led to meaningless

answers; it is only now that these difficulties, which arose twenty years ago, are beginning to become resolved, as a result of careful examination and reorganization of the mathematical theories. Returning to the topic at hand, the need for a sophisticated use of the theory of probability in its application requires that this theory be given an exposition as a mathematical theory, rigorous, and bare of phenomenological interpretation.

Thus, I regard my task to be the brief exposition of this theory.

Set Theory

Every mathematical theory must contain undefined terms. We shall talk of "points,"without saying what they are, and collections of points, which we shall call "sets." If A and B are two sets, we say that $A < B$ if every point of A is also a point of B. The set $A + B$ consists of all points which are in either A or B. The set AB consists of all points which are in both A and B. By the null set, φ, we mean the set which contains no points. For every set A, $\varphi < A$. In fact, assume $\varphi \not< A$ (this should be read "it is not true that $\varphi < A$"); then there must be a point, a, which is in φ but not in A. But φ contains no points.

All of our sets will be assumed to be subsets (contained in; $<$) of a universal set, Ω. If A is a set, by $\bar{A}$ we mean the points in Ω which are not in A. $A\bar{A} = \varphi$. In general, if $AB = \varphi$, then A and B are said to be disjoint.

If $A_1, A_2, \ldots$, are a finite or infinite sequence of sets, by $A_1 + A_2 + \ldots$, we mean all points which are in at least one of the A_n; by $A_1A_2 \ldots$, we mean all points which are in all the A_n. It is easy to verify that $\overline{(A_1 + A_2 + \ldots)} = \bar{A}_1\bar{A}_2 \ldots$; and $\overline{(A_1A_2\ldots)} = \bar{A}_1 + \bar{A}_2 + \ldots,$.

Borel Fields

Let $\underline{B}$ be a collection of subsets of Ω. Let $\underline{B}$ satisfy the following conditions:

(1) Ω and φ are in $\underline{B}$.

(2) If $A_1, A_2, \ldots$, is a finite or infinite sequence of sets in $\underline{B}$, then $A_1 + A_2 + \ldots$ is in $\underline{B}$, $A_1A_2 \ldots$ is in $\underline{B}$.

(3) If A is in $\underline{B}$, $\overline{A}$ is in $\underline{B}$.

These conditions are, in fact, redundant; there are mathematicians who enjoy reducing collections of axioms to a smallest possible collection. Since the collection of all subsets of Ω satisfies (1), (2), and (3), these conditions are consistent. $\underline{B}$ is said to be a Borel field.

Probability Measure

Let $\underline{B}$ be a Borel field. Let P be a function which associates with each set, A, in $\underline{B}$, a real number P(A), subject to the following conditions:

(1) $0 \leq P(A) \leq 1$

(2) If $A_1, A_2, \ldots$, form a finite or infinite sequence in $\underline{B}$, which are mutually disjoint ($A_kA_j = \varphi$ if $k = j$), then $P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \ldots$.

(3) $P(\Omega) = 1$.

Then P is called a probability measure.

It is immediately possible to derive from these properties, further properties of P:

(1) $P(A) + P(\overline{A}) = 1$

Proof: $P(A) + P(\overline{A}) = P(A + \overline{A}) = P(\Omega) = 1$.

(2) $P(\varphi) = 0$

Proof: $P(A) + P(\varphi) = P(A + \varphi) = P(A)$.

(3) $P(A + B) = P(A) + P(B) - P(AB)$

Proof: $A = A\bar{B} + AB$ $\qquad P(A) = P(A\bar{B}) + P(AB)$

$B = AB + \bar{A}B$ $\qquad P(B) = P(AB) + P(\bar{A}B)$

$$P(A) + P(B) = 2P(AB) + P(A\bar{B}) + P(\bar{A}B)$$

$$P(A) + P(B) - P(AB) = P(AB) + P(A\bar{B}) + P(\bar{A}B)$$

$$= P(AB + A\bar{B} + \bar{A}B)$$

$$= P(A + \bar{A}B)$$

$$= P(A + B)$$

Our purpose in proving these theorems is to illustrate the fashion in which an elaborate theory can be developed from some fairly simple assumptions.

It is time now to catch our breath, and discuss the question of applying the above material. It should first be observed that the formulas we have just written down resemble the calculus of probability we discussed in the beginning of the section, provided the word "event" is substituted for "set." Let us take a concrete example. Suppose that we wish to give a probabilitistic discussion of the experiment of tossing a coin. There are two outcomes of the experiment: "heads" and "tails." (I am sure that someone is thinking 'what about standing on edge?'. In reply, I apologize for over-simplifying the problem - we have also neglected the case in which the coin fails to come down, so that we must regard this as only an approximate theory.) According to our primitive scheme we wish to assign a probability to these two outcomes. If we follow our primitive scheme through, we must then toss the coin a large number of times, and acting on some sort of faith that the ratios involved in the definition have a limit which is approximated in a reasonably small number of tosses, we arrive at two probabilities, P(H), P(T), which presumably we shall now

make use of in some future work. Probabilities arrived at in this fashion are called a posteriori probabilities.

Now let us proceed differently. Let the concrete representation of the set, Ω, be the "points," H, T. The Borel field will consist of four sets: $\{H\}$, $\{T\}$, $\{H + T\}$, and φ. The set, $\{H + T\}$, is to be interpreted as "heads or tails," and in Ω itself. Now let P(H) and P(T) be any two non-negative numbers such that $P(H) + P(T) = 1$. Let $P(\varphi) = 0$ and $P(\Omega) = 1$. This function has now been defined for all sets in B and fulfills the demands of a probability measure. We then build up a theory of the experiment by using this probability measure. One of the predictions of this theory would be that in a sufficiently large number of repetitions of the experiment, the ratio of the number of heads to the number of tosses would be approximately P(H). It may seem that we have gained nothing, but I point out that this prediction is the result of the mathematical "law of large numbers," and is arrived at within the mathematical system; for different choices of the numbers, P(H) and P(T), we obtain different theories and different predictions, any one of which can be tested by experimentation. Because of the simplicity of this experiment, the power of this reorientation toward the subject may not be apparent.

Let us go over to a more complicated situation; the case of quantum statistical mechanics. Let us think of a gas, counting of N identical particles. Each of these molecules will be described by a set of k numbers, which specify its position, orientation, linear and angular momentum, and any other dynamic properties we choose to consider. Thus, it requires kN numbers to completely specify the state of the system at any given time. The fact that we are dealing with a quantized theory

requires that we break up the kN dimensional space into a discrete set of "boxes," each of the same "volume." This set of boxes are the "points" of Ω, and we must now assign a value of P to each of them, which obeys the conditions of being a probability measure. Now admittedly, this has been a hasty description, but it is adequate to bring out the main point. According to the choice of P, a unique theory will emerge. It is known that for different types of particles, different choices of P must be made to secure a successful theory. The "correct" choice of P is determined by the insight of the investigator. As a matter of record, there are essentially two different choices known which are valuable: the Fermi-Dirac choice and the Einstein-Bose choice. How these choices were made we can only guess, but they have both produced fruitful theories. What is more important, the nature of the phenomena is such that no experiment could be designed to compute the a posteriori probabilities. The procedure of guessing the correct probabilities, making predictions, and testing is the only feasible one.

Relative Probability

We now turn to a question which, I am told, has led to many disagreements among scientific practitioners. It is the question of Baye's theorem. This theorem is often cited in the study of the validity of hypotheses, and it seems that the disagreement must arise from a lack of understanding as to what the theorem says.

Let B be a Borel field subsets of Ω, and let P be a probability measure on B. Let H be in B, for which $P(H) \neq 0$. If A is in B, let us define

$$P_H(A) = \frac{P(AH)}{P(H)}$$

$P_H(A)$ is called: the probability of A, relative to H. It is easy to verify that $P_H(A)$ is a new probability measure for B. It has the interesting property that if $H < A$, then $P_H(A) = 1$.

Now let H and A be two sets in B, where $P(H)$ and $P(A) \neq 0$. Now by definition

$$P_A(H) = \frac{P(AH)}{P(A)} \qquad \text{and } P_H(A) = \frac{P(AH)}{P(H)} ;$$

then

$$P_A(H)\, P(A) = P_H(A)\, P(H), \text{ or}$$
$$P_A(H) = \frac{P_H(A)\, P(H)}{P(A)} .$$

This formula is called Baye's rule. It is of great practical importance in evaluating the "probability that H has occurred if A occurs." All confusions resulting from the application of this rule stem from the fact that in it the word "probability" occurs with three different meansings. For P_A, P_H, and P are distinct probability measures. But if this is understood, there is no further bar to an intelligent application of Baye's rule.

Independent Events

Let A and H be two subsets of B. Intuitively, we think of A and H as being independent if $P_H(A) = P(A)$. In words, the probability of A does not change with the knowledge of the occurence of H. But, referring to the first formula in the preceding section, this leads to the definition:

A and H are independent if $P(AH) = P(A)\, P(H)$.

This is a rule of the probability calculus which we "derived" in our introduction. In our revised theory, we use a formula to define the concept of independence. Again, I observe that the the usefulness of this definition depends on the success of its application.

Repeated Experiments

Let Ω be a set and $\underline{B}$ a Borel field of subsets for which we have a probability measure, P. Let Ω', $\underline{B}'$ and P' be another such triple. By $\Omega \times \Omega'$, we mean a set consisting of all pairs (E, E'), where E and E' are respectively in Ω and Ω'. There are certain subsets of $\Omega \times \Omega'$ of the form $A \times A'$, where A is in $\underline{B}$, and A' is in $\underline{B}'$. These subsets do not necessarily form a Borel field. Nevertheless, we define a function, $P^*(A \times A') = P(A)\, P(A')$. It is a known theorem that if $\underline{B}^*$ is the smallest Borel field in $\Omega \times \Omega'$ which contains all sets of the form $A \times A'$, (the intersection of all Borel fields with this property), then there exists a unique probability measure defined on $\underline{B}^*$, call it $\overline{P}^*$, such that $\overline{P}^*(A \times A') = P^*(A \times A')$. We say that P^* has been extended to $\overline{P}^*$. It should be observed that if the set $A \times \Omega'$ in $\underline{B}^*$ is identified with the set A in $\underline{B}$, and the set $\Omega \times A'$ is identified with A' in $\underline{B}'$, the set $(A \times \Omega') \cdot (\Omega \times A') = (A \times A')$ has the probability measure $P(A)P'(A')$, which is consistent with the assertion that A and A' are independent. This is rather heady material and I shall give an example:

Let Ω be the set we have previously considered, containing the points H and T. The Borel field is the sets, $\{H\}$, $\{T\}$, Ω, φ; let $P(H) = p$, $P(T) = q$, where $p + q = 1$. Let Ω', $\underline{B}'$, and P' be the same as Ω, $\underline{B}$, and P. We can compute the set $\Omega \times \Omega'$ to be the points: (H,H), (H,T), (T,H), (T,T). The Borel field, $\underline{B}^*$, is the collection of all subsets formed, using these four points, together with $\Omega \times \Omega'$ itself and φ. The measure, P^*, has the respective values p^2, pq, qp, q^2 on these four points, and to verify that $\overline{P}^*$ is a probability measure, it is sufficient to observer that $p^2 + pq + qp + q^2 = (p + q)^2 = 1$.

If we care to, we can interpret this in the following fashion: Let a coin be tossed with p the probability of it turning up heads, $[P(H) = p]$, and q the probability of it turning up tails, $[P(T) = q]$. If it is tossed twice, the probability of two heads is $p^2[P^*(H,H) = p^2]$; the probability of two tails is $q^2[P^*(T,T) = q^2]$; and the probability of a head and a tail is $2pq\left[\bar{P}^*[(H,T) + (T,H)] = \bar{P}^*(H,T) + \bar{P}^*(T,H) = P^*(H,T) + P^*(T,H) = 2pq\right]$.

We see that this gives us a machinery for considering the probabilities associated with the repetition of an experiment. Although we have only sketched the ideas involved, it is clear that we are expanding our theory so that it can have useful application.

Bernoulli Trials

The ideas in the preceding section can be generalized to the case of n distinct sets, $\Omega^1, \ldots, \Omega^n$, each with its onw probability measure. We shall restrict this generalization to the case in which the Ω^j are identical, each being the set considered above, whose points are H and T, and we shall assume that the probability measure in all of the Ω^j is the same; namely, $P(H) = p$, $P(T) = q$, where $p + q = 1$. This situation has a name. It is called n Bernoulli trials. If a concrete example is desired, it can be thought of as a description of n successive tosses of a coin, or, more generally, n repetitions of an experiment whose outcomes can be described as one of two possibilities. The points of $\Omega \times \Omega \times \Omega \ldots \Omega$, can be listed as all n-tuples of the form $(E_1, E_2, \ldots E_n)$, where each E_j can be either H or T, and $P^*(E_1, E_2, \ldots, E_n) = P(E_1)\, P(E_2) \ldots P(E_n)$. $\bar{P}^*$ is then defined by requiring that it be additive (property 2). Suppose we let S_k^n represent the event "k H's." The points of S_k^n are those n-tuples, $(E_1, E_2, \ldots, E_n)$ in which exactly k of the E_j are H. The $P^*(E_1, \ldots, E_n)$, for such points, are all $p^k\, q^{n-k}$. $\bar{P}^*(S_k^n)$ is equal to the sum of the

$P^*(E_1, \ldots, E_n)$, for all $(E_1, \ldots, E_n)$ in S_k^n. How many of these are there? This is a familiar problem studied in elementary algebra courses. It can be answered by the following argument. We have k H's at our disposal. The first of these can be placed in any one of n positions. When this is done, the second can be placed in any one of n - 1 positions; continuing this procedure, we find that the H's can be distributed among the E's in exactly $n(n-1)\ldots(n-k+1)$ different ways. How many of these distributions give the same $(E_1, \ldots, E_n)$? Clearly, this is the number of ways in which the k H's can be arranged among themselves. By the same argument, this number is $k(k-1), \ldots, (2)(1)$. Thus, the number of $(E_1, \ldots, E_n)$ which have exactly k H's among the E's is

$$\frac{n(n-1),\ldots,(n-k+1)}{k(k-1),\ldots,(2)(1)} \quad .$$

These numbers are the familiar binomial coefficients, and are denoted by the symbol C_k^n. If 0! is defined by the relation 0! = 1, and (n+1)! is defined as (n+1)! = (n+1)(n!), for n = 0, 1, 2, ..., then we may write

$$C_k^n = \frac{n!}{(n-k)!k!} \quad \text{for } k = 0, 1, 2, \ldots, n.$$

It will be convenient to also let

$$C_k^n = 0 \text{ if } k < 0 \text{ or } k > n.$$

Thus, we have answered our question: $\overline{P}^*(S_k^n) = C_k^n p^k q^{n-k}$. This formula is alternatively called the Binomial Distribution or the Bernoulli Distribution.

Now let us consider in what way this theory coincides with the "primitive" theory outlined in the introduction. Suppose that n = 8, and p = 1/2. Again, to be concrete, imagine we toss a coin 8 times, and the probability of heads is 1/2. Then in 8 tosses, our primitive theory predicts that there will occur 4 heads. Actually, what is $\overline{P}^*(S_4^8)$? It is a

simple calculation to show that

$$\bar{P}^*(S_4^8) = \frac{C_4^8}{2^8} = \frac{70}{2^8} = \frac{70}{256}$$

which is a number much less than 1. Hence, we are led to the statement that tossing 4 heads in 8 trials is an unlikely event, even though the probability of a head is 1/2. However, if we compute

$$\bar{P}^*(S_3^8 + S_4^8 + S_5^8),$$

we find it is $\frac{182}{256}$, which is much closer to 1. So we are led to the statement that tossing between 3 and 5 heads in 8 tosses is a likely event.

These are rough ideas, but they can be made precise in the following fashion. Let us consider n Bernoulli trials, with probability p for H. Let S_n be the number of heads which actually occur. The primitive theory predicts that $\lim_{n\to\infty} \frac{S_n}{n} = p$. Does our present theory make a similar prediction? What we require is a measure of the difference between $\frac{S_n}{n}$ and p. If $\epsilon > 0$ is chosen arbitrarily, we should like to know that $p - \epsilon < \frac{S_n}{n} < p + \epsilon$ for sufficiently large n. But, of course, our present theory cannot tell us this. It makes no prediction of S_n, it simply assigns probability measures to events. But $p - \epsilon < \frac{S_n}{n} < p + \epsilon$ *is* an event. In fact it is the sum of the events S_k^n where k ranges over the integers between $(p - \epsilon)n$ and $(p + \epsilon)n$. Thus, $\bar{P}^*(p - \epsilon < \frac{S_n}{n} < p + \epsilon)$ can be computed as the sum of the corresponding $\bar{P}^*(S_k^n)$. It is then possible to prove, since these last numbers are known explicitly, that

$$\lim_{n\to\infty} P^*(p - \epsilon < \frac{S_n}{n} < p + \epsilon) = 1.$$

This result is one of the forms of a *law of large numbers*; by means of it, we have come full circle. The primitive calculus of probability is preserved; but, philosophically, we have accomplished much more. We have a rigorous mathematical discipline, which, when applied to phenomena, has

in many instances provided a predictive theory of those phenomena.

Distributions

In the preceding section we spoke of the numbers S_k^n as a binomial distribution. As a matter of fact, if we think of a new space, whose points are the integers and if $P(k) = S_k^n$ and if $P(k)$ is extended to a $\overline{P}$ on the Borel field of all subsets of the integers, $\overline{P}$ is a probability measure. To verify this it is only necessary to observe that

$$\overline{P}(\Omega) = \overline{P}(0) + \overline{P}(1) + \ldots + \overline{P}(n) = \overline{P}(1) + \ldots + \overline{P}(n)$$
$$= C_0^n p^n + C_1^n p^{n-1} q + \ldots + C_n^n q^n = (p + q)^n = 1.$$

In general, if we let Ω be the integers, and $\underline{B}$ the Borel field of subsets of Ω, then a probability measure, P, on $\underline{B}$, is called a discrete, one-dimensional distribution. If $f(k)$ is any function defined on the integers, by the expected value of $f(k)$ is meant the sum

$$< f(k) > = \sum_{k=-\infty}^{\infty} f(k)P(k),$$

when this sum exists. Of particular interest are the expected values

$$< k^t >, \; t = 1, 2, 3, \ldots,$$

which are called the t^{th} moments of the distribution. The first moment, $< k >$, is sometimes called the center of mass or mean value of the distribution, and the second moment $< k^2 >$, is sometimes called the moment of inertia of the distribution, in obvious analogy with the related mechanical problem of a system of rigidly connected masses, $P(k)$, located at the points, k. An interesting and useful expectation is

$$\sigma^2 = < (k - < k >)^2 >.$$

σ is called the standard deviation, or $\sqrt{\text{variance}}$, of the distribution and is useful as a measure of the degree of concentration of the distribution about its mean. Observe that $\sigma^2 = \sum_{k=-\infty}^{\infty} (k^2 - 2k < k > + < k >^2) P(k) = < k^2 > - 2 < k >^2 + < k >^2 = < k^2 > - < k >^2$, which sometimes provides a

useful way of computing σ. Let us compute, for example, $< k >$, $< k^2 >$, and σ for the binomial distribution $P(k) = C_k^n\, p^k\, q^{n-k}$.

$$< k > = \sum_{k=0}^{n} k\, C_k^n\, p^k\, q^{n-k} = \sum_{k=0}^{n} \frac{n!}{(n-k)!(k-1)!}\, p^k\, q^{n-k}$$

$$= \sum_{t=0}^{n-1} \frac{n!}{(n-1-t)!t!}\, p^{t+1}\, q^{(n-1)-t} = np \sum_{t=0}^{n-1} C_t^{n-1}\, p^t\, q^{(n-1)-t}$$

$$= np$$

This provides another link between the present theory and the primitive one.

$$< k^2 > = \sum_{k=0}^{n} \frac{k^2 n\, p^k\, q^{n-k}}{(n-k)!k!}$$

$$= np \sum_{t=0}^{n-1} (t + 1)\, C_t^{n-1}\, p^t\, q^{(n-1)-t}$$

$$= (np)\,[(n-1)p + 1] = np\,(np + q)$$

$$\sigma^2 = n^2p^2 + npq - n^2p^2 = npq, \text{ or } \sigma = \sqrt{npq}.$$

Thus, we see, in a rough way, that the distribution, which extends from zero to n, tends to be concentrated in a region about np of length $\sqrt{npq}$, which is $\sqrt{\frac{pq}{n}}$ of the total length of the region on which the distribution is not zero. This is another law of large numbers, loosely stated.

If $P(k)$ and $P'(k)$ are two discrete one-dimensional distributions, the <u>joint distribution</u> $(P * P')(k)$, is defined as

$$(P * P')(k) = \sum_{n=-\infty}^{\infty} P(k-n)P'(n).$$

It is easy to verify that $(P * P')(k) = (P' * P)(k)$, and that $\sum_{k=-\infty}^{\infty}(P * P')(k) = 1$. This last is needed to be sure that $(P * P')(k)$ is a distribution.

Let $< k >$, $< k >_1$, $< k >_2$, be the means of $P \times P'$, P, and P' respectively.

$$< k > = \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} k\, P(n)\, P'(k-n)$$

$$= \sum_{n=-\infty}^{\infty} P(n) \sum_{n=-\infty}^{\infty} kP'(k-n)$$

$$= \sum_{n=-\infty}^{\infty} P(n) \sum_{k=-\infty}^{\infty} (k+n)P'(k)$$

$$= \langle k \rangle_1 + \langle k \rangle_2$$

Using similar notation,

$$\langle k^2 \rangle = \sum_{n=-\infty}^{\infty} P(n) \sum_{k=-\infty}^{\infty} (k+n)^2 P'(k) = \langle k^2 \rangle_2 + 2 \langle k \rangle_2 \langle k \rangle_1 + \langle k^2 \rangle_1$$

Thus, $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2 = [\langle k^2 \rangle_2 - \langle k^2 \rangle^2] + \langle k^2 \rangle_1 - \langle k \rangle^2_1] = \sigma_1^2 + \sigma_2^2$

If $P(k) = C_k^m p^k q^{m+k}$,

$P'(k) = C_k^1 p^k q^{1-k}$,

$$(P * P')(k) = \sum_{n=0}^{1} C_{k-n}^m C_n^{(1)} p^k q^{m+1-k}$$

$$= [C_k^m C_0^{(1)} + C_{k-1}^m C_{(1)}^1] p^k q^{m+1-k}$$

$$= \frac{m!}{(m-k)!k!} + \frac{m!}{(m+1-k)!(k-1)!} p^k q^{m+1-k}$$

$$= \frac{(m+1)}{(m+1-k)!k!} p^k q^{m+1-k}$$

$$= C_k^{m+1} p^k q^{m+1-k}$$

which is the binomial distribution associated with m+1 trials. This is just the proof of the familiar "Pascal triangle" method of finding binomial coefficients. In general the joint distribution of the distributions for m and m' Bernoulli trials is the distribution for m + m' Bernoulli trials. Because the means and the square of the variance add when joint distributions are taken, this general method is quite useful when it is generalized to the study of the large number of successive performances of an experiment whose outcomes can be described by a distribution.

Continuous Distributions

Let $D(x)$ be a function defined on the real axis, $-\infty < x < \infty$. For simplicity, assume $D(x)$ is a non-negative continuous function. Let $\int_{-\infty}^{\infty} D(x)dx = 1$. If $P[a,b) = \int_a^b D(x)dx$, where $[a,b)$ stands for the interval, $a \leq x < b$, $P[a,b)$ can be extended to a probability measure on the smallest Borel field containing such intervals. $P(x)$ is called a continuous distribution function. Expected values, moments, joint distributions can be defined in a similar fashion to that in the case of discrete distributions, where integration is used in place of summation.

The Normal Distribution

The classical normal distribution is given by the formula:

$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

The mean of the normal distribution is a, and its variance is σ^2. It is the well-known "bell-shaped curve." The normal distribution plays a central role in the application of the theory of probability to the repetition of identical experiments. For, let $D(x)$ be a distribution function, subject to certain hypotheses regarding the existence of its moments. Let $D(x)$ have a mean a and variance σ^2. Then, if $D^n(x)$ is defined by $D^n(x) = (D^{n-1} * D)(x)$, $n = 2, 3, \ldots$, it can be shown that $D^n(x)$ can be approximated, in a sense we shall not describe precisely, by a normal distribution with the mean na, and with variance $n\sigma^2$. These facts, when stated precisely, are another form of the law of large numbers.

There are many topics, important in the theory, which I have omitted. There are many texts on the theory of probability, such as that of Feller, and on mathematical statistics, and to preempt them would be presumptuous. I have attempted to give an outline of what seems to me to be the highlights of the mathematical theory, and the emphasis I have given to the

various laws of large numbers has been motivated by a hope that I have made clear that, by means of these laws, the philosophic link between the axiomatic theory and what I have termed the primitive theory can be established. And what is science without a philosophy of science?

A Brief Introduction to Decision Theory

John A. Swets, Massachusetts Institute of Technology

This chapter provides a brief introduction to the theory of statistical decision. A simple game of chance is described to illustrate the main ideas of the theory as applied to signal detection by human observers. A formal account of decision theory has been presented by Wald (Ref. 1). Less technical expositions of the theory have been advanced by Braithwaite (Ref. 2) and Bross (Ref. 3). This introduction borrows to some extent from each of these sources.

The theory of statistical decision is founded upon the more special theory of testing statistical hypotheses. A statistical hypothesis is to be contrasted with a universal hypothesis, i.e., a hypothesis of the form: "all A's are B's," or "all swans are white." In a common form of statistical hypothesis, the assertion is made that "some (specified) proportion of A's are B's," for example, "51% of swans are white."

At one time statistical hypotheses could be regarded as necessary evils in the less advanced sciences until they could be replaced by universal hypotheses. Presently, statistical hypotheses are of greater interest, since the most sophisticated theories of the most advanced sciences have postulated an irreducibly statistical form of explanation. Recall that whereas Pascal introduced the theory of probability in the seventeenth century, it had not been widely applied until the work of K. Pearson and R. A. Fisher in the present century. Braithwaite has suggested that the delay in the further development and in the application of probabilistic concepts was due to the tenor of the times. In the Age of Reason, universal hypotheses, which form a part of deductive

logic, were easily regarded as ultimate; statistical hypotheses did not fit the premise that reason alone was capable of solving all human problems, and they were considered, therefore, as temporary substitutes.

These two kinds of hypotheses differ principally with respect to the manner in which they may be rejected. The rejection of a universal hypothesis is, of course, a matter of deductive logic. We may reject the hypothesis "all swans are white" if we find a single black swan. A statistical hypothesis, on the other hand, requires a special rule of rejection. It is with these special rules that this chapter is primarily concerned. We may first note that although a universal hypothesis can be definitively rejected, a statistical hypothesis usually cannot. If we follow the procedure of rejecting a universal hypothesis if and only if we find an A that is not a B, we will never make a mistaken rejection. Unless we examine the entire population in question, which is usually not a practical possibility, it is possible to mistakenly reject a statistical hypothesis. Hence, the special rules of rejection used with statistical hypotheses provide only a provisional rejection.

Since the work of R. A. Fisher in 1912, mathematicians have developed a variety of rules for the rejection of statistical hypotheses. These have been devised for different types of problems and have been called by different names. Rules for deciding whether or not the hypothesis 'that the value of a certain probability differs from a certain number' should be preferred to the hypothesis 'that the value is equal to the specified number' have been called "significance tests." Rules for establishing the preferability of one value over any other have been called "tests of estimation." These tests were originally advanced as ad hoc tests for particular problems arising in applied statistics, and were accompanied by little rational justification. Only in the last 25

years have general principles governing such rules been developed, beginning with the work of J. Neyman and E. S. Pearson. In his book 10 years ago, and in some papers 10 years before that, Abraham Wald presented a completely general decision procedure. This procedure, developed independently by Peterson and Birdsall (Ref. 4), is the basis for a unified theory of signal detectability.

In testing simple statistical hypotheses (which, as we shall see, is the case relevant to the fundamental detection problem), two and only two possibilities are considered: that of rejecting a hypothesis and that of accepting it; or, alternatively, that of accepting one or the other of two mutually exclusive and exhaustive hypotheses. It is assumed that the problem is to determine which of these actions to take when the observed evidence is of certain sorts. The problem is usually represented geometrically: the possible observation values are represented as points in a sample space which is to be divided into a region of acceptance and a region of rejection. (With respect to the null hypothesis, the region of rejection is referred to as the "critical region.")

Before discussing the example we have chosen to illustrate the general problems and procedures, we must add one more thing. The statistical hypotheses of principal interest are not singular probability statements, such as "51% of swans are white," but rather are statements about functional laws of continuous probabilities. As an example of such a hypothesis we might state the probability that an Englishman has height x as a certain function of x in which the parameters take certain values.

As indicated, we shall consider a simple game of chance as an expository device. We shall first describe the game in a very general way to establish the scope of discussion. We shall then work through

the game more carefully, making the computations required. In this game, three dice are thrown. Two of the dice are ordinary dice. The third die is unusual in that on each of three of its sides it has 3 spots, whereas on its remaining three sides it has no spots at all. You, as the player of the game, do not observe the throws of the dice. You are simply informed, after each throw, of the total number of spots showing on the three dice. You are then asked to state whether the third die, the unusual one, showed a 3 or a 0. If you are correct - that is, if you assert a 3 showed when it did in fact, or if you assert a 0 showed when it did in fact - you win a dollar. If you are incorrect - that is, if you make either of the two types of errors possible - you lose a dollar.

How do you play the game? Certainly you will want a few minutes to make some computations before you begin. You will want to know the probability of occurrence of each of the possible totals two through twelve in the event that the third die shows a 0, and you will want to know the probability of occurrence of each of the possible totals five through fifteen in the event that the third die shows a 3. Let us ignore the exact values of these probabilities and grant that the two probability distributions in question will look much like those sketched in Fig. 1.

Realizing that you will play the game many times, you will want to establish a policy which defines the circumstances under which you will make each of the two decisions. We can think of this as a criterion or a cutoff point along the axis representing the total number of spots showing on the three dice. That is, you will want to choose a number on this axis such that whenever it is equaled or exceeded you will state that a 3 showed on the third die, and such that whenever the total number of spots showing is less than this number, you will state that a 0 showed

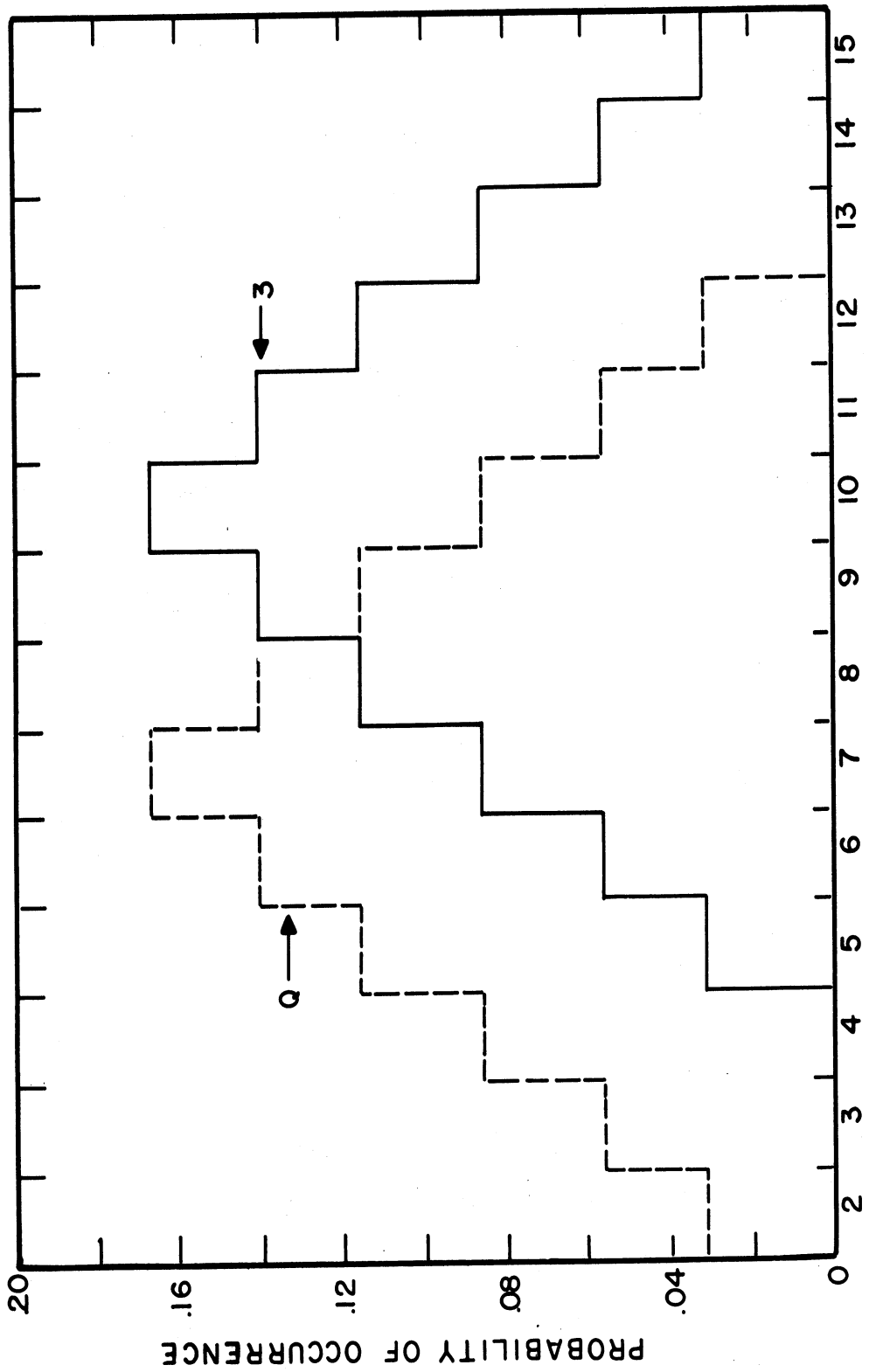
PROBABILITY OF OCCURRENCE
.20
.16
.12
.08
.04
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
TOTAL OF THREE DICE
3
Q

FIGURE I

on the third die. For the game as described, with the a priori probabilities of a 3 and a 0 equal, and with equal values and costs associated with the four possible decision outcomes, it is intuitively clear that the optimal cutoff point is that point where the two curves cross. You will maximize your winnings if you choose this point as the cutoff point and adhere to it.

Now, what if the game is changed? What, for example, if the third die has 3 spots on five of its sides and a 0 on only one? Certainly you will now be more willing to state, following each throw, that the third die showed a 3. You will not, however, simply state more often that a 3 occurred without regard to the total showing of the three dice. Instead, you will lower your cutoff point: you will accept a smaller total than before as representing a throw in which the third die showed a 3. Conversely, if the third die has 3 spots on only one of its sides and 0s on five sides, you will do well to raise your cutoff point - to require a higher total than before for stating that a 3 occurred.

Similarly, your behavior will change if the values and costs associated with the various decision outcomes are changed. If it costs you five dollars every time you state that a 3 showed when in fact it did not, and if you win five dollars every time you state that a 0 showed when in fact it did (the other value and the other cost in the game remaining at one dollar), you will raise your cutoff to a point somewhere above the point where the two distributions cross. Or if, instead, the premium is placed on being correct when a 3 occurred, rather than when a 0 occurred as in the immediately preceding example, you will assume a cutoff somewhere below the point where the two distributions cross.

Again, your behavior will change if the difference between the means of the two distributions is changed. You will assume a different cutoff than you did in the game as first described if the three sides of the third die showing spots now show 4 spots rather than 3.

This game was suggested by Tanner to exemplify the type of problem for which the theory of statistical decision was developed. The theory specifies the optimal behavior in a situation in which one must choose between two alternative statistical hypotheses on the basis of an observed event. In particular, it specifies the optimal cutoff, along the continuum on which the observed events are arranged, as a function of: a) the a priori probabilities of the two hypotheses, b) the values and costs associated with the various decision outcomes, and c) the difference between the means of the distributions that constitute the hypotheses. According to the general theory of signal detectability developed by Peterson and Birdsall (Ref. 4), the problem of detecting signals that are weak, relative to a background of interference, is like the one faced by the player of our dice game. In the fundamental detection problem, the observer must decide, for each observation, whether it arose from random interference, i.e., noise alone, or from the presence of a signal in the noise. The signal is regarded as simply effecting a shift in the mean of the noise distribution, the extent of the shift depending upon the strength of the signal.

In the game we have described, our policy, or rule, for choosing the criterion was that of maximizing the expected value, or the total payoff, of the game. Though no attempt will be made here to treat specifically the other policies that have been put forward, one of them should be mentioned. This policy, suggested by Neyman and Pearson (Ref. 5), takes

historical precedence over the others, and its general familiarity may facilitate understanding of the present essay. In Fig. 2 are portrayed two statistical hypotheses, designated $f_0(x)$ and $f_1(x)$. Sample values, or observation values, denoted x, are arranged along the abscissa, and the ordinate represents the probability (strictly, the probability density, since x is continuous) associated with each of the values of x under each of the hypotheses. Neyman and Pearson dealt with the problem in which the permissible probability of mistakenly rejecting $f_0(x)$ is fixed, i.e., in which the α error, or the Type I error, or the "significance level," is fixed. Their aim, given this restriction, was to select the test, or rule, or critical region, to maximize the probability of accepting $f_1(x)$ when it is true, this probability being referred to as the power of the test. Figure 2 makes clear that good tests and poor tests are possible - if the critical region is defined by $x > a$, the power of the test $[\int_a^\infty f_1(x)dx]$ is far greater than in the case in which the critical region is defined by $b < x < c$ [the power of the test is this case being equal to $\int_b^c f_1(x)dx$], although the Type I error is the same in the two cases, namely, $\int_a^\infty f_0(x)\,dx = \int_b^c f_0(x)dx$.

It might be noted parenthetically, as Braithwaite has, that, in general, a decision rule involves some concept of value. Generally, we want the course of action we decide upon to accomplish some designated purpose. If this were not true, the decision problem would be trivial. All sorts of mechanisms exist that could be used for making a purposeless choice - we could flip a coin or draw from a hat. Since we care whether we are right or wrong when we reject one hypothesis and accept another, we place values on the various decision outcomes, and our policies for choosing between the two hypotheses incorporate a comparison of these

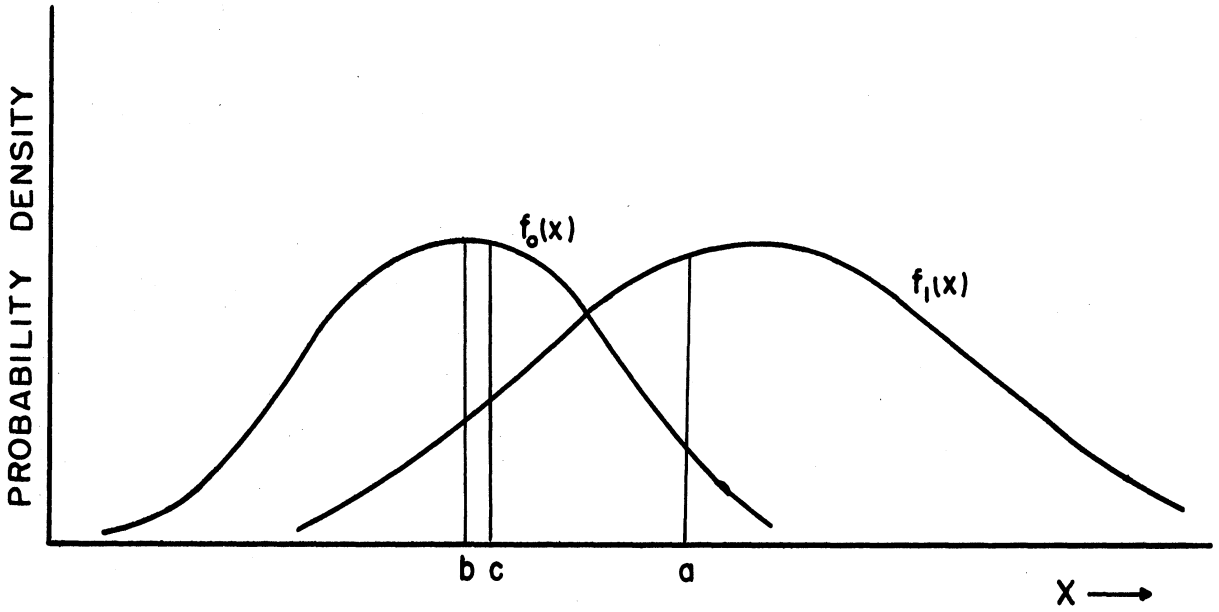

FIGURE 2

values. Thus, the outcomes considered in decision theory have two numbers associated with them - a probability and a desirability. (This is most explicit in the expected value policy, in which the expected value is defined as the sum, over the various outcomes, of the products of probability and desirability.) In the case of probabilities, we have come a long way in the transition from words to numbers - but not so with desirabilities. About the only scale of value available is the monetary scale. J. S. Mill suggested that the pleasure-pain dimension might be quantified, but no one has succeeded in doing so. It seems interesting to note that, logically, this chapter would be preceded not only by a chapter on the theory of probability, but also by a parallel account of the theory of desirability, but there is none.

Let us return now to the dice game to consider it in more detail. The first problem is to determine the probability of each of the sums

of the three dice (and let us denote the sum as x) given a 0 on the third die, and given a 3 on the third die. The ways in which each sum can possibly occur are shown in Table I in the Appendix. In Table Ia, which assumes a 0 on the third die, and in Table Ib which assumes a 3 on the third die, we read the positive diagonals to determine the relative frequency of each sum. We see that the probability of a 2 showing is 1/36 = .0277, that the probability of a 3 is 2/36 = .0555, and so forth. The probability of each x, given 0 on the third die, $P_0(x)$, is shown in the second column of Table II. Similarly, $P_3(x)$ is given in the third column. (Ignore for now the fourth column of Table II.)

Now, assume that the occurrence of a 0 and of a 3 are equally likely, and that we are confronted with the sum x = 7. We note that:

$$P(0.7) = P(0)P_0(7) = .5 \times .1667 = .0833,$$

$$P(3.7) = P(3)P_3(7) = .5 \times .0833 = .0417,$$

(the probability of a joint occurrence of a 0 on the die and x = 7 is equal to the a priori probability of a 0 times the conditional probability of x = 7 given a 0) and that

$$P(7) = .0833 + .0417 = 9/72 = .125$$

is the total probability of an occurrence of x = 7. We can determine the a posteriori, or inverse, probabilities by Baye's theorem. These are equal to the joint probability divided by the total probability, thus:

$$P_7(0) = \frac{.0833}{.125} = .667$$

$$P_7(3) = \frac{.0417}{.125} = .333$$

It is clear that when a 0 and a 3 are equally likely to occur, and when we observe x = 7, we should always state that a 0 occurred on the third die, for the probability that a 0 occurred, given x = 7, is greater than

the probability that a 3 occurred. We can make the same analysis for each sum. Given that a 0 and a 3 are equally likely, whenever the probability that a 0 led to the x observed is greater than the probability that a 3 led to the x observed, we should say 0; otherwise, we should say 3. So in this game (see Table II), for sums of 8 or less, we say 0; for sums of 9 or more, we say 3. The optimal cutoff, or criterion, is at $x = 8.5$.

If the a priori probabilities of a 0 and a 3 are not equal, we must take them into account explicitly in setting the criterion. This is very simple. We have said that we decide in favor of the hypothesis 3 when

$$P(3)P_3(x) > P(0)P_0(x)$$

Rearranging, we have:

$$\frac{P_3(x)}{P_0(x)} > \frac{P(0)}{P(3)} .$$

We thus decide in favor of the hypothesis 3 whenever the quantity to the left of the inequality, designated the likelihood ratio, $\ell(x)$, is greater than the ratio of the a priori probabilities. The ratio of the a priori probabilities establishes the optimal cutoff value of $\ell(x)$, call it β. The likelihood ratios for the game under discussion are tabulated in the fourth column of Table II.

In our game, if the third die contains a 0 on one of its sides and 3s on the other five sides, then

$$\frac{P(0)}{P(3)} = \frac{1/6}{5/6} = 0.20$$

and we should say 3 whenever $\ell(x) > 0.20$, that is, for $x \geq 5$. Similarly if there are two 0s and four 3s on the third die, i.e., if

$$\frac{P(0)}{P(3)} = \frac{2/6}{4/6} = 0.50$$

One of the chief features of the concept of likelihood ratio, which was Wald's insight, and Peterson's and Birdsall's, is that the criterion can be defined in terms of likelihood ratio for any of the policies, i.e., for any of the definitions of optimum, that we might consider. It permits a completely general decision procedure that may be applied irrespective of the particular quantity that we wish to maximize. It applies to the expected value decision rule that we have emphasized, to the policy of maximizing the power of a test given a significance level as described by Neyman and Pearson, to the policy of maximizing the number of correct decisions, and so on.

We have one more major topic to consider, that of operating characteristics. As we have mentioned, we can speak of the power of a test, $P_3(III)$, in terms of our dice game, and of the power function of a series of tests. Figure 3 shows power functions for each of two one-tailed tests, $\theta > \theta_0$(a) and $\theta < \theta_0$(b), and for a two-tailed test of $\theta \neq \theta_0$(c). Statisticians frequently speak of operating characteristics as 1 minus the power function. Figure 4 shows the operating characteristics for the three tests whose power function is represented in Fig. 3. The ordinate, again in terms of our game, is $P_3(\overline{0})$, i.e., the probability of accepting the null hypothesis when it is false, or the probability of a Type II error.

There are some reasons for preferring a slightly different plot to represent the relations among the quantities in question, namely, a plot of power, $P_3(III)$, or 1 minus the probability of a Type II error, vs. the probability of a Type I error, $P_0(III)$, with $\theta - \theta_0$ as the parameter. This plot for θ = the mean of a distribution, is shown in Fig. 5. (The exact form of the plot, as shown, assumes that the distributions $f_{\theta_0}(x)$

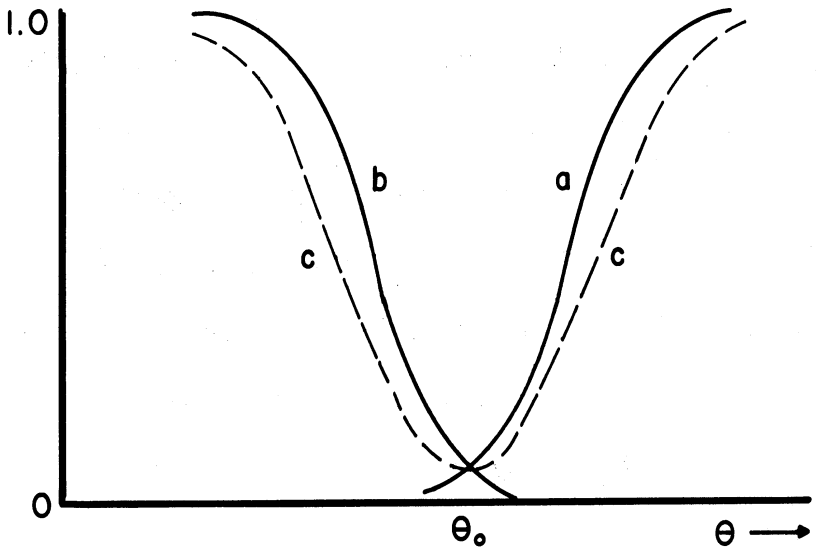

FIGURE 3

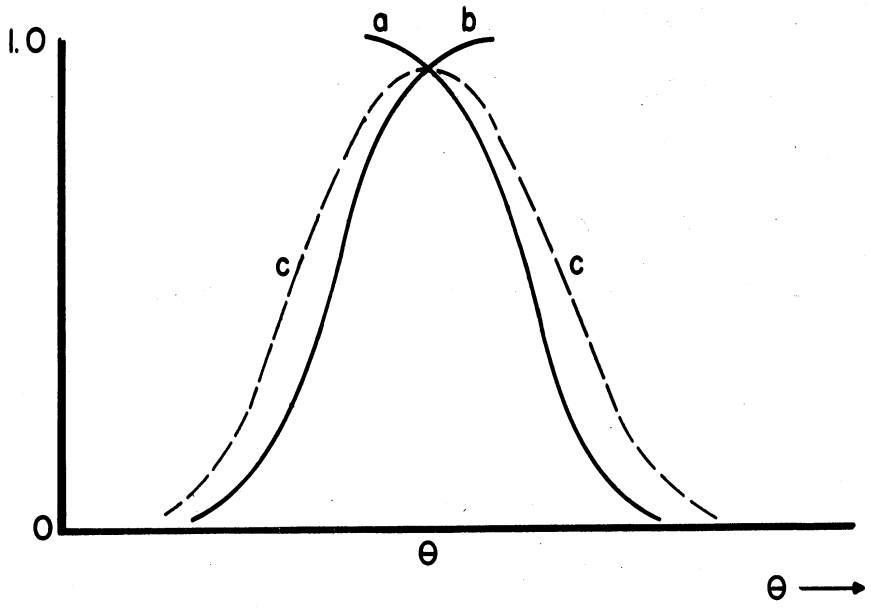

FIGURE 4

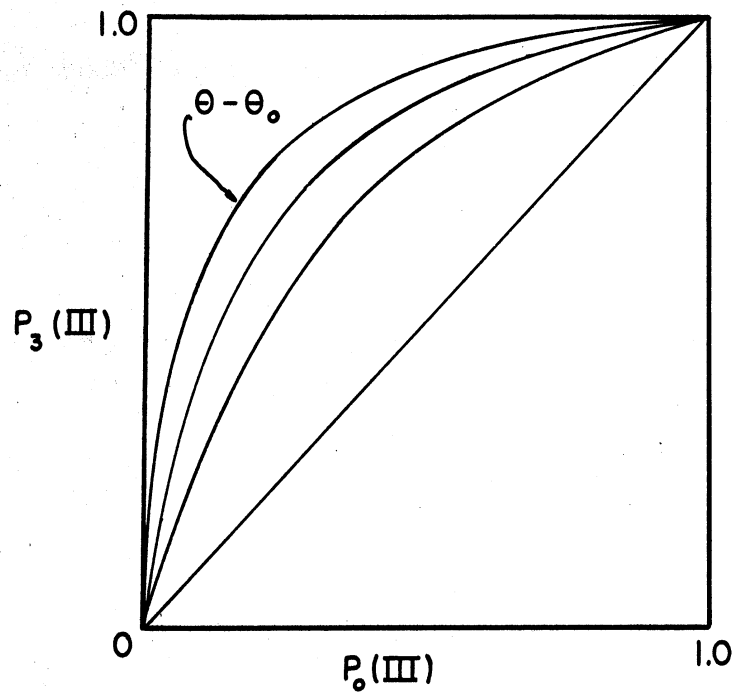

FIGURE 5

and $f_{\theta_1}(x)$ are normal and of equal variance.) The general form of the operating characteristic gives a prominent and parallel place to both types of errors; the general form shown in Fig. 4 was suggested by the Neyman-Pearson test in which the Type I error is arbitrarily fixed.

It is interesting to note, in conclusion, the relationship of the optimal criterion, or the optimal "operating level," to the operating characteristic. As we have indicated, the general aim in hypothesis testing is to maximize some quantity. In particular, the aim is to maximize the expression, $P_3(III) - \beta P_0(III)$ (in the terms of our example), in which the value of β (the critical value of likelihood ratio) depends on the quantity to be maximized. (The value β, in the terminology of the signal detection problem, weights the "hits" relative to the "false alarms.") It can be shown that the expression, $P_3(III) - \beta P_0(III)$, defines a utility line of slope β, and that the point of

tangency of this line to a particular operating characteristic curve is the optimal operating level in a situation in which that curve applies. More important, the choice of a criterion made by a decision maker, even if not the optimal one, can be determined from his behavior by means of such a plot. He yields a pair of values, $P_3(III)$ and $P_0(III)$, i.e., a point on an operating characteristic curve. The slope of the curve at this point corresponds to the value of likelihood ratio at which he has located his criterion. Thus, from a 2 x 2 table - in the case of our dice game:

		Event: 0	Event: 3
Statement	$\overline{0}$	$P_0(\overline{0})$	$P_3(\overline{0})$
	III	$P_0(III)$	$P_3(III)$
		1.00	1.00

or in the case of the detection problem:

		Event: Noise Alone (N)	Event: Signal plus Noise (SN)
Statement	Noise Alone (B)	$P_N(B)$	$P_{SN}(B)$
	Signal + Noise (A)	$P_N(A)$	$P_{SN}(A)$
		1.00	1.00

the theory of statistical decision permits one to extract two quantities, one equal to $\theta - \theta_0$ (or d' as it has been called in the detection problem), that is, an index of the effective signal strength or of the sensitivity of the observer, and another equal to β, that is, an index of his criterion. The ability to separate these two aspects of detection performance is one of the principal advantages of the application of statistical decision theory to the detection problem.

APPENDIX

Table I(a)

		Number Facing Up on One Die					
		1	2	3	4	5	6
Number Facing Up on Second Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	8	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table II(b)

		Number Facing Up on One Die					
		1	2	3	4	5	6
Number Facing Up on Second Die	1	5	6	7	8	9	10
	2	6	7	8	9	10	11
	3	7	8	9	10	11	12
	4	8	9	10	11	12	13
	5	9	10	11	12	13	14
	6	10	11	12	13	14	15

Table II

Total of 3 dice	$P_0(x)$	$P_3(x)$	$\ell(x) = P_3(x)/P_0(x)$
2	.0277	.0000	0.00
3	.0555	.0000	0.00
4	.0833	.0000	0.00
5	.1111	.0277	0.25
6	.1388	.0555	0.40
7	.1667	.0833	0.50
8	.1388	.1111	0.80
9	.1111	.1388	1.25
10	.0833	.1667	2.00
11	.0555	.1388	2.50
12	.0277	.1111	4.00
13	.0000	.0833	∞
14	.0000	.0555	∞
15	.0000	.0277	∞

REFERENCES

1. Wald, A. Statistical Decision Functions. New York: Wiley, 1950.
2. Braithwaite, R. B. Scientific Explanation. London: Cambridge University Press, 1953.
3. Bross, I. D. J. Design for Decision. New York: Macmillan, 1953.
4. Peterson, W. W., Birdsall, T. G., and Fox, W. C. "The Theory of Signal Detectability," Trans. IRE, 1954, PGIT-4, 171-212.
5. Neyman, J. and Pearson, E. S. "On the Problem of the Most Efficient Tests of Statistical Hypotheses," Philosophical Transactions of the Royal Society (London), Ser. A, vol. 231, 1933, 289.