Operating Characteristics Determined by Binary Decisions and by Ratings

James Egan, , Arthur I. Schulman, and , and Gordon Z. Greenberg

Citation: The Journal of the Acoustical Society of America **31**, 768 (1959); doi: 10.1121/1.1907783
View online: http://dx.doi.org/10.1121/1.1907783
View Table of Contents: http://asa.scitation.org/toc/jas/31/6
Published by the Acoustical Society of America

Operating Characteristics Determined by Binary Decisions and by Ratings*

JAMES P. EGAN, ARTHUR I. SCHULMAN, AND GORDON Z. GREENBERG
Hearing and Communication Laboratory, Indiana University, Bloomington, Indiana

(Received February 2, 1959)

With the theory of signal detectability as a framework, two psychophysical experiments were conducted in which each observation interval was well defined for the listener. Each interval contained noise, and it either did or did not ($p=0.5$) contain a signal (1000 cps, 0.5 sec in duration). In separate sessions of the first experiment, either the listener gave a yes-no decision or he responded with a rating (1-4) after each observation interval. Operating characteristics were obtained with E/N_0 equal to 15.8. It is clear from the data that the trained listener can perform as well when he adopts the multiple criteria necessary for the rating method as when he adopts the single criterion required by the binary-decision procedure. In the second experiment, only the rating method was used to determine the relation between d' and E/N_0. The resulting function, for $d'\leq 3.0$, approximates a straight line which passes through the origin and which has nearly the same slope as that obtained in other laboratories.

THE theory of signal detectability provides a framework for the study of the detectability of a signal by the human listener.[1–10] Within this framework, it is possible to derive a measure of performance which is relatively independent of the particular procedure employed.[11] This measure of performance is d', and, unlike the threshold as it is usually measured, the value of d' is largely unaffected by those circumstances that may influence the criterion adopted by the listener.

One of the procedures that has been developed within the framework of the theory of signal detectability has been called the "fixed-interval observation experiment." For each observation by the listener, a single temporal interval is well defined. These intervals consist of two types: (1) those that contain the signal plus Gaussian noise, and (2) those that contain noise alone. Thus, there are two stimulus situations, and there is no longer any special meaning intuitively ascribed to the presentation of a "stimulus." After each interval, the listener responds with *yes* or *no*. Each of these decisions is made according to some fixed arbitrary criterion adopted by the listener. The two types of intervals, SN and N, and the two responses, y and n, result in four stimulus-response conjunctions. The important descriptive probabilities are as follows: $p(SN)+p(N)=1.0$, $p(y|SN)+p(n|SN)=1.0$, and $p(y|N)+p(n|N)=1.0$. By empirical tests, it turns out that the three, formally independent, probabilities $p(y|SN)$, $p(y|N)$, and $p(SN)$ give a comprehensive description of the behavior of the real listener in the detection of a signal in noise.

For fixed average conditions of signal and noise, the two conditional probabilities $p(y|SN)$ and $p(y|N)$ are determined for several different criteria, and the relation between these two criterial probabilities is termed the *operating characteristic*. From the operating characteristic, the measure of detectability d' is derived.

Actually, for the determination of d', a temporal forced-choice procedure (with two or more observation intervals per trial) is more economical than the procedure that uses a single fixed interval for an observation. However, for certain purposes, the nature of the problem may demand that a single interval for observation be employed. In such cases, the basic procedure requires that the subject make a binary decision. According to this binary-decision procedure, the listener adopts one criterion for a series of observation intervals. He then adopts a different criterion for the next series of intervals, and this process is continued until a sufficient number of pairs of values of the two criterial probabilities is obtained. Now, if the listener adopts a given criterion for a long series of observation intervals, many of his yes-no decisions are well within and many well outside that criterion. Consequently, the listener should be able to order his yes-no decisions and thereby assign ratings to each of his responses. The rating method utilizes the fact that, during a single series of

* This research was supported by the U. S. Air Force under Contract No. AF 19(604)-1962, monitored by the Operational Applications Laboratory, Air Force Cambridge Research Center. This is Rept. No. AFCRC-TN-59-50.

[1] W. W. Peterson and T. G. Birdsall, "The theory of signal detectability," Tech. Rept. No. 13, Electronic Defense Group, Department of Electrical Engineering, University of Michigan, Ann Arbor, Michigan (June, 1953).

[2] M. Smith and E. A. Wilson, Psychol. Monogr. **67**, No. 9, 1–35 (1953).

[3] W. P. Tanner, Jr., and J. A. Swets, Psychol. Rev. **61**, 401–409 (1954).

[4] W. P. Tanner, Jr., J. Acoust. Soc. Am. **28**, 882–888 (1956).

[5] T. Marill, "Detection theory and psychophysics," MIT. Research Laboratory of Electronics, Tech. Rept. No. 319 (1956).

[6] Green, Birdsall, and Tanner, J. Acoust. Soc. Am. **29**, 523–531 (1957).

[7] J. P. Egan, "Message repetition, operating characteristics, and confusion matrices in speech communication," AFCRC-TR-57-50, ASTIA Document No. AD 110064, Hearing and Communication Laboratory, Indiana University (1957).

[8] I. Pollack and L. R. Decker, J. Acoust. Soc. Am. **30**, 286–292 (1958).

[9] J. P. Egan, "Recognition memory and the operating characteristic," AFCRC-TN-58-51, ASTIA Document No. AD 152650, Hearing and Communication Laboratory, Indiana University (1958). For an earlier instance in which the investigators partitioned the subject's ratings between "old" and "new" stimuli, see Table I, p. 46, in the paper by J. C. R. Licklider and I. Pollack, J. Acoust. Soc. Am. **20**, 42–51 (1948).

[10] W. P. Tanner, Jr., and T. G. Birdsall, J. Acoust. Soc. Am. **30**, 922–928 (1958).

[11] John A. Swets, J. Acoust. Soc. Am. **31**, 511 (1959).

observation intervals, the human listener is capable of adopting multiple criteria. For comparable reliability, the rating method with 4 categories should require about one-third the number of trials as that used for the binary-decision procedure, provided the listener can in fact maintain several criteria during a single series of observation intervals. If it can be shown that nearly the same value of d' is obtained by the rating method as is obtained by the binary-decision procedure, then a very considerable saving will result in future determinations of d'.

The relation between the binary-decision procedure and the rating method is as follows. In the rating method, the assignment of 1 represents a "yes" under a strict criterion. If the observation interval in fact contains the signal, then $p(r_1|SN)$ is equivalent to $p(y|SN)$, where r_1 is the assigned rating 1. Similarly, for the noise-alone intervals, $p(r_1|N)=p(y|N)$.

Next, the conditional probability of the assignment of 2 to an SN interval is computed. The reasonable assumption is then made that all "1's" would fall within the criterion designated by the assignment of "2's." More generally speaking, the assumption is the following: observation intervals that are accepted as containing the signal under a given criterion will also be accepted under a less strict criterion.[12] Therefore, the second value of $p(y|SN)$ is $p(r_1|SN)+p(r_2|SN)$. Thus, the value of $p(y|SN)$ corresponding to the criterion established by the rating c is the cumulative probability

$$p(y|SN)=\sum_{i=1}^{c} p(r_i|SN).$$

For the same criterion, the corresponding value of $p(y|N)$ is

$$p(y|N)=\sum_{i=1}^{c} p(r_i|N).$$

Computation of $p(y|SN)$ and $p(y|N)$ over the entire rating scale, $i=1, 2, \cdots, r$, generates the operating characteristic.

The following two experiments were conducted in order to compare the values of d' and (σ_{SN}/σ_N) as they are obtained by two methods, the binary-decision procedure and the rating method.

GENERAL PROCEDURE

A "fixed-interval observation experiment" is one in which (1) the listener knows that a signal, if presented, will occur in a well-delimited interval of time, and in which (2) a decision of some sort is required after each observation. A single temporal interval for observation by the listener is the principal part of a *trial*, and Fig. 1 diagrams the sequence of events. In the experiments, a trial began with a white light (W) which served as a warning signal and which remained on for 1.5 sec. The last 0.5 sec of W was the fixed interval for observation, and was marked by the presence of a red light (R). The signal, when presented, also had a duration of 0.5 sec with its onset and offset synchronous with the onset and offset of R. The response interval was 5 sec, and the listener responded with a binary decision or with a rating, as the case might be, by pressing the appropriate key during the early part of this interval. The subject was not informed whether his decision was correct or incorrect.

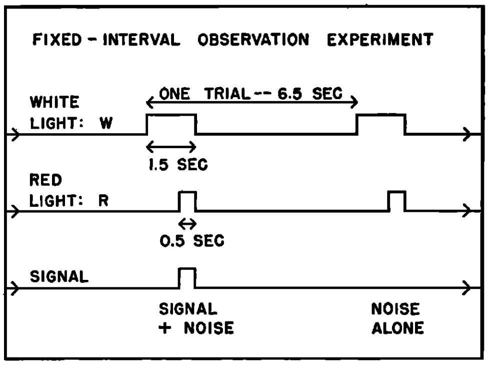

FIG. 1. Schema showing the sequence of events in each trial. The red light (R) marked the interval for observation, and the signal of 1000 cps was presented with a probability of 0.5. The listener responded by pressing the appropriate key after the termination of W and R. Although the white noise is not represented, it was on continuously throughout the session.

The signal was a "pure tone" of 1000 cps. The signal voltage was turned on without regard for phase and without the use of special devices, so that the (negligible) transients were determined by the response of the earphone (Permoflux Corporation, PDR-10). In both experiments, the *a priori* probability of the signal $p(SN)$ was 0.5, and the listener knew this fact. The white noise was generated by a 6D4-tube (Noise Generator, Model 455-B, Grason-Stadler Company). The over-all sound pressure level of the noise was about 65 db *re* 0.0002 μbar, and it was held constant for all tests at this comfortable level for listening. The signal and the noise were mixed electrically and then presented over a binaural headset with the two earphones wired in parallel and in phase. The noise was present continuously throughout the session.

The primary environmental condition will be specified in the present experiments by the value of E/N_0, in which E is the signal energy, or the time integral of power, and N_0 is the noise power per unit band width.

EXPERIMENT I

In separate sessions, data were obtained by the binary-decision procedure and by the rating method. In order to obtain a quite different criterion from one series of trials to the next for the binary-decision procedure, the listeners were first trained to adopt

[12] The rating scale as here applied is assumed to have only ordinal properties; no assumption is made with regard to the numerical properties of the rating "scale" other than that of order.

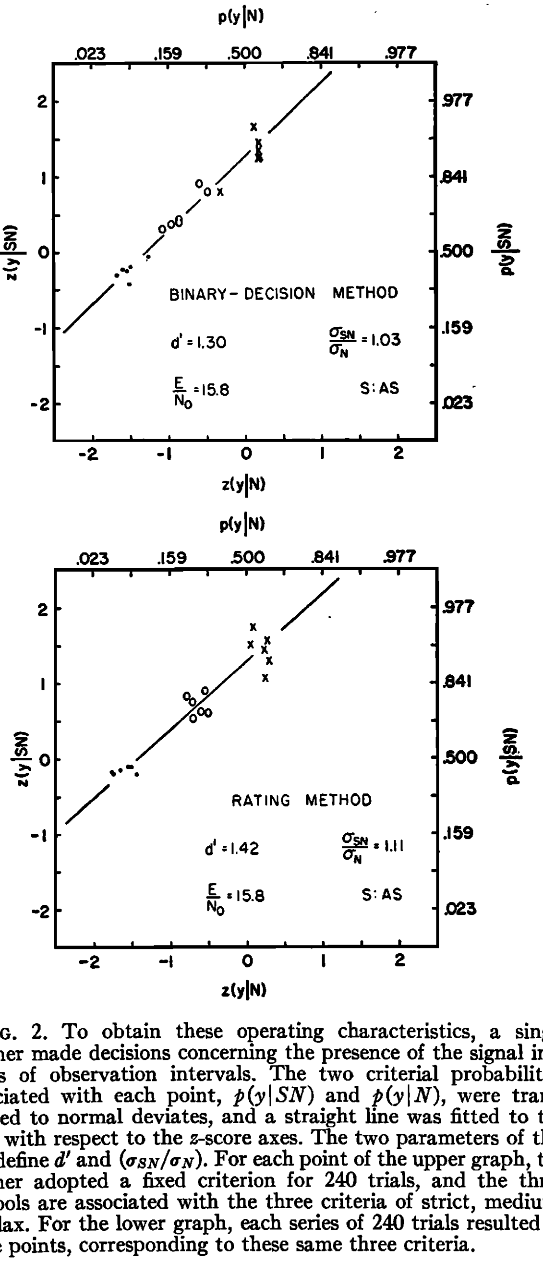

FIG. 2. To obtain these operating characteristics, a single listener made decisions concerning the presence of the signal in a series of observation intervals. The two criterial probabilities associated with each point, $p(y|SN)$ and $p(y|N)$, were transformed to normal deviates, and a straight line was fitted to the data with respect to the z-score axes. The two parameters of this line define d' and (σ_{SN}/σ_N). For each point of the upper graph, the listener adopted a fixed criterion for 240 trials, and the three symbols are associated with the three criteria of strict, medium, and lax. For the lower graph, each series of 240 trials resulted in three points, corresponding to these same three criteria.

three different criteria which were specified as "strict," "medium," and "lax." For purposes of instructing the listener during the experiment proper, the following conditions were chosen: (1) *strict*, $0.05 \leq p(y|N) \leq 0.10$; (2) *medium*, $p(y|SN)+p(y|N)=1.00\pm0.07$; (3) *lax*, $0.90 \leq p(y|SN) \leq 0.95$. After each block of 240 trials, all at a given criterion, the experimenter calculated the conditional probabilities of a "yes." If the obtained value(s) fell within the appropriate range indicated in the foregoing, the listener was told that his performance was "good." If the value(s) fell outside the specified range, the listener was told that he had been "too strict" or "too lax" as the case might be. These remarks were repeated to the listener the next time he used that criterion in an effort to induce him to operate within the desired range. For the rating method, a similar instructional procedure was used. The listener divided his responses into four categories. On this scale, a "1" was meant to correspond approximately to a "yes" given as a response when the listener was making binary decisions under a strict criterion; a "2" corresponded to a "yes" with a medium criterion; a "3" to a "yes" with a lax criterion; and a "4" to a "no" with a lax criterion. No other instructions were given. The listener was not informed of the *level* of his performance at any time during the experiment.[13]

Nine daily sessions were administered individually to three listeners. On days 1–3 and 7–9 the listener used binary decisions. On day 4 the listener was given practice on the rating method, and on days 5 and 6 he was tested on the rating method. Each daily session consisted of 9 test periods of 80 trials each, separated by short rest intervals. In order to reinforce the listener's memory for the signal frequency, an extra trial was given before each test period with the signal intensity increased by 10 db over its normal level.

For each binary-decision session, three consecutive periods, giving a total of 240 trials, were devoted to each of the 3 criteria. Over the 6 sessions, each listener was presented once with each of the 6 ways of ordering the 3 criteria. All told, 4320 test trials under the binary-decision procedure were administered to each listener. Each of these same listeners rated 1440 test trials. The computation of each pair of criterial probabilities was based upon the 240 trials that occurred in 3 consecutive periods. In about one-half of these 240 trials, the observation interval contained the signal. Perhaps it should be pointed out that, in the binary-decision procedure, only one point for the operating characteristic was provided by 240 trials; on the other hand, in the rating method, 3 points were available from 240 trials.

All tests for this experiment were conducted with $E/N_0=15.8$; expressed in decibels this value is

$$10\log(E/N_0)=12 \text{ db}.$$

The results for each listener are shown separately as operating characteristics in Figs. 2–4. The data obtained by the binary-decision procedure and by the rating method are shown, respectively, in the upper and the lower graphs of each figure. The data for each operating characteristic are plotted on normal-normal coordinates.

[13] The measure d' depends upon how well the listener *partitions* each of his response categories between the two stimuli, SN and N. In situations that require a binary decision, the response probabilities are simply $p(y)+p(n)=1.0$; the corresponding probabilities for the rating method are: $p(r_1)+\cdots+p(r_4)=1.0$. These probabilities are easily manipulated by a host of variables, including the instructions given to the listener. Some psychologists are concerned with the social, motivational, and learning variables which affect these response probabilities. However, it should be clear by now that, in the study of signal detection, the great virtue of d' is its independence from such sociopsychological factors.

For this purpose, each probability scale, $p(y|SN)$ and $p(y|N)$, is so transformed that the corresponding normal deviates (z scores) are linearly spaced. In the figures, these z-score axes are referred to as $z(y|SN)$ and $z(y|N)$.

In effect, a curve was fitted to the data by passing the criterion cut through two normal curves of equal area plotted on a common decision axis. When the operating characteristic is generated in this way, it is linear on a z-score plot. Two parameters, based upon these two normal curves, were then adjusted to give a least-squares fit to the data. One of these parameters is the difference between the means of the two normal curves, expressed as follows: $d' = (M_{SN} - M_N)/\sigma_N$. The other independent parameter is (σ_{SN}/σ_N).

Actually, a straight line was fitted to the data by that method of least squares which minimizes the sum of the squared perpendicular distances between the data points and the straight line. The two constants of this straight line determine the two parameters required to specify the two normal curves referred to previously. For purposes of simplicity, the value of d' will be taken as the negative of the "x intercept," that is to say, d' is here defined as the negative of the coordinate $z(y|N)$ for the point at which the operating characteristic intersects the horizontal line, $z(y|SN) = 0$. The ratio (σ_{SN}/σ_N) is given by the reciprocal of the slope of the straight line with respect to the z-score axes.[14]

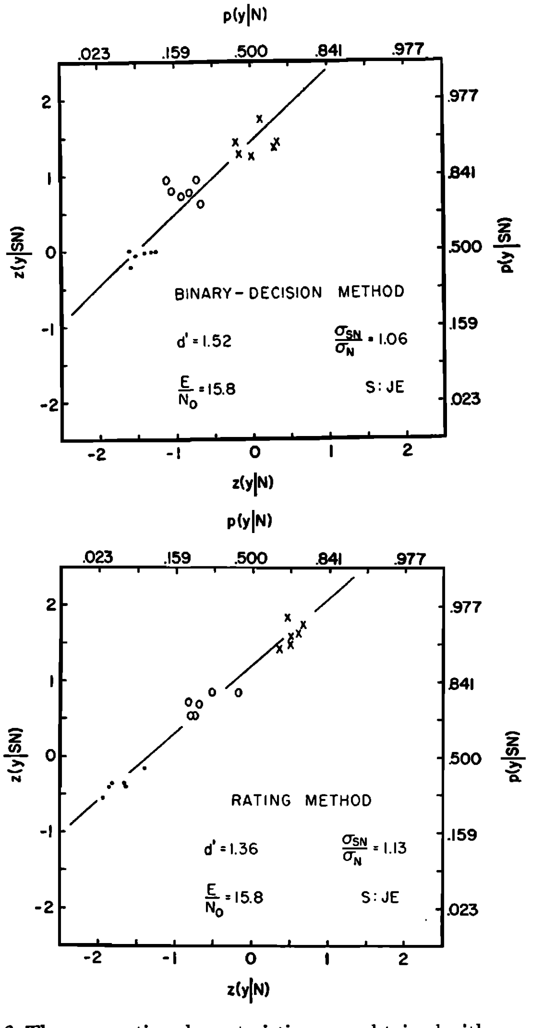

FIG. 3. These operating characteristics were obtained with a second listener under the same conditions as those for Fig. 2.

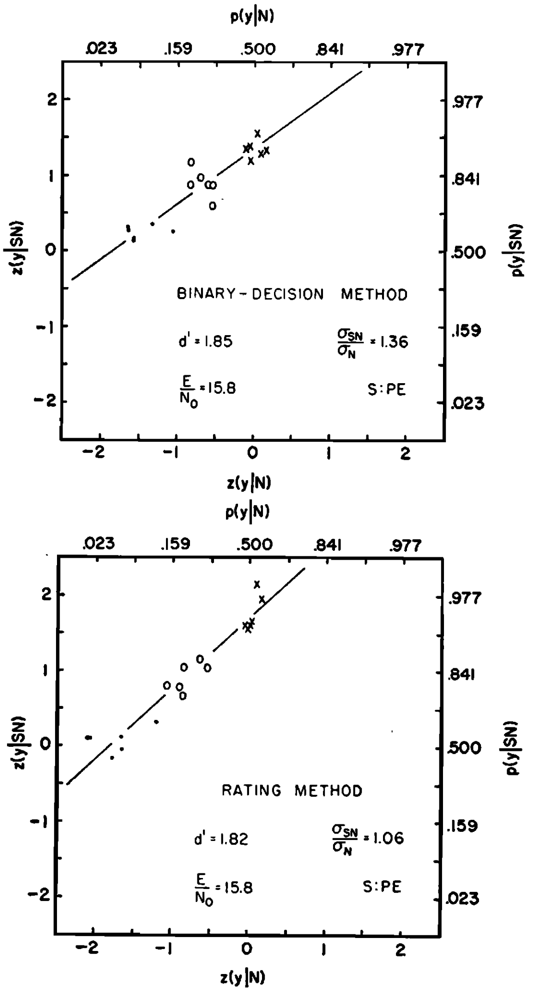

FIG. 4. These operating characteristics were obtained with a third listener under the same conditions as those for Fig. 2.

[14] Strictly speaking, we are taking the value of d' at which $p(y|SN)$ is 0.5. When the slope of the straight line of best fit is not one, d', as defined in reference 10, becomes a function of the criterion adopted. Nevertheless, the negative of the "x intercept" and the reciprocal of the slope are two constants that give a comprehensive description of the listener's behavior. In a personal communication, Clarke, Birdsall, and Tanner have recently defined a measure of performance which they denote by $\sqrt{d_e}$. The value of $\sqrt{d_e}$ is computed as follows. The coordinates with respect to the z-score axes of the ***point of intersection*** between the ROC curve and the negative diagonal are determined. The ordinate minus the abscissa of this point of intersection is the value of $\sqrt{d_e}$. In Experiment I, the mean value of $\sqrt{d_e}$ for the binary-decision procedure is 1.45, and the corresponding value for the rating method is 1.46.

In reference 9, the advantages of specifying performance in terms of the coordinates of the point of intersection are discussed.

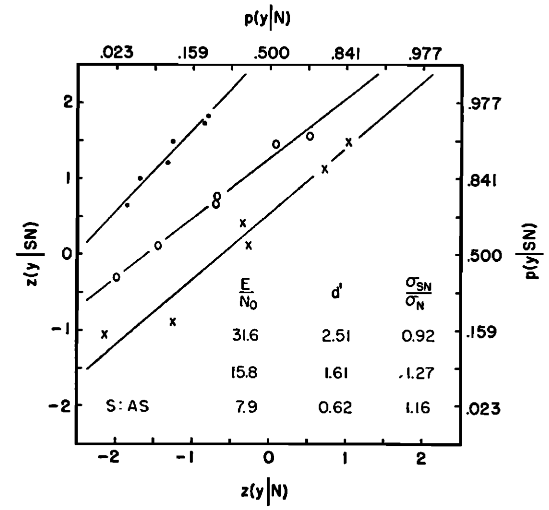

FIG. 5. These operating characteristics were obtained by the rating method. At each of two sessions, 480 test trials were conducted at a given value of E/N_0. These 480 test trials resulted in three points for a given operating characteristic. A straight line was fitted to the data with respect to the z-score axes, and the two parameters of this line define d' and (σ_{SN}/σ_N).

It is fairly obvious from an inspection of Figs. 2–4 that the data fall along a straight line. Of course, the obtained proportions plotted in Figs. 2–4 are subject to random sampling errors. The scatter of each set of points is well within the limits expected on the basis of chance fluctuations.

Each operating characteristic shown in Figs. 2–4 has a value of d' and of (σ_{SN}/σ_N). These values are also presented in Table I for comparative purposes. It is clear that the trained listener can perform as well when he adopts the multiple criteria necessary for the rating method as when he adopts the single criterion required by the binary-decision procedure.

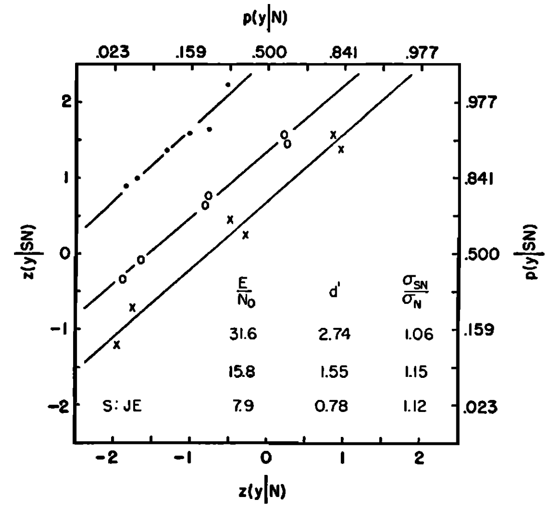

FIG. 6. These operating characteristics were obtained with a second listener under the same conditions as those for Fig. 5.

The theory of signal detectability makes it possible to set an exact upper bound upon the detectability of the signal. Therefore, the efficiency of the real listener may be expressed in terms of the performance of an ideal observer.[10] The amplitude of the signal used for the listener is attenuated until the performance of the ideal observer is the same as that of the listener. The ratio of the corresponding signal-energies defines the efficiency η of the listener.[15] On the basis of the average d' for the three listeners (Table I), the value of η is 0.076, or −11 db.

EXPERIMENT II

In the second experiment, only the rating method was used to determine the relation between d' and E/N_0. The same three listeners were employed. For two of the listeners (JE and AS), operating characteristics were determined at each of the following values of E/N_0: 7.9, 15.8, and 31.6. For the other listener (PE), the three E/N_0's were: 6.3, 12.6, and 25.1. Each subject was run for 6 sessions, 2 at each of the three E/N_0's. A session consisted of 9 periods of 80 trials each. The first 3 periods of each session were "practice," and, on the basis of these 240 trials, the subject was informed whether or not he had adopted the appropriate criteria for each of the rating categories. These criteria corresponded to those established for the binary-decision procedure in Experiment I. The criterial probabilities for the operating characteristic were estimated from the 480 test trials given in the last 6 periods of each session. As in Experiment I, the probability that a signal would be presented was 0.5, so that each criterial probability is based upon about 240 test trials.

The sequence of warning signals and the corresponding temporal intervals were the same as those shown in Fig. 1.

Figures 5–7 display the results as operating characteristics on normal-normal coordinates. The straight

TABLE I. Comparison of values of d' and of values of (σ_{SN}/σ_N) as obtained by the binary-decision method (BD) and by the rating method (R). The data are taken from Figs. 2–4 in which E/N_0=15.8.

	d'		(σ_{SN}/σ_N)	
Subject	BD	R	BD	R
AS	1.30	1.42	1.03	1.11
JE	1.52	1.36	1.06	1.13
PE	1.85	1.82	1.36	1.06
Mean	1.56	1.53	1.15	1.10

When the *a priori* probabilities of the two types of intervals are equal, then, at the point of intersection, (1) one-half of the intervals are in the criterion of acceptance, and (2) the rate of correct acceptance is equal to the rate of correct rejection.

[15] In this computation, it is assumed that the signal is known exactly. If the phase at the onset of the signal had been known to the listener, he almost certainly could not have used this information in the situation being studied here. Therefore, part of the inefficiency of the listener, relative to the ideal observer, is his inability to use phase information.

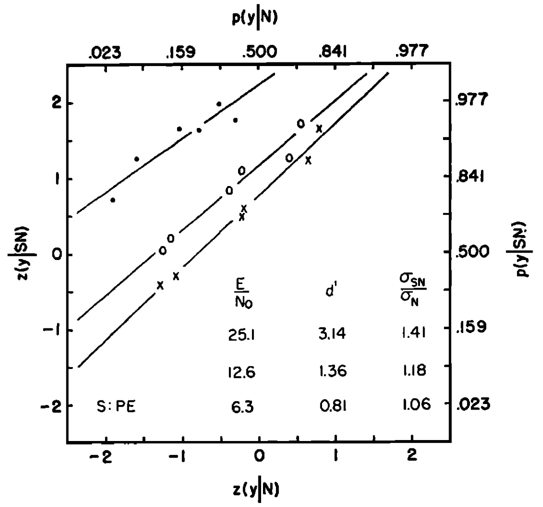

FIG. 7. These operating characteristics were obtained with a third listener under the same conditions as those for Fig. 5.

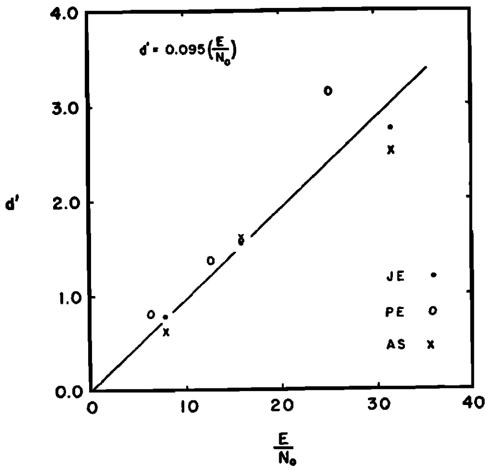

FIG. 8. These data are based upon the operating characteristics of Figs. 5–7. The derived measure d' is plotted against E/N_0, and, for $d' \leq 3.0$, the resulting function may conveniently be considered as linear.

line passing through the data points was fitted by the same method of least squares as that used in Experiment I. The constants, d' and (σ_{SN}/σ_N), associated with each operating characteristic, are also shown. For two of the listeners, JE and AS, one of the values of E/N_0, 15.8, was used in Experiment II as well as in Experiment I. Both listeners showed a small improvement in performance from the first to the second experiment.

The values of d' taken from Figs. 5–7 are plotted as a function of E/N_0 in Fig. 8. A straight line, constrained to pass through the origin, was fitted to the data by the standard method of least squares. The equation of this line is $d' = 0.095(E/N_0)$. The constant, 0.095, differs from that obtained in other laboratories by the equivalent of about 1 db in the value of E/N_0. On the basis of all available data, this constant appears to be about 0.08.

On the basis of the foregoing experiments, it is concluded that the rating method (4 categories) is superior to the binary-decision procedure for the following reasons. (1) Nearly the same measures of performance are obtained by the two methods. (2) Although only one-third as many trials were administered for the rating method as for the the binary-decision procedure, these two methods provide about the same reliability.

ACKNOWLEDGMENTS

The writers are indebted to Dr. James D. Miller and to Mr. Charles S. Watson for their assistance in the design and construction of Audiac, a flexible assembly of components for the investigation of the detection of signals presented at fixed or at random times.