Perception & Psychophysics
1974, Vol. 16 (3), 478-488

Canonical ratings*

B. SAKITT
Stanford University, Stanford, California 94305

The concept of canonical ratings is introduced in which each S describes all the visual sensations produced by signal and noise trials in the expected spatial and temporal location of the stimulus. After many practice sessions, the S assigns one and only one numerical rating to each visual sensation. These canonical ratings are determined by the S, not the E, and are abbreviations for verbal descriptions of subjectively distinct visual sensations. The data consisted of canonical ratings at absolute visual detection for dim visual stimuli (signal) and blank (noise) trials containing no light at all. The physical stimulus is discrete since it is made up of absorptions of quanta of light that result in isomerizations of rhodopsin molecules or thermal decompositions of rhodopsin which are discrete noise events that mimic the action of quantal absorptions. Under these conditions, it is known from the laws of physics that these quantum-like events (absorptions plus thermal decompositions) follow a Poisson distribution. Previously, it had been shown that the canonical ratings follow the same Poisson distributions that the quantum-like events do. It was also shown that the data for one S were consistent with the hypothesis that the rating on any trial was equal to the number of quantum-like events that had occurred and for two other Ss, either one less or two less than this number. A signal detection theory analysis of these canonical ratings is performed, resulting in ROC curves and estimates of d'. In addition, it is shown that the Poisson canonical rating distributions can be approximated by cutoff Gaussian distributions. Hence it is possible to use a probit analysis, which is computationally simple, to calculate the maximum likelihood solutions for all means, standard deviations, d', and b, as well as the standard errors of all these estimates. The rating is shown to be a linear function of the internal decision variable. The internal criteria are all greater than the mean of the noise distribution and they are all separated by steps of equal size. The probit analysis may be used whenever all the individual rating distributions are Gaussian in order to obtain the maximum likelihood estimates and standard errors of all parameters for each Gaussian distribution. Thus, this analysis may be applied to rating experiments, other than the one described here.

In the theory of signal detection, it is usually assumed that the underlying signal and noise distributions are Gaussian, and d′ is defined as the difference between the means of the signal and noise distributions divided by the standard deviation of the noise distribution. For Gaussian distributions, if one plots the ROC curve on double probability paper, then d′ will be equal to the horizontal intercept (Green & Swets, 1966).

In practice, one doesn't know if the underlying distributions are Gaussian, but one can still plot the ROC curve on double probability paper. If it appears to be a straight line, then the data are consistent with the Gaussian hypothesis and one usually estimates the parameters of the assumed Gaussian distributions from the horizontal intercept and slope of the ROC curve. However, it must be remembered that a straight line on double normal paper does not prove that both the signal and noise distributions are Gaussian. Nachmias and Kocher (1970) plotted ROC curves based on theoretical Poisson distributions and found that they appeared to be fitted by straight lines on double normal paper. This was true even for Poisson distributions with relatively small means, where each individual distribution is not that well approximated by a Gaussian. Thus, the shape of the ROC curve is not a sensitive test for the underlying distributions.

I would like to consider the application of ROC curves to a process whose underlying distributions are known to be Poisson rather than Gaussian. For example, suppose that signal detection theory was applied to an ideal photometer. This photometer would have some noise since its photosensitive substance would have some thermal decompositions and dark current. It would be able to count every photon absorption plus noise event without being able to distinguish between the two. Instead of giving a confidence rating, the photometer would give the number of electrons emitted (current) over a certain fixed time interval. In the absence of any external stimulus, there will be a distribution in the number of electrons emitted (noise distribution). In the presence of a stimulus, there will be a different distribution in the number of electrons emitted in the same time interval (signal distribution). In order to obtain the ROC curve, one plots the probability that i or more

*This research was supported by Grant EY 00276 from the National Eye Institute, U.S. Public Health Service. It is a pleasure to thank Horace Barlow, David Krantz, R. Duncan Luce, and Jacob Nachmias for many helpful criticisms and comments during the preparation of this manuscript.

electrons were emitted when the stimulus was presented (hit rate) vs the probability that i or more electronics were emitted when only noise was present (false alarm rate). Each value of i produces one point on the ROC curve.

Because of the quantum fluctuations of the light stimulus, the number of quantal absorptions will follow a Poisson distribution. Most probably the noise events will also be Poisson distributed (if the noise is due to a very large number of independent events like thermal decompositions of molecules of photosensitive substance). So even for an ideal photometer than can count every quantum absorption plus noise event, there will be variability in the response due to the variability of the stimulus itself. If the variability of the stimulus is sufficient to account for the variability of the response, then one concludes that the photometer being tested is perfect in the sense of counting every event.

If the photometer counted every event (quantum absorption or thermal noise event) making no distinction between kinds of events, then one could say that it is giving canonical counts because it is counting every event that it is possible to do, given the physical limitations for even ideal performance. A photometer could do a lot worse than canonical counting. For example, it could count only in bunches of 1,000 quanta. Or it could turn itself off intermittently and randomly guess during these periods. The possibilities for error are myriad.

Instead of a photometer, let us consider a human O. Using conditions of maximum sensitivity of the human eye, Hecht, Shlaer, and Pirenne (1942) demonstrated that humans can see when relatively few quanta are absorbed. Their most sensitive O had a criterion of five quantal absorptions in order to "see." They showed that the variability in seeing could be completely accounted for by the known quantum fluctuations of light. Thus the O with a criterion of five absorptions had a frequency of seeing curve in the shape of a cumulative Poisson with a criterion of five.

The experiments to be discussed in this paper were done on human Os under conditions similar to those of Hecht, Shlaer, and Pirenne, using the concepts and techniques of signal detection developed and investigated by Tanner and Swets (1954), Barlow (1956), and Nachmias and Kocher (1970). Some analysis of these experiments has been previously reported (Sakitt, 1972). It was shown there that an O can, in the best case, detect a single quantum of light and that subjectively distinct visual sensations result when 1, 2, 3, ..., etc., quantum events occur. Furthermore, the probabilities of these distinct sensations occurring followed cumulative Poisson curves with criteria of 1, 2, 3, ..., etc. Thus human Os acted as ideal photometers, as discussed above. In order to get these results, Os were asked to use canonical ratings, whereby a separate rating must be used for every subjectively distinct visual sensation so that there is a one-to-one correspondence between ratings and sensations. The concept of canonical ratings assumes that the possible sensations are discrete, not continuous, as the stimuli were. The fact that the probabilities of the different sensations occurring were equal to the probabilities of different numbers of quantum events occurring lends support to this assumption. The canonical ratings are discussed in detail under Methods.

It is the purpose of this paper to analyze the data of Sakitt (1972) at absolute visual detection from the point of view of signal detection theory. First it will be assumed that the underlying signal and noise distributions are Gaussian. All ROC curves will be drawn, and d′ and the slope, b, will be extracted on the basis of the usual Gaussian assumptions. The motivation for this is to push the Gaussian hypothesis as far as one can and then reconcile it to the Poisson distributions of quantum events that were previously obtained.

A new type of analysis of rating data is introduced which is based on canonical ratings. The individual rating distributions are analyzed with a probit analysis. This leads to more information than ROC curves provide. Also, a relatively simple computation can be used to get the maximum likelihood solutions and standard errors for d′ and for the parameters of each signal and noise distribution. This analysis will work under circumstances where the rating distributions can be approximated by Gaussians.

METHODS

The data used for the present paper were previously analyzed in a non-SDT context (Sakitt, 1972). Since the methods are described in that paper, I will only briefly describe the essentials here.

Stimulus

The stimulus was a 29-min disk located about 7 deg in the temporal retina of the left eye. It was a 16-msec blue-green flash (470-520 nm) corresponding to either 66 or 55 photons, on the average, at the cornea, hereafter called the strong and weak stimuli. Os dark-adapted for about 1 h before each experiment, and no background illumination was used at all. In addition to the strong and weak stimuli were blank trials which corresponded to no light.

Apparatus

The experiments were done on a Maxwellian view optical system illustrated and described elsewhere (Sakitt, 1971, 1972). A wheel containing two neutral density filters was placed in the beam which allowed the E to set either the strong or the weak stimulus. In the target plane (conjugate to the retina) was a wheel which permitted the E to insert either a 29-min disk or an opaque stop. Both wheels were moved for every trial, even if they had to be returned to their original positions. They were very quiet and could not be heard above the sounds of fans of power supplies in the room. A shutter was placed in the beam which was electronically controlled by a switch held by the O and opened within milliseconds of the release of the switch.

Observers

None of the Os needed corrective lenses. The author (B.S.) and a hired O (L.F.) were used as the basic Os. They were both very

experienced in visual psychophysical experiments. A third hired O, K.D., who was relatively inexperienced in vision experiments, was used in the preliminary runs but was unable to remain an O for reasons unrelated to the experiment. During the preliminary run with K.D., the wheels controlling the target and intensity were noisy but the noises were uniform and did not seem to give any cues to the O. These auditory noises were eliminated before the runs using the other two Os. The data for all three Os were used for this study. A fourth O gave erratic data and was not used further.

Calibration

Briefly, an SEI meter was used which was calibrated against a standard source whose calibration was traceable to the National Bureau of Standards. All filters were calibrated with either a Perkin-Elmer spectrophotometer or with a Gamma photometer. The absolute values for the average intensities of the strong and weak stimuli were estimated to be scotopically equivalent to 66 and 55 photons of wavelength 507 nm incident at the cornea per flash. The ratio of the two intensities is known with more accuracy than the absolute values. It was estimated to be 1.20 within a few percent. The blank stimulus corresponded to zero quanta at the cornea.

Procedure

An experimental run consisted of a block of 160 trials composed of 80 blank trials, 40 strong stimuli and 40 weak stimuli. The O knew these a priori probabilities, but these trials were presented in a random order. The order was selected before the experiment according to random-number tables. The E looked at the instructions, moved both wheels as described above, and then signaled "okay" to the O. The O released the shutter only after carefully fixating. Brief breaks were taken occasionally during the run. After a block was finished, the O rested for 5-10 min and another block was run. Two blocks were done ach day.

Canonical Ratings

Confidence ratings have been used in signal detection experiments by Pollack and Decker (1958) in speech communication, by Egan, Schulman, and Greenberg (1959) in audition, by Swets, Tanner, and Birdsall (1961) in vision, and by many other workers. It is a well-known and well-used procedure by now. Although the instructions vary from one experiment to another, the usual practice is for the E, not the O, to determine the number of categories. Also, different ratings are meant to represent different subjective estimates of the confidence that the signal was presented.

Canonical ratings differ from the usual confidence ratings because there is a one-to-one correspondence between canonical ratings and subjectively distinct sensations. The system of canonical ratings used in the present study is new and is the key to the entire experiment. It is a system that takes practice to use successfully. Hence, it will be described in detail.

First of all, Os attempted to rate their sensory impressions only in the area where the stimulus was anticipated and only at the time that they opened the shutter. After rating each trial, the O was told what the trial was—strong, weak, or blank. This was done during all the practice sessions as well as during the actual experimental sessions.

During the practice sessions, the task of each O, including the author, was to consciously think about one's own process of rating stimulus and noise trials. We tried to develop as many categories of sensory impressions as possible that we felt we could reliably use. After much practice, we assigned numerical ratings to these different visual sensations. The ratings are called "canonical ratings" because *each rating corresponds to one, and only one, visual sensation.* No rating was used for more than one sensation and no sensation was given more than one rating. Different ratings corresponded to subjectively distinct sensations.

It is difficult to verbalize the quality of these sensations, especially those that correspond to the lowest ratings. At first, the Os were instructed to use verbal descriptions for the different trials. A typical O would begin with only the three classes of "nothing there," "not sure if I saw anything," and "a little light." However, the initial task was to maximize the number of subjectively distinct categories. Using feedback and hundreds of trials over many days, Os felt able to make further distinctions. Typical descriptions then became "a tiny pinpoint of light," or "a vague feeling," or "a sort of feeling of something," etc.

In order to shorten the time for each trial, the verbal descriptions were eventually shortened and standardized to not seen, very doubtful,slightly doubtful, dim, moderate, bright, and very bright. Then these categories were replaced by the numerical ratings of 0, 1, 2, 3, 4, 5, and 6, respectively. The shorthand words were probably not the best choices, but it should be remembered that they are, indeed, only shorthand for the original verbal descriptions which were themselves only feeble attempts to use ordinary English to describe visual sensations that are not ordinarily described. Nevertheless, it would probably have been better to have used a different personalized verbal shorthand for each O and to have encouraged Os to keep refining their categories until they felt they had reached their limits.

Two Os (B.S. and L.F.) had a large number of practice sessions (thousands of trials), and data were taken when a final stable rating system was developed. Another O (K.D.) had few practice sessions (for reasons unrelated to the study) but seemed to give stable ratings. A fourth O (C.L.) had a rating system that was so unstable it was not included in the final data, although the other three Os who were used all seemed able to give stable ratings from day to day. Data were taken for this experiment on 5 days. However, in another rating study, in which B.S. and L.F. participated (to be discussed elsewhere), it was found that they continued to use the same rating systems.

Notation

The usual notation in SDT is to use the letters S and N to refer to the signal and noise distributions. This is inconvenient here because there are three stimulus conditions corresponding to 0, 55, and 66 photons, on the average, at the cornea per flash. The first stimulus condition will be called the blank or noise condition since it corresponds to no external input. The other two stimuli will be called the weak and strong stimuli, respectively. The letters S, W, and B will be used to denote the strong, weak, and blank conditions.

$\overline{P}(i \mid S)$ = probability of saying i or greater when the strong stimulus is presented.

$\overline{P}(i \mid W)$ = probability of saying i or greater when the weak stimulus is presented.

$\overline{P}(i \mid B)$ = probability of saying i or greater when the blank stimulus is presented.

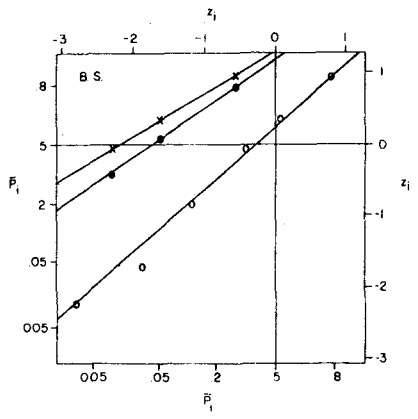
For each O, three ROC curves are obtained by plotting these cumulative probabilities on double probability paper: $\overline{P}(i \mid S)$ vs $\overline{P}(i \mid B)$, $\overline{P}(i \mid W)$ vs $\overline{P}(i \mid B)$, and $\overline{P}(i \mid S)$ vs $\overline{P}(i \mid W)$. For each curve, the different points are obtained from the different values of i.

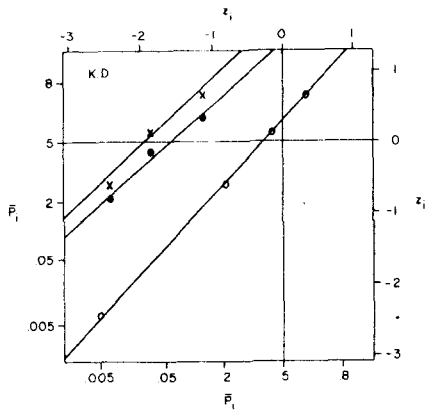
When referring to any ROC curve, the symbols S-B, W-B, and S-W will be used to designate which conditions are being compared. For example, b(S - B) is the slope of the ROC curve obtained from plotting $\overline{P}(i \mid S)$ vs $\overline{P}(i \mid B)$ on double-probability paper. It will also be convenient to use $\overline{P}_i$ and z_i as the cumulative probability of saying i or more and the normal deviate of $\overline{P}_i$ without specifying a particular stimulus condition.

RESULTS AND ANALYSIS

ROC Curves

The data for each O consisted of 400 responses for both the strong and weak stimuli and 800 responses





for the blank stimulus. The cumulative probabilities, $\overline{P}(i \mid S)$, $\overline{P}(i \mid W)$, and $\overline{P}(i \mid B)$, for giving a rating of i or more for the strong, weak, and blank stimuli were calculated for each O and are plotted against each other as ROC curves on normal-normal graph paper in Fig. 1. Straight lines were fitted by eye to the data and d′, and the slope b was extracted from each ROC curve in the usual manner (see Green & Swets, 1966). That is, it was assumed that the underlying distributions were Gaussian so that d′ is the horizontal intercept. Table 1 gives these results under the heading "ROC curves."

The results in Table 1 are fairly typical of ROC

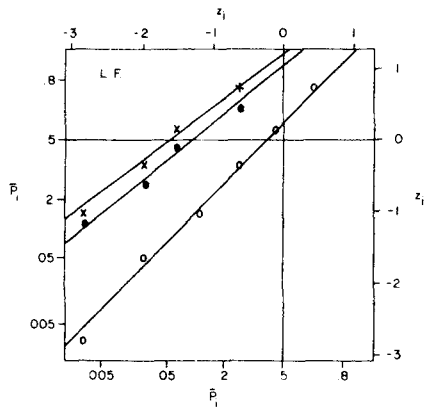


Fig. 1. ROC curves on double-normal paper. Experimental points are X for strong vs blank, ● for weak vs blank, and O for strong vs weak.

curves obtained in many previous studies. The ratio of d′(S-B) to d′(W-B) was between 1.24 and 1.26 for all Os, whereas the ratio of the intensities of the strong and weak stimuli was 1.20. The standard errors of d′ are too large to determine, whether or not d′ was linear with intensity, which would not be typical.

The results from these ROC curves are consistent with the usual Gaussian hypothesis. The hypothesis will be pushed further in the next section.

The Internal Decision Variable

The advantage of using ROC curves is that they give some information about the signal and noise distributions even though the critical sensory and decision variables are not known. In a certain sense, the ROC curve eliminates the most interesting information: the unknown internal decision variable, x. Although it is useful because usually one does not know what x is, it is often useless in trying to discover the underlying physiological mechanisms that produce it.

The purpose of this section is to analyze the individual rating distributions. Figure 2 consists of some unusual curves. For each stimulus (strong, weak, and blank), the cumulative probabilities $\overline{P}_i$ for

Table 1
Values of d′ and b as Obtained from ROC Curves of Figure 1 and as Obtained from the Probit Regression Analysis of Figure 2

	ROC Curves			Probit Analysis		
	B. S.	L. F.	K. D.	B. S.	L. F.	K. D.
$d'(S-B) = [\mu(S) - \mu(B)]/\sigma(B)$	.16	1.61	1.93	2.11 ± .16	1.57 ± .13	2.03 ± .27
$b(S-B) = \sigma(B)/\sigma(S)$	.60	.76	.96	0.67 ± .05	0.84 ± .06	0.71 ± .10
$d'(W-B) = [\mu(W) - \mu(B)]/\sigma(B)$	1.71	1.28	1.56	1.74 ± .14	1.25 ± .12	1.57 ± .23
$b(W-B) = \sigma(B)/\sigma(W)$	.68	.81	.88	0.70 ± .05	0.80 ± .06	0.83 ± .11
$d'(S-W) = [\mu(S) - \mu(W)]/\sigma(W)$	.26	.23	.27	0.26 ± .05	0.26 ± .05	0.38 ± .07
$b(S-W) = \sigma(W)/\sigma(S)$	.88	1.04	1.08	0.96 ± .05	1.05 ± .06	0.86 ± .11
$d'(S-B)/d'(W-B)$	1.26	1.26	1.24	1.21 ± .08	1.26 ± .11	1.29 ± .17

Note—Standard errors are given after estimates from the probit analysis.

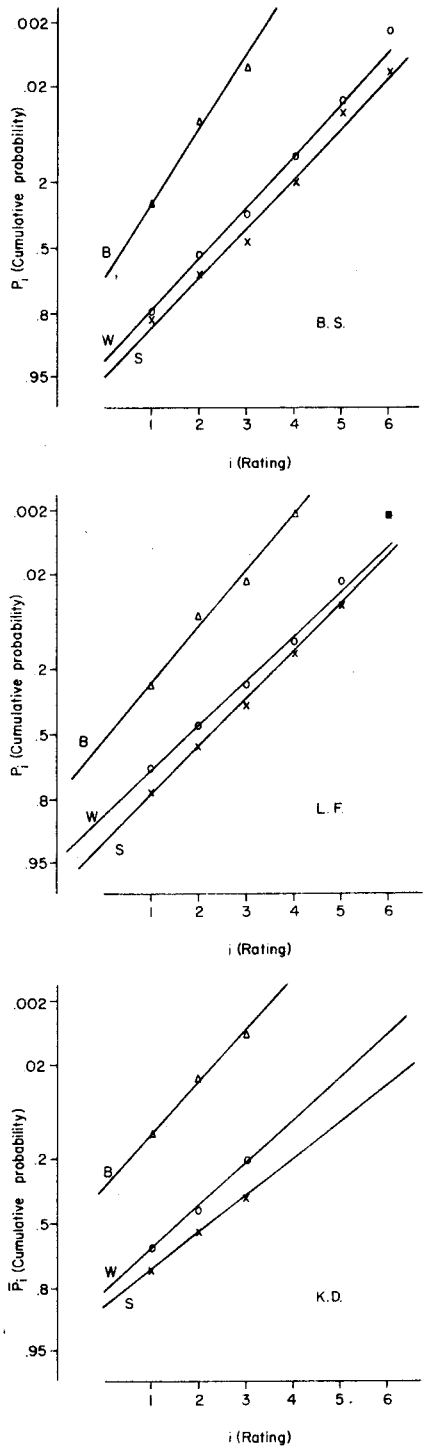


Fig. 2. Cumulative probability vs rating on normal probability paper. Experimental points are △ for blank, ○ for weak, and X for strong. Lines through points are best fitting ones obtained by the probit regression analysis.

saying i or greater are plotted on probit paper against the confidence rating i.

Probit paper transforms a cumulative Gaussian into the straight line

$$Y - 5 = (i - \mu)/\sigma, \tag{1}$$

where i is the independent variable, Y is the probit, μ is the mean of the distribution, and σ is the standard deviation of the distribution. Hence the mean is that value of i that makes $Y = 5$ (same as $\overline{P} = .5$) and σ is the reciprocal of the slope of the line.

It can be seen that the experimental points in Fig. 2 seem to be fitted by straight lines, which means that the underlying strong, weak, and blank distributions are Gaussians. This is a much more stringent test than straight lines on double-normal paper, since ROC curves based on non-Gaussian functions can sometimes be fitted by straight lines on double-normal paper. It should be pointed out that the Gaussian distributions obtained here are dependent on the use of canonical ratings, as will become more obvious later.

Although one can extract some information by an "eye" fit in Fig. 2, there exists an objective method of analyzing such graphs. This straightforward technique, called probit analysis, is described by Finney (1964) in the first four chapters of his book. The advantages of a probit analysis are: (1) it gives an objective statistical test of how well the points are fitted by Gaussians; (2) it provides the best fitting regression line which determines the maximum likelihood solutions for the estimates of the mean and standard deviation of the Gaussian distribution; and (3) it enables one to calculate the standard errors of both these estimates.

The probit analysis of the data of Fig. 2 was done with the aid of a 6400 CDC computer. However, a computer is not necessary. In the appendix of Finney's (1964) book, he shows a detailed numerical example of a complete probit analysis using a desk calculator. The results of the probit analysis, rounded off to three significant figures, are given in Table 2. The results include the best estimates of the mean and standard deviation of each distribution with the standard errors of these estimates. From the values of chi squared, it is seen that Gaussian distributions give a moderate fit to the data. It should be noted here that, when doing a probit analysis, one can drop categories that have less than five expected occurrences. Finney points out that if one uses such classes, the chi-squared distribution may give misleading results. Eliminating such a class does not necessarily "improve" the fit, since it means that chi squared has less degrees of freedom. The number of degrees of freedom, N, in Table 2 is two less than the maximum rating used. K.D. never used the ratings 5 and 6 and had so few occurrences of the

Table 2
Results from the Probit Analysis Where Y is the Probit and i is the Independent Variable

		Probit Line Y =	$\mu \pm S(\mu)$	$\sigma \pm S(\sigma)$	χ^2 (N)	p
B. S.	Strong	3.34 + 0.64 i	2.62 ± .051	1.57 ± .057	14.2(4)	.1%- 1%
	Weak	3.52 + 0.66 i	2.23 ± .052	1.51 ± .058	4.7(4)	30%-50%
	Blank	4.61 + 0.95 i	0.41 ± .075	1.05 ± .066	4.5(1)	2%- 5%
L. F.	Strong	3.61 + 0.61 i	2.28 ± .056	1.64 ± .063	5.0(4)	10%-30%
	Weak	3.94 + 0.58 i	1.83 ± .065	1.73 ± .074	4.6(4)	30%-50%
	Blank	4.92 + 0.73 i	0.11 ± .105	1.38 ± .085	6.5(2)	2%- 5%
K. D.	Strong	3.88 + 0.48 i	2.30 ± .081	2.07 ± .197	0.1(1)	70%-90%
	Weak	4.09 + 0.56 i	1.63 ± .073	1.78 ± .150	1.5(1)	10%-30%
	Blank	5.46 + 0.68 i	−0.677 ± .228	1.47 ± .151	0.6(1)	30%-50%

Note—For each rating distribution, μ and σ are the estimates of the mean (corresponds to mean P = .50) and the standard deviation; $S(\mu)$ and $S(\sigma)$ are the standard errors of these estimates. The last two columns give the chi-squared test for N degrees of freedom.

rating 4 that only rating categories up to 3 were used. However, B.S. and L.F. used all the ratings up to 6, and no points were dropped for them.

Figure 3 shows the best fitting Gaussian distributions obtained for the weak and blank distributions for O B.S. The abscissa is the variable i and is a continuous extension of the ratings i. The probit regression analysis indicated a moderately good fit of the experimental points to the Gaussians shown in Fig. 3. According to signal detection theory, the rating of i is given when x is between the i^{th} and $(i + 1)^{st}$ criterion. Therefore, the abscissae of Fig. 3 can also be marked with the values of x. The rating 0 will be given when x is less than the lowest criterion x_1; the rating 1 will be given when x is between x_1 and x_2; the rating 2 will be given when x is between x_2 and x_3; etc. The ratings i must be a linear function of the internal criteria, x_i, since the only difference between i and x on Fig. 3 can be a scale change and/or a horizontal shift. Therefore, since the ratings are equally spaced apart, the internal criteria must also be equally spaced apart, beginning with x_1.

It is not necessary to draw all the distributions for all the Os in order to make this point. The straight lines in Fig. 2 and the probit analysis indicate that the rating distributions for each of the three stimulus conditions are Gaussians.

According to SDT, the criteria can be chosen arbitrarily. From Fig. 3 and Table 2, it is seen that, for all Os, *all the criteria were above the mean of the noise distribution and all the criteria were equally spaced apart.* It is difficult to believe that all Os chose equally spaced criteria from a continuum of possibilities. The more likely hypothesis is that the decision variable itself is discrete. A very likely candidate would be the number of action potentials in a critical neuron, or the number of discrete bursts of action potentials. Perhaps the rating i is given when k + ni pulses occur in the critical neuron, k and n being integers.

It is important to remember that the Os used canonical ratings (see Methods section). Each canonical rating described one and only one visual sensation. This was a difficult task and took much practice. This is not the usual procedure in rating experiments where Os may, in principle, use categories they feel they cannot always subjectively distinguish, or use a single category for more than one subjective sensation. For example, in the present study, it would have been possible to combine the

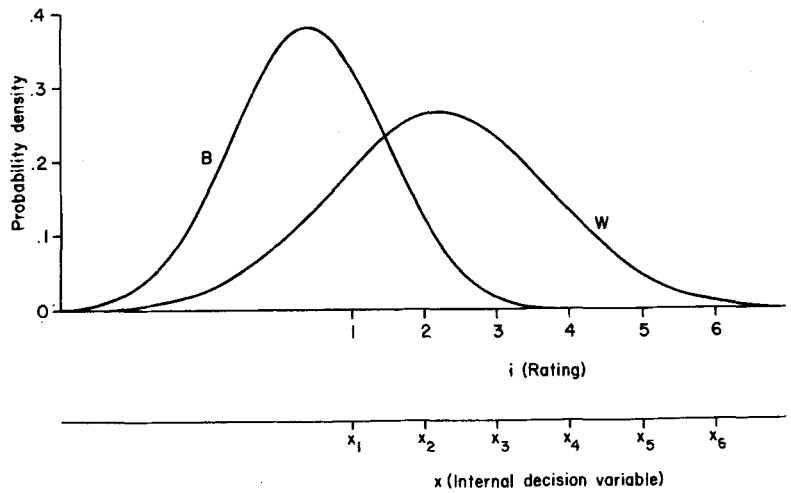


Fig. 3. Best fitting Gaussians for the blank and weak distributions for the B.S. data obtained from the probit analysis. The means are $\mu(B) = 0.41$ and $\mu(W) = 2.23$. The standard deviations are $\sigma(B) = 1.05$ and $\sigma(B) = 1.05$ and $\sigma(W) = 1.51$.

categories corresponding to the ratings 2 and 3 into one category even though these were subjectively different sensations. If an O did that, the curves in Fig. 2 would *not* be straight lines and the above analysis would be inappropriate. But the motivation in the present study is to see the best that Os may do and compare that behavior to an ideal device.

Each individual rating distribution could be moderately well fit to a Gaussian when canonical ratings were used. This seems to lend support to the Gaussian hypothesis. However, it is more accurate to say that the rating distributions were fit to the ordinates of Gaussians at equally spaced abscissae. Furthermore, as seen in Fig. 3, the ratings are quite asymmetrical with respect to the means of the underlying Gaussian distributions.

Maximum Likelihood Estimates and Standard Errors of d′ and b

In the previous section, it was shown how the maximum likelihood estimates and standard errors of all the means and standard deviations could be calculated. The analysis was dependent upon the equal spacing of the experimental points on probit paper. From these estimates, it is possible to calculate d′ and b, as well as their standard errors. An example will illustrate how the quantities were calculated. From the probit analysis, the estimates of the means of the weak and blank distributions for L.F. were 1.833 and 0.111, respectively. The estimate of the standard deviation of the blank distribution was 1.38. Hence the estimate of d′(W-B) is [1.833 - 0.111]/1.38 = 1.722/1.38 = 1.25. Similarly, all other quantities were calculated and tabulated in Table 1 under the "probit analysis" heading. It is seen that the values of d′ and b, as calculated from the probit analaysis, agree with those estimated from the ROC curves. This lends support to the argument that the ratings were linear with the internal decision variable, x.

The probit analysis yielded not only the best estimates of the mean, μ, and the standard deviation, σ, for each distribution, but also the standard error of the mean, $S(\mu)$, and the standard error of the reciprocal of the standard deviation, $S(1/\sigma)$, for each distribution. The 95% confidence limits (or any other limits) can be calculated exactly for μ and $1/\sigma$. Since d′ and b are derived from μ and $1/\sigma$, it is possible to calculate the 95% confidence limits for d′ and b for each distribution. This calculation can be done exactly, although numerically. However, when a random variable has a narrow distribution (if the variance is small compared to the square of the mean) and has a moderately well-behaved distribution (no peculiar discontinuity, etc.), then one can derive analytic expressions for the standard error of functions of this random variable. This procedure was employed here. The derivations and calculations are given in the appendix, and the values obtained for the standard errors of d′ and b are given in Table 1.

The analysis described here is only applicable whenever all the individual rating distributions are Gaussian. In that case, the probit analysis could be applied to each individual distribution, and then from those results, the best estimates and standard errors of d′ and b can be calculated.

Recently, Fortran programs have been developed for finding the maximum likelihood estimates of parameters of ROC curves generated by rating data (Dorfman & Alf, 1969; Dorfman, Beavers, & Saslow, 1973). Their method has the advantage of being quite general. It is not necessary that the individual rating distributions be Gaussians. However, where the analysis of the present paper can be used, it has the advantage of being computationally simple, requiring only a desk calculator.

The Physical Interpretation of the Ratings

Up to this point, the data have been shown to be consistent with the Gaussian hypothesis, The ROC curves in Fig. 1 are straight lines. Even the individual canonical rating distributions are straight lines on probit paper, as seen in Fig. 2.

Yet, in the introduction it was mentioned that a previous analysis of this data (Sakitt, 1972) fitted the rating distributions approximately to Poissons. The same data for B.S. as used in this paper were consistent with the hypothesis that the rating on each trial was the number of effective quantum absorptions plus noise events. For Os L.F. and K.D., the same data as used in this paper were shown to be consistent with the hypothesis that the rating on each trial was either one less or two less than the number of effective quantum absorptions plus noise events. Both the ratings and quantum events followed the same Poisson distributions.

In 1956, Barlow suggested that, in the dark, thermal decompositions of rhodopsin molecules, or other noise events, mimic the action of quantal absorptions. Therefore, the total number of quantum-like events would be those due to quantal absorptions plus those due to noise events. If there are an average of Q quanta incident at the cornea per flash, only some of the incident quanta will be transmitted to the retina, be absorbed by rhodopsin, and result in an isomerization of a rhodopsin molecule which is necessary in order to produce a rod signal. If a is the average number of quantum-like events (or rod signals) and f is that fraction of the quanta incident at the cornea which result in a rod signal, then

$$a = fQ + y \qquad (2)$$

where y is the average number of noise events occurring over the retinal region being considered

during a certain time period. The number of effective quantum absorptions follows a Poisson distribution because of the quantum fluctuations of the light itself. It is probable that the noise events also do so. Therefore, the number of quantum-like events probably has a Poisson distribution, with the average number determined by Eq. 2.

Suppose that the rating i is given whenever c or more quantum-like events (or rod signals) occur (effective quantum absorptions plus noise events). Then the probability of giving a rating of i or greater when a is the average number of rod signals is equal to $\bar{P}(c,a)$, the cumulative Poisson probability that c or more rod signals occur when a is the average number occurring. Thus, using Eq. 2, for each value of i, one gets

$$\bar{P}(i \mid S) = \bar{P}(c, 66\,f + y) \quad (3)$$

$$\bar{P}(i \mid W) = \bar{P}(c, 55\,f + y) \quad (4)$$

$$\bar{P}(i \mid B) = \bar{P}(c,y) \quad (5)$$

Using the B.S. data for the ratings 1, 2, 3, 4, and 5 produced 15 equations which could be approximately solved with the *same* value of f (.0274), the *same* value of y (0.36), and with the value of c always being the rating i. Thus the 15 experimental points were fitted to successive Poisson curves with only two parameters, f and y. Furthermore, the estimate of .0274 found for f, the fraction of quanta at the cornea that are effective in producing a rod signal, is consistent with previous estimates made by workers using independent physical methods. In a similar manner, it was found that the data of L.F. and K.D. were consistent with counting either one less or two less than the number of rod signals. The reader who is interested in the details of the quantum counting analysis is referred to Sakitt, 1972.

Figure 4 shows the experimental cumulative probabilities of giving a rating of i or greater, the

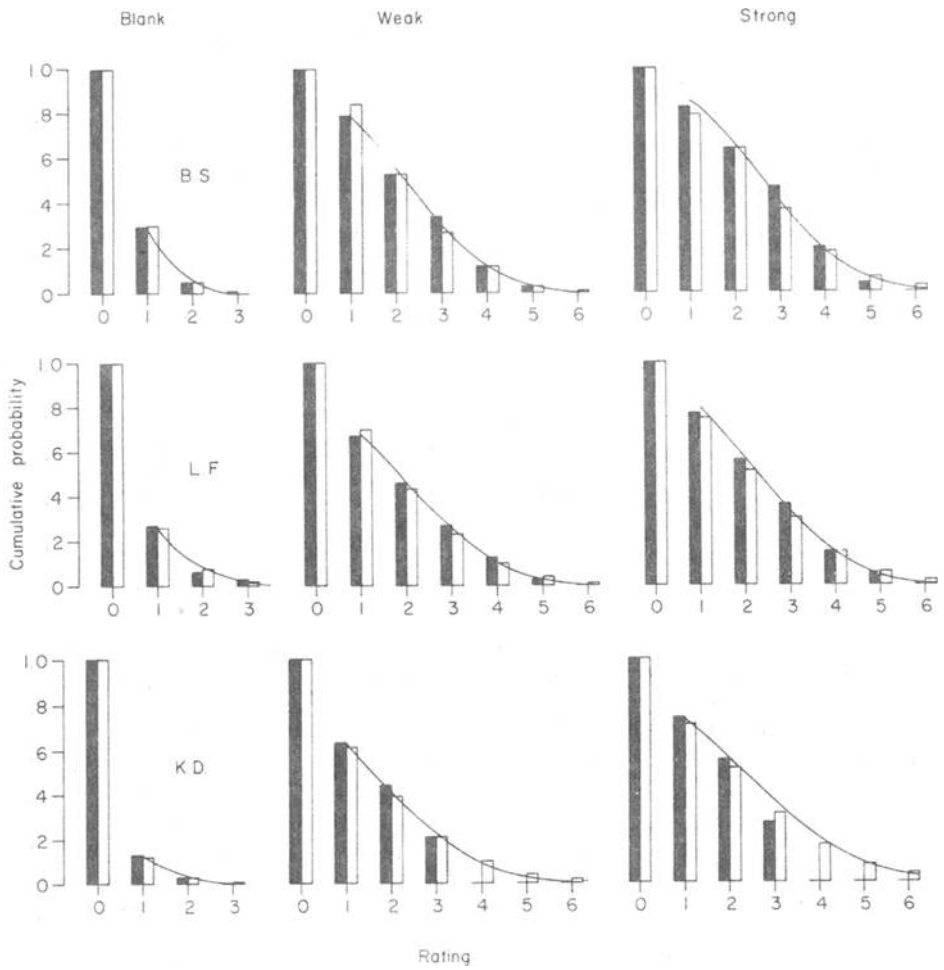


Fig. 4. Cumulative probabilities for giving the rating i or greater vs the rating i. The black bars are the experimental values and are slightly offset to the left for clarity. The white bars are the predictions from the quantum counting hypothesis and are slightly offset to the right for clarity. The continuous curve is the prediction from the Gaussians obtained by the probit analysis.

predictions of these probabilities from the Gaussian hypothesis, and the quantum counting predictions, all on the same graph. The Gaussian predictions for the cumulative probabilities are from the probit analysis of an earlier section of this paper, and the moderate goodness of their fit can be seen by the chi-squared test listed in Table 2. The quantum counting predictions are the cumulative Poisson probabilities that i or more rod signals occur as described above and in Sakitt (1972). The quantum counting predictions have a poor chi-squared fit ($p < .01$). That means that the departure of the experimental probabilities from quantum counting is greater than would be expected by purely sampling error. In addition to quantum counting, an O might be inattentive on some trials and just guess on those occasions, or an O might have a bias against giving the rating 1 and not always use it when it might be appropriate. It would not require a large amount of such bias to account for the small discrepancy between the quantum counting predictions and the actual experimental points, as can be seen by looking at Fig. 4.

Both the Gaussian and Poisson distributions give plausible fits to the data, with the Gaussians giving a better fit. However, the Poisson distributions are based on fewer parameters. But more important, the Poisson distributions are based on the physical stimulus, not an abstract model. The Poisson distributions resulted from the hypothesis that each distinct visual sensation was a result of a different number of quantum-like events and that canonical ratings report all visual sensations.

Thus, a human O can almost act as an ideal photometer, counting the number of quantum-like events that occurred. Although one knows the stimulus has a Poisson distribution, it is theoretically possible that the transduction process is such that a Gaussian represents the distribution of the decision variable. Theother possibility is that the Poissons derived from the quantum counting model are the true sensory distributions but that inattention and bias cause deviations of the data.

In the latter case, the Gaussian distributions can still be used as convenient approximations. But the Gaussian distribution which best approximates a Poisson distribution is *not* obtained by using the mean of the Poisson but by doing the probit analysis, except when the mean is large. When the Poisson to be approximated by a Gaussian has a small mean, the mean of the best fitting Gaussian will be appreciably different from the mean of the Poisson but be closer to the median of the original Poisson.

This can be best seen by an example. Consider a Poisson distribution of mean 1.70. From Poisson tables, the median of this distribution is 2.0. In fact, for a Poisson of mean 1.70, the probabilities of 1, 2, 3, 4, and 5 or more events occurring are .82, .51, .24, .09, and .03, respectively. If one takes a Gaussian of mean and variance equal to 1.70, the cumulative probabilities are .71, .41, .16, .04, and .01, respectively—not a very good match. However, if one takes a Gaussian of mean and variance equal to 2.00, the cumulative probabilities are .76, .50, .24, .08, and .02—a much better fit. Of course, the best Gaussian is not found by using the median but by doing the probit analysis, although intuitively one can see that the median will be more important than the mean since the horizontal intercept is defined in terms of a 50% response.

For O B.S., the best fitting Poissons to the data had means of 2.19, 1.82, and .36 for the strong, weak, and blank distributions. The best fitting Gaussians had means of 2.62, 2.23, and 0.41. Figure 4 indicates that these Poissons and Gaussians are fairly well matched (white bars vs smooth curves), although their means all differ. Similar results occurred for all Os.

Canonical Ratings

Using the canonical rating system, each sensation corresponded to some fixed number of quantum-like corresponded to some fixed number of quantum-like events. Os could therefore reliably distinguish, for example, between three quantum-like events occurring as opposed to four such events, etc.

Barlow, Levick, and Yoon (1971) recorded from individual ganglion cells in the cat retina and found that the information about the number of quantum-like events was preserved, essentially completely. By using different numbers of impulses as criteria, they obtained results on the cat that are very similar to the human results obtained from canonical ratings. (The reader who is interested in the detailed comparisons is referred to Sakitt, 1972).

In summary, canonical ratings produced by subjectively distinct visual sensations in humans produce data similar to those obtained by using different numbers of impulses in an individual ganglion cell of the cat. The simplest explanation of the human data from canonical ratings is that each quantum-like event produced one action potential in a critical neuron and that the canonical ratings were a shorthand for the subjectively distinct visual sensations generated by different numbers of action potentials occurring in this neuron over a certain time period.

There will be some conditions under which the distribution of impulses will not even be approximately Poisson. Barlow and Levick (1969) have shown that in the cat ganglion cell, the distribution of number of impulses follows a Gaussian at moderately high light levels but looks like the human distribution for blanks in the dark. They show that the mean divided by the variance of the distribution, for an on-center unit, increases from about one in the dark (like a Poisson) to higher values when the luminance level is increased. Hence, at high

luminance levels, a Poisson will never fit the data.

In the human, the number of impulses must be a highly compressive function of the number of quantum absorptions in order to see over the enormous range that we do. Therefore, the distribution of the number of impulses must be scaled from the distribution of absorptions and cannot be Poisson at high luminances. Since the distributions for single cat ganglion cells are similar to human rating distributions at absolute visual detection, it is possible that the distributions will also be similar at high luminance levels. If that is so, human canonical rating distributions at high luminance levels will be Gaussian, so a probit analysis may be helpful.

CONCLUSION

The concept of canonical ratings has been introduced where each rating corresponds to one and only one subjective sensation. The present study has shown how canonical ratings can be used to find the underlying neural distributions at absolute visual detection. Hopefully, this technique can be applied to other visual conditions and to other sensory modalities.

When the canonical rating distributions are approximately Gaussian, it was shown that a simple probit analysis can determine the maximum likelihood estimates and the standard errors of all parameters. There is some reason to believe that this statistical analysis can be applied to other canonical rating experiments.

APPENDIX

In order to calculate the standard errors, it is necessary to use the following relationships:

$$V(y) = [S(y)]^2, \tag{6}$$

where V(y) is the variance of the random variable y and S(y) is the standard error of the random variable y, and it is assumed that the sample size, n, is large enough so that n - 1 is approximated by n. If x and y are random independent variables (for example, the means of the weak and blank rating distributions), then

$$V[x - y] = V(x) + V(y), \tag{7}$$

as shown in any elementary statistics book. Another useful relation is

$$V[f(y)] = V(y)[f'(y)]^2 + \text{terms of order } [S(y)/\bar{y}]^2, \tag{8}$$

where f(y) is a moderately well-behaved function of y and $\bar{y}$ is the mean value of y. Equation 8 can be derived by expanding y about $\bar{y}$ in a Taylor series.

The random variables that are of interest here are all the differences in means—for example, $[\mu(W) - \mu(B)]$—and the quantities $1/\sigma$ for each rating distribution. The ratio of the standard error to the mean for all these quantities is always less than 0.2, and hence the correction term in Eq. 8 is at most 4% and usually is much less. Therefore, Eq. 8 can be approximated by

$$V[f(y)] = V(y)[f'(y)]^2. \tag{9}$$

In particular, Eq. 9 can be applied to 1/y and ln y to get

$$V(1/y) = V(y)/\bar{y}^4 \tag{10}$$

and

$$V(\ln y) = V(y)/\bar{y}^2. \tag{11}$$

Since the probit analysis yielded $V(1/\sigma)$, Eqs. 6 and 10 were used to calculate $S(\sigma)$, and these values are listed in Table 2.

Three examples will illustrate the method of calculating the standard errors for b and d'. Consider

$$b(W - B) = \sigma(B)/\sigma(W). \tag{12}$$

Take the natural logarithm of both sides and then, taking the variances and using Eq. 7, one gets

$$V[\ln b(W - B)] = V[\ln \sigma(B)] + V[\ln \sigma(W)]. \tag{13}$$

Using Eq. 11 yields

$$\frac{V[b(W - B)]}{[b(W - B)]^2} = \frac{V[\sigma(B)]}{[\sigma(B)]^2} + \frac{V[\sigma(W)]}{[\sigma(W)]^2} \tag{14}$$

From Eq. 6, the variance is approximated by the square of the standard error. Hence all the quantities in Eq. 10 have already been calculated except V[b(W - B)]. For example, using Tables 1 and 2 and doing the calculation for the L.F. data yields

$$\frac{V[b(W - B)]}{(.80)^2} = \frac{(.085)^2}{(1.38)^2} + \frac{(.074)^2}{(1.73)^2}$$

$$= .005590 = (.075)^2$$

Therefore, S[b(W - B)] = (.80)(.075) = .060. In Table 1, under the heading "probit analysis," there appears the entry 0.80 ± .060 as the estimate of b(W - B) plus or minus S[b(W - B)], the standard error of b(W - B) for L.F. In a similar manner, the standard errors for all the estimates of b were calculated and tabulated.

Next consider another example.

$$d'(W - B) = [\mu(W) - \mu(B)]/\sigma(B). \tag{15}$$

Take the natural logarithm of both sides and use Eqs. 7 and 11 to get

$$\frac{V[d'(W - B)]}{[d'(W - B)]^2} = \frac{V[\mu(W)] + V[\mu(B)]}{[\mu(W) - \mu(B)]^2} + \frac{V[\sigma(B)]}{[\sigma(B)]^2} \tag{16}$$

Since the variance is just the square of the standard error, all the quantities in Eq. 16 have already been calculated, except for V(d'(W - B)], which can now be done. Looking up all the values in Tables 1 and 2 for the L.F. results yields S[d'(W - B)] = .118. This is listed for L.F. under the probit analysis heading after the estimate of d'(W - B) in the form 1.25 ± .12. In a similar manner, the standard errors of all other estimates of d' were calculated and tabulated. The other quantity of interest is the ratio, R,

$$R = d'(S - B)/d'(W - B) = [\mu(S) - \mu(B)]/[\mu(W) - \mu(B)] \tag{17}$$

Taking the natural logarithm of both sides and using Eqs. 7 and 11 yields

$$\frac{V(R)}{R^2} = \frac{V[\mu(S)] + V[\mu(B)]}{[\mu(S) - \mu(B)]^2} + \frac{V[\mu(W)] + V[\mu(B)]}{[\mu(W) - \mu(B)]^2} \quad (18)$$

All the quantities in Eq. 16 have been previously calculated except for V(R). Looking up all the values for L.F. in Tables 1 and 2 yields S(R) = 0.113. This is tabulated in the last row of Table 1 as 1.26 ± .11, which is of the form R ± S(R).

REFERENCES

BARLOW, H. B. Retinal noise and absolute threshold. *Journal of the Optical Society of America*, 1956, **46**, 634-639.

BARLOW, H. B., & LEVICK, W. R. Changes in the maintained discharge with adaptation level in the cat retina. *Journal of Physiology*, 1969, **202**, 699-718.

BARLOW, H. B., LEVICK, W. R., & YOON, M. Responses to single quanta of light in retinal ganglion cells of the cat. *Vision Research*, 1971, **11**, Suppl. No. 3, 87-101.

DORFMAN, D. D., & ALF, E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of Mathematical Psychology*, 1969, **6**, 487-496.

DORFMAN, D. D., BEAVERS, L. L., & SASLOW, C. Estimation of signal-detection theory parameters from rating-method data—a comparison of the method of scoring and direct search. *Bulletin of the Psychonomic Society*, 1973, **1**, 207-208.

EGAN, J. P., SCHULMAN, A. I., & GREENBERG, G. Z. Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America*, 1959, **31**, 768-773.

FINNEY, D. J. *Probit analysis*. Cambridge, England: Cambridge University Press, 1964.

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

HECHT, S., SHLAER, S., & PIRENNE, M. H. Energy, quanta and vision. *Journal of General Physiology*, 1942, **25**, 819-840.

NACHMIAS, J., & KOCHER, E. C. Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America*, 1970, **60**, 382-389.

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

SAKITT, B. Configuration dependence of scotopic spatial summation. *Journal of Physiology*, 1971, **216**, 513-529.

SAKITT, B. Counting every quantum. *Journal of Physiology*, 1972, **223**, 131-150.

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

TANNER, W. P., JR., & SWETS, J. A. A decision-making theory of visual detection. *Psychological Review*, 1954, **61**, 401-409.

(Received for publication March 22, 1974;
accepted July 9, 1974.)