Perception & Psychophysics
1981, 30 (2), 169-174

An optimum value for detection probability

HARVEY DILLON
National Acoustic Laboratories, Sydney 2000, Australia

In many psychoacoustic experiments, an experimenter wishes to determine the effects of one or more treatments (to the subjects, or to the stimuli, or to both) with as little experimental labor as possible. The experimenter is often able to control the similarity of the stimuli and is thus free to choose a base level of performance before the treatment is applied. On the basis of the model presented in this paper, the base level of performance that best enables any change in performance due to the treatment to be detected is predicted. A relatively high base-level performance of 95% correct discrimination results from the model, although practical constraints may dictate a somewhat lower value. An experiment (involving pitch discrimination) that provides data supporting the model is also reported. The model is applicable only to the two-alternative/forced-choice experimental paradigm.

Often, in forced-choice experiments, it is desirable to know if a given treatment of either the stimuli or the subjects has any effect on the ease with which the subjects discriminate between the two stimuli. This change in discriminability will appear as a change in the probability of a correct response after the treatment is applied. The problem addressed here is that of finding the optimum probability of discrimination in order that any change in probability due to the treatment is most easily detected. This will be accomplished in two stages. First, the discrimination probability that maximizes the change in probability will be calculated. Second, the criterion will be changed by taking into consideration the test statistic that is used for testing the significance in changes in discrimination probability. The base discrimination probability that leads to the most significant change in probability for a given number of replications or subjects will then be found. Intuitively, one would not make the task "too easy" or "too hard" since any change in discriminability may not be noticed, that is, the subjects may continue to respond at the 50% or 100% level.

The model applies only when one wishes to detect a change in performance in a two-alternative/forced-choice paradigm. If one wishes to detect changes in the parameters of a psychometric function arising from a yes-no task, suitable strategies are given by Bush (1963).

The work outlined in this paper was performed while the author was at the University of New South Wales, Australia. The author wishes to thank D. McNicol and W. H. Holmes for the valuable suggestions made during the preparation of the paper. Requests for reprints should be sent to National Acoustic Laboratories, 5 Hickson Road, Millers Point, Sydney 2000, Australia.

THE MODEL

Conventional Thurstonian scaling theory is assumed (e.g., Torgerson, 1958). The equations will be developed with reference to the two-alternative/forced-choice task. It is postulated that the effect of the treatment is to alter either the variance or the mean, or both, of the Gaussian probability density function f(x) of each stimulus on the psychological scale. Figure 1a shows how the two stimuli, A and B, map onto the scale before and after the treatment (cases 1 and 2, respectively). Since the subject is deciding which of the two stimuli, A or B, he perceives as being the greater, we are interested in the difference distribution, $g(x_B - x_A)$, which is shown in Figure 1b. The relationships between the parameters of this distribution and the original are as follows:

$$\mu_1 = \mu_{B_1} - \mu_{A_1} \qquad \sigma_1^2 = \sigma_{A_1}^2 + \sigma_{B_1}^2$$

$$\mu_2 = \mu_{B_2} - \mu_{A_2} \qquad \sigma_2^2 = \sigma_{A_2}^2 + \sigma_{B_2}^2.$$

The hatched area represents the probability of a correct response. These distributions are then normalized to a zero mean and unity standard deviation, and are shown redrawn in Figure 1c as $F_1(x)$ and $F_2(x)$. Finally, these distributions are superimposed in Figure 1d. The change in probability, ΔP, that is to be maximized is shown as the doubly hatched region. In the limit of small changes in μ and σ, that is, from μ_1 to μ_2 and from σ_1 to σ_2,

$$\Delta P = \left(\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}\right) \cdot \left[F\left(\frac{\mu_1}{\sigma_1}\right) + F\left(\frac{\mu_2}{\sigma_2}\right)\right] \Big/ 2$$

$$\approx \left(\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}\right) \cdot F\left(\frac{\mu_1}{\sigma_1}\right). \qquad (1)$$

0031-5117/81/060169-06$00.85/0

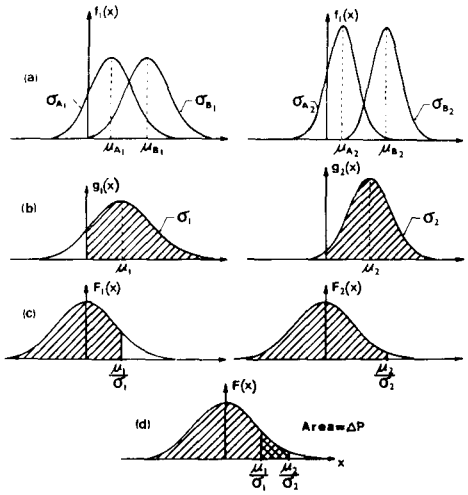

Figure 1. (a) Distribution of two stimuli, A and B, on the subjective scale before and after treatment (suffixes 1 and 2, respectively). (b) Distributions obtained by subtracting the distributions of the original two stimuli. (c) Normalized version of the distributions in (b). (d) The two graphs in (c) superimposed to show the change in discrimination probability, ΔP.

As a result of the treatment, the mean and standard deviation of the difference distribution may change by a small amount.

Putting $\mu_2 = (1+k) \cdot \mu_1$ and $\sigma_2 = (1+k') \cdot \sigma_1$, then

$$\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1} = \frac{\mu_2\sigma_1 - \mu_1\sigma_2}{\sigma_2\sigma_1}$$

$$= \frac{(1+k)\mu_1\sigma_1 - (1+k')\mu_1\sigma_1}{(1+k')\sigma_1^2}$$

$$= \frac{\mu_1(k-k')}{(1+k')\sigma_1}$$

$$\approx (k-k')\frac{\mu_1}{\sigma_1}. \qquad (2)$$

The term $(k-k')$ in Equation 2 represents the fractional changes made to the standard deviation and mean. For any given changes in σ and μ, we wish to find the maximum ΔP as the probability P (or equivalently, μ_1/σ_1) varies. Substituting Equation 2 into Equation 1:

$$\Delta P = (k-k')\frac{\mu_1}{\sigma_1} F\left(\frac{\mu_1}{\sigma_1}\right).$$

Thus, the quantity to be maximized is $x \cdot F(x)$, where $x = \mu_1/\sigma_1$. Differentiating ΔP with respect to x and setting the derivative equal to zero,

$$\frac{d(\Delta P)}{dx} = (k-k')\frac{d}{dx}[x \cdot F(x)] = 0.$$

Substituting

$$F(x) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$$

(the normalized Gaussian distribution) then gives

$$\frac{\Delta\sigma}{\sigma_1} \cdot \frac{1}{\sqrt{2\pi}}\left[\exp\left(\frac{-x^2}{2}\right) - x^2\exp\left(\frac{-x^2}{2}\right)\right] = 0.$$

$$\therefore x^2 = 1$$

$$\therefore x = \pm 1.$$

Since x is the ratio of mean to standard deviation in the difference distribution, a value of unity implies that the maximum change in discrimination probability will occur when the two stimuli map into regions spaced apart by one standard deviation. This corresponds to a discrimination probability of 84%. As the assumption regarding the manner of change of the variances and means will be later justified, the only assumption remaining is that the difference distribution is Gaussianly distributed. If the result is to be of any practical use, then it is important that the optimum percentage obtained be not too strong a function of the distribution assumed. To check this, the above procedure has been repeated for several other initial distributions. The calculations are more complex, since the initial distributions of the two stimuli must be convolved with each other to obtain the difference distribution, so only the initial distributions and the final discrimination probabilities that maximize ΔP are shown in Figure 2. For this wide range of possible distributions, the discrimination

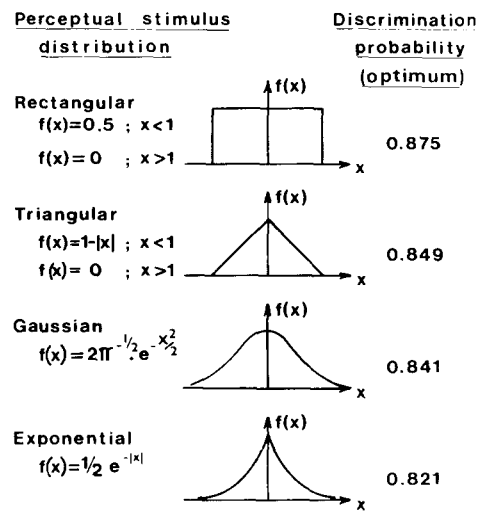

Figure 2. Four possible distributions of the stimulus on the subjective scale and the discrimination probability which maximizes the change in probability due to a given treatment.

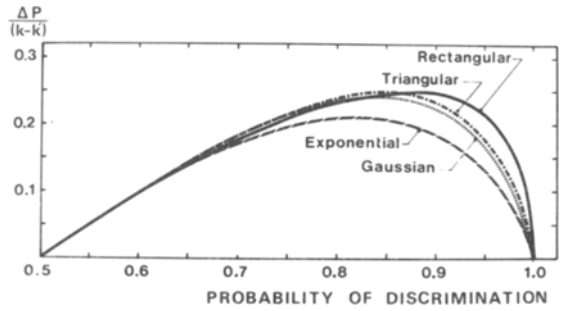

Figure 3. Ratio of change in probability to change made in stimulus distribution vs. the average probability of discrimination for the probability density functions described in Figure 2.

probability which maximizes the change in probability varies only between 82% and 88%.

For a range of distributions, Figure 3 shows the decrease in change of probability observed when other discrimination probabilities are used. The term $\Delta P/(k-k')$ represents the ratio of the change in probability measured to the change made in the stimulus distributions. "Probability of discrimination" refers to the average of the two probabilities measured before and after the treatment.

MODIFIED CRITERION TO ALLOW FOR A MEASUREMENT ERROR

So far it has been shown that the change in discrimination probability will be maximized for initial probabilities of around 84%. However, this change in probability will normally have a test of significance applied to it. If the variance of the test statistic is also a function of discrimination probability, then a new criterion allowing for this will result in an optimum discrimination probability different from the 84% result previously found. For initial and final probabilities P_1 and P_2, and number of observations in each group N_1 and N_2, the standard deviation of the distribution of the difference between the two probabilities is given by:

$$\sigma_D = \sqrt{P \cdot (1-P) \cdot \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}, \qquad (3)$$

where

$$P = \frac{N_1P_1 + N_2P_2}{N_1 + N_2}.$$

In Figure 4, σ_D is shown as a function of average discrimination probability P (for $N_1 = N_2 = 1$). To test for a significant change in the proportion discriminated, one computes the ratio $z = (P_1 - P_2)/\sigma_D$ and assumes that it is normally distributed (Hays, 1969). Thus, the new criterion is that this ratio z is to be maximized for a given treatment and number of observations. It is clear from Figure 4 that higher values of P are to be preferred in order to achieve a lower value of σ_D. The required ratio, $z = \Delta P/[(k-k') \cdot \sigma_D]$ is shown in Figure 5 as a function of P, with $\Delta P/(k-k')$, the normalized change in discrimination probability as previously derived for the Gaussianly distributed case. The peak has now shifted up to the rather high value of 95% discriminability due to the lower value of σ_D in this region. This, then, is the most efficient region in which to conduct tests that investigate the effect of a treatment. The results for the other distributions mentioned previously are similar, with the exception of the rectangular distribution, which shows a peak at 100% discrimination probability.

DISCUSSION OF ASSUMPTIONS

The optimum probability of discrimination is independent of any equal variance assumptions and relatively independent of the mapping distribution assumed. The only assumptions made are that Thurstonian-type scaling theory is an appropriate model, and that if the mean of the difference dis-

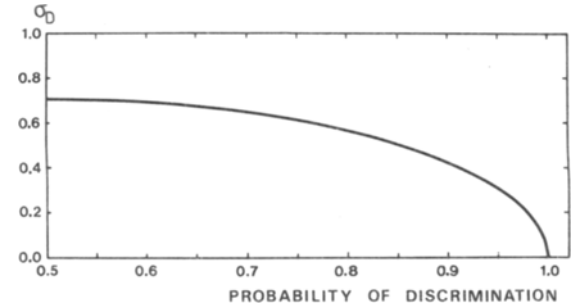

Figure 4. The standard deviation of the difference between two probabilities (normalized to one observation per probability) vs. the average of the two probabilities.

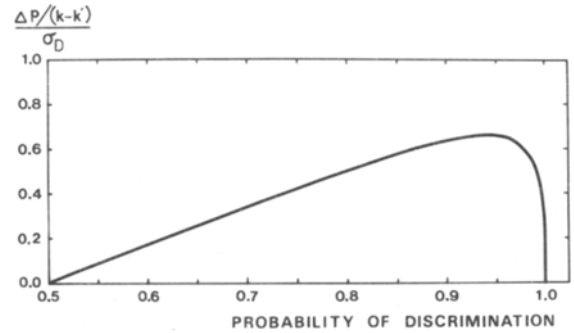

Figure 5. The ratio of the normalized change in discrimination probability to the normalized standard deviation used to test the significance of such a change. ΔP is normalized with respect to the change made to the stimulus distribution (k—k') and with respect to the number of observations.

tribution is affected by the treatment, then any change in this mean is proportional to the initial difference between the means of the individual distributions. If only the variance is affected, then the results are the same as already derived, irrespective of whether the standard deviation is changed by a fixed increment or by a multiplicative constant, as in both cases the change in μ_1/σ_1 is proportional to the initial value of μ_1/σ_1. However, if it is postulated that μ_1 is affected by the treatment, then the assumption that it changes by a fixed increment (independent of μ_1) leads to a result that is different from that which follows from the assumption of a proportionate increase in μ_1.

We will now show that any change in μ_1 will be of the proportionate type if the signals are reasonably close together. This is so because no matter what shape the function relating the scale values after the treatment to the scale values before the treatment is, any small section of it may be considered linear. Since the stimuli used in discrimination tasks are, of necessity, reasonably close together on the psychophysical scale, the following approximation may be used.

Let $\mu' = S(\mu)$, where μ is the scale value of any stimulus before the treatment and S is the function relating μ to the corresponding scale value μ' after the treatment. Using the same terminology as in Figure 1,

$$\mu_{B_2} - \mu_{A_2} \approx (\mu_{B_1} - \mu_{A_1}) \cdot \left.\frac{dS}{d\mu}\right|_{\mu=\mu_{A_1}}$$

that is,

$$\mu_2 \approx \mu_1 \cdot \left.\frac{dS}{d\mu}\right|_{\mu=\mu_{A_1}}$$

$$\mu_2 \propto \mu_1.$$

Thus, the change in the difference between the two means must be approximately proportional to the original difference between the two means. Note that the mean values referred to above are values on a psychophysical scale. They do not refer to the parameter of a decision or attention process that may be changed by experimental manipulations.

PRACTICAL IMPLICATIONS

With very few constraints or assumptions additional to those inherent in scaling theory, it has been shown that, for the most efficient measurement of a change in discriminability of two signals, the probability of discrimination should be in the 93%-96% region. Caution is required, however, before this result is applied to experimental design. The reason for this is that the analysis has been via a small-signal, differential approach, thereby assuming small shifts in discrimination probability. For medium-sized shifts ($\Delta P < .2$), the results will still be appropriate if the range of probabilities involved encompasses the most efficient region. For higher values of ΔP, such gross shifts in probability will be easily detected with little experimental labor and maximally efficient experiments will not be so important.

A further note of warning arises from the ever-present spread of subjects' abilities or biases found in psychophysical experiments. If the average subject were to be operating at around the 95% region, then a high percentage of the subjects would be found to be discriminating perfectly both before and after the treatment has been applied. No information would be obtained from these more highly skilled observers and an inefficient experiment would result. The results may also be biased due to the exclusion of this group of subjects. Notice also, from Figure 5, that the sensitivity of the test falls off extremely rapidly as the proportion correct increases above the optimum value, but only slowly as the proportion decreases below this value. Thus, it would seem wise to aim at a value somewhat lower than the optimum proportion correct.

Thus, a suitable summary of the practical application of this result would be: Provided that only an insignificant number of subjects are responding with near perfect discrimination (either with or without the treatment being present), the effect of a treatment will be most efficiently detected if the stimuli are chosen such that the subjects will respond at as high a level of discrimination as possible.

The procedure is therefore of greatest relevance to those two-alternative/forced-choice designs in which stimulus levels are adjusted for individual subjects before beginning an experiment in which the experimenter wishes to detect changes from this baseline condition.

REFERENCES

Atkinson, R. C. A variable sensitivity theory of signal detection. *Psychological Review*, 1963, 70, 91-106.

Bush, R. R. Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

Hays, W. L. *Statistics*. London: Holt, 1969.

Konig, E. Effect of time on pitch discrimination thresholds under several psychophysical procedures: Comparison with intensity discrimination threshold. *Journal of the Acoustical Society of America*, 1957, 29, 606-612.

Thornton, A. R., & Raffin, M. J. M. Speech discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research*, 1978, 21, 507-518.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

APPENDIX

EXPERIMENTAL VERIFICATION

There are two features of the model which may require experimental verification. First, it may be that the subjects' responses are not truly binomially distributed, so that the variability does not decrease with increasing probability of discrimination in the manner shown in Figure 4. Second, the assumption about the nature of the change in the mean of the distribution when the treatment is applied may not be valid. If this is the case, then the change in probability will not vary with average probability in the manner shown in Figure 3. Thus, the purpose of this experiment is to verify the two separate results that have been presented graphically as Figures 3 and 4.

A complete verification would require the performance of all conceivable psychophysical experiments. This being impractical, only one experiment was performed. The discrimination of small pitch differences was chosen as the test experiment for the simple reason that the necessary equipment was readily available. Different levels of discrimination can be easily obtained by using different frequency separations, and a "treatment" can be readily applied to the stimuli by decreasing the duration of the tones. This is known to cause a decrease in pitch discrimination performance (Konig, 1957). Note that the theoretical model presented earlier did not make use of the known properties of pitch perception or, indeed, of any particular discrimination ability.

Experimental Design

Pairs of successive tones were presented to the subjects. Each pair of tones had one of three different frequency separations and one of two different durations. Thus, six different stimulus pairs were presented to each subject. These represent three levels of difficulty, each presented with or without the treatment (a change of duration). The subjects were instructed to choose whether the first or second tone of the pair had the higher pitch and to press the corresponding button. The sinusoidal tones were generated by an Adret CS 201 synthesizer controlled on line by a Hewlett-Packard 9830 computer. They were gated on with rise and fall times of 10 msec and a duration of either 100 or 300 msec. Beyer DT-48 headphones were used to present the stimuli at 70 dB SPL (measured on a continuous tone). The stimuli were presented as 20 practice trials followed by 11 blocks of 60 trials each, with rest periods between each block. As the results from the first of these blocks were discarded before analysis, 600 trials remained upon which to perform the analysis. This represents 100 replications of each of the six stimulus pairs, with 10 replications of each pair arranged randomly within each block.

The tones used had frequencies of between 400 and 410 Hz; the actual frequency separations were selected to be commensurate with each subject's pitch discrimination abilities as determined in the first (discarded) block of trials.

Eight first-year psychology students were used as subjects. Each subject was tested for 2 h, and feedback was given after each trial in an attempt to maintain accuracy and motivation.

Analysis

A population standard deviation can only be estimated if several scores from that population are available. For this reason, the 100 responses to each stimulus pair were broken up into 10 groups of 10. For each stimulus pair, a group consisted of all responses made to that pair in a particular block. A value for the standard deviation for that stimulus pair can then be estimated from the resulting 10 scores by the usual method:

$$\hat{\sigma} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N-1}},$$

where $N = 10$ (the number of groups) and x_i is the proportion of correct responses for the ith group. If the responses are truly from a binomial distribution, the theoretically expected value will be

$$\sigma = \sqrt{\frac{P(1-P)}{M}},$$

where P equals the probability of correct discrimination and M equals the number of trials upon which that probability is based, in this case 10.

The change in probability, ΔP, is more easily calculated. For each frequency separation, the proportion correct (based upon 100 replications) for the short-duration stimuli is subtracted from the score for the long-duration stimuli.

Results

Since six different stimulus pairs were presented to each of eight subjects, altogether 48 estimates of σ could be made. These estimates correspond to values of probability in the range of .5 to 1.0 and are shown as the crosses in Figure A1.

The expected value of σ, $\sqrt{[P \cdot (1-P)]/M}$, is shown as the curve in Figure A1. Note that this curve has not been in any way "fitted" to the experimental data, but that it shows the trend of the data quite well. Clearly, high discrimination scores are accompanied by relatively low values of variability. There appears to be a slight tendency for the binomial distribution to underestimate the average variability. Two reasons for this are possible. First, any learning or fatigue effects in the experiment will increase the

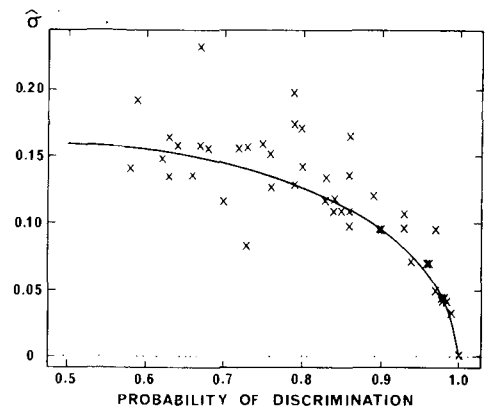

Figure A1. Experimentally determined estimates of the standard deviation of a proportion, $\hat{\sigma}$ (when measured in a two-alternative/forced-choice task), as a function of the experimental estimate of the true proportion. The smooth curve shows the theoretical expected value.

observed score variability. This happens because the scores that were used to estimate the variability were obtained from consecutive blocks of trials, as explained earlier. Second, Atkinson (1963) has shown that sequential dependencies can affect the response probabilities, especially when the subjects are provided with information feedback. Under such conditions, the assumption of independence of events (which is made when applying the binomial distribution) is violated. Although no special techniques (such as trial spacing or counterbalancing) were intentionally used to overcome sequential effects, it is clear that the effect of sequential dependencies on the score variability is not large, in this experiment at least. The appropriateness of the binomial distribution in estimating the variance of a particular score has also been confirmed for speech intelligibility tests (Thornton & Raffin, 1978).

We now turn to the confirmation of the main result of this paper—that a discrimination score of around 84% leads to the greatest change in discrimination when measuring the effect of a treatment. For each of the eight subjects, a score was obtained for both long and short stimuli at each of three degrees of difficulty. Thus, three differences in discrimination (or effects) were obtained for each subject. Evaluation of the data proved difficult for two reasons. First, the change of duration affected the subjects by different amounts. Figure A2 shows the data for the two subjects who were most and least affected by the change of stimulus duration. The change in detection probability is on the ordinate, and the average detection probability is on the abscissa. Each solid curve shown is a third-order least squares fit to the data, but constrained to pass through $\Delta P = 0$ at average probabilities of .5 and .1. (For either perfect discrimination or random performance both before and after the treatment, a change in discrimination of zero must result). Curves such as these were fit to the data of each subject, although coincidence of the curve and data points was not generally as close as in the two cases shown. The peak height of each curve was measured, and each set of raw data was multiplied by a factor such that all subjects had a peak equal to the group average. This transformation thus gave each set of data the same peak sensitivity without affecting the average probability at which that peak occurred. The resulting data and least squares fit (solid curve) are shown in Figure A3.

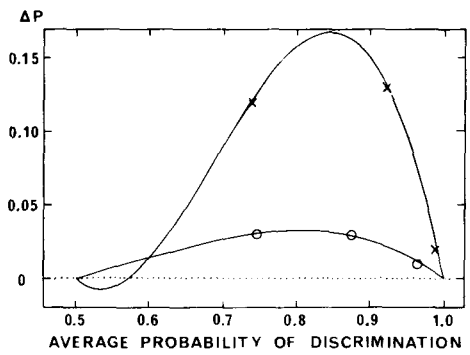

Figure A2. The effect of a treatment vs. the difficulty of the task for two subjects. Values on the ordinate show the increase in observed correct pitch discriminations ΔP, when the duration was changed from 100 to 300 msec. Values on the abscissa show the proportion correct when averaged over these two durations. The smooth curves are third-order least squares fits to the data.

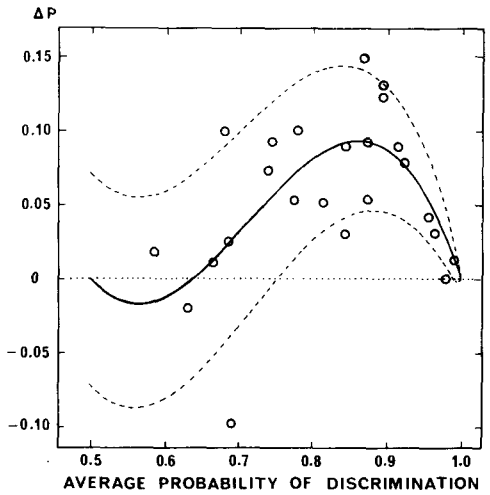

Figure A3. The same as for Figure A2, except that the data is for all eight subjects and has been normalized (see text).

The large scatter in Figure A3 is caused by the second reason why the model evaluation proved difficult. Each point in the figure is based on the difference between two probability estimates, with each estimate based on 100 trials. The expected standard deviation of the experimental points is given by Equation 3, and the ± 1 standard deviation limits are shown by the dotted lines in Figure A3. A closer conformity of the data points to the line of best fit can thus not be expected from this experiment. The measurement error could be halved only by increasing the duration of the experiment from 2 to 8 h per subject. Despite the measurement error, it is evident that larger effects are observed for higher average scores than for lower average scores. If the curve of best fit is taken to represent the data, then the peak effect has occurred at an average score of 86%. Agreement with the theoretical value is very close—much closer than could be expected considering the scatter of experimental points. It may be of interest to note that the eight individual curves rose to a maximum at average percentage scores of 90, 87, 86, 85, 81, and 65. Six of the eight scores at which the maximum effect occurred were thus within three percentage points of the theoretically expected value, although the measurement error prevents drawing a conclusion from any individual curve.

In summary, although the theoretical model did not involve any of the known properties of pitch perception, the experimental data provided support for the model. No systematic deviation from the theory was observed and the random measurement error was of a size commensurate with binomial theory predictions for initial probability estimates based on 100 observations.

The two theoretical relationships (between average probability and effect size, and discrimination probability and data variability) have thus been separately verified. It follows that the composite result, presented in Figure 5, has also been verified (to the extent that it can be verified by any one experiment).

(Manuscript received October 13, 1980;
revision accepted for publication April 28, 1981.)