METHODS & DESIGNS

Maximum-likelihood estimation of the accuracy of auditory discrimination in the absence of an overt response to every stimulus

L. H. BROEKHOVEN, B. BROOKER, M. CZIGLER, and M. W. DONALD
Queen's University, Kingston, Ontario, Canada

In research on human performance in which techniques for estimating hits and false-alarm rates are used in conjunction with other measurements, such as evoked potentials, it is necessary to devise special data-collection methods. This paper describes such data-collection methods, together with a statistical method for summarizing the data. The statistical method consists of setting up a probability model whose parameters describe the process and then estimating the parameters using maximum likelihood.

In research on human discrimination performance, techniques for estimating hit and false-alarm rates have depended upon the availability of some measure of response on every stimulus presentation. The response may take the form of a same-different or yes-no judgment, or it may take the form of a buttonpress or some other motor movement. In this way, it is possible to compare the response on a given trial with the response that should have been made.

There are, however, occasions that demand a measure of performance, but that do not allow an assessment of the response to every stimulus. For example, when evoked potentials are measured during the performance of a discrimination task, a time-locked motor movement or vocalization following every target stimulus could introduce myogenic artifact, as well as other time-locked confounds, into the averaged EEG record. The possibility of such confounds would preclude any form of response by the subject during the recording and averaging periods, and would prevent a trial-by-trial record of performance.

One solution to this problem is to conduct parallel experiments, that is, to collect the behavioral data on one run and the electrophysiological data on another. Another solution is to obtain only a very crude estimate of performance accuracy, such as an overall count of the number of targets presented during a recording session. Neither solution is really adequate: In the first case, the

direct relation between the electrophysiological and behavioral data is not subject to examination, and in the second case, guessing behavior has not been assessed, since guessing cannot be differentiated from actual target detection.

We have attempted a solution to this problem by devising a selective listening paradigm that utilizes the maximum-likelihood estimation technique (see, e.g., Kendal & Stuart, 1967, and Sprott & Kalbfleisch, 1965). The experimental session is broken into brief 10-sec trials, during which a number of targets might be presented. The subject responds to the end of each trial with a number, which is an estimate of the number of targets presented during the trial. Since the subject does not respond during the trial, evoked potentials averaged during a trial are free of contamination. This sequence is repeated for whatever number of trials may be needed. At the end of such a series of trials, the experimenter is presented with a string of numbers, which constitute evidence of performance on a number of trials.

These numbers are, of course, ambiguous, just as a single overall count of targets would have been. Any given number can represent some combination of guesses and actual target detections: For example, if four targets are presented, and the subject responds with the number 4, the number might represent either perfect performance or a lucky guess. Although this holds for a single trial, it is possible to estimate for a number of trials what a given guessing strategy might do to the distribution of responses. By use of the maximum-likelihood estimation technique, it is possible to fit a model to the subject's pattern of performance over a period of time.

This model was applied to the data from an experiment involving dichotic presentation of pure tone "pips" of four different frequencies. The subjects ranged in age from 6 to 40 years. One ear received a stream of

We are indebted to a referee for suggesting the incorporation of the Fletcher method of optimization. This research was supported in part by NSERC Grant A7029 to M. W. Donald. The authors are listed in alphabetical order. B. Brooker and M. W. Donald are with the Department of Psychology, and L. H. Broekhoven and M. Czigler are with the Department of Mathematics & Statistics, all at Queen's University, Kingston, Ontario K7L 3N6, Canada.

Table 1
Typical Response Data From One Subject

Trial Number	Right 1500		Left 800		Left 1500		Right 800	
	Response	Target	Response	Target	Response	Target	Response	Target
1	3	3	2	4	2	3	4	4
2	4	4	3	3	2	4	2	3
3	4	0	2	4	2	0	3	4
4	2	0	0	2	0	0	1	2
5	2	1	1	0	3	1	1	0
6	1	2	1	2	2	2	2	2
7	2	1	3	1	1	1	2	1
8	0	3	0	0	3	3	0	0
9	2	2	2	3	1	2	3	3
10	3	4	2	4	4	4	3	4
11	2	0	4	3	3	0	2	3
12	3	1	0	1	0	1	2	1
13	2	4	3	2	3	4	2	1
14	4	3	1	0	1	3	0	0
15	1	2	0	1	2	2	1	1
16	2	3	2	0	2	3	2	0
17	1	1	1	3	2	1	4	3
18	2	4	2	2	4	4	0	2
19	2	0	4	4	1	0	2	4
20	3	2	1	1	1	2	3	1

Note—Four experimental conditions were presented, requiring attention to tones in the designed space-frequency categories shown. Twenty trials were run in each condition. Actual responses are shown on the left; expected responses (targets) are shown on the right.

800-Hz tone pips ("standards") that contained infrequently occurring 840-Hz tone pips ("rares"). Rares ranged in probability from 5% to 25% and were positioned randomly in the stimulus stream. A similarly constructed stream of higher pitch standard (1500-Hz) and rare (1560-Hz) tones was presented to the other ear. The subject's task was to pay attention to the tones in one ear, while ignoring tones in the other ear, and to count silently the number of rare ("target") tones in the attended ear.

An experimental trial consisted of 10 sec of tone presentations at an overall presentation rate of about 6 tones/sec, followed by 6 sec during which the subject reported the number of target tones he detected. There were 20 trials for each experimental condition, and four experimental conditions corresponding to the allocation of attention to the left or right ear and to the high- or low-pitch tones. An example of data from one subject is given in Table 1.

The problem was to reduce these data to some simple statistics that adequately described them. It was evident from the raw data that sometimes the subjects counted the same number of targets that had been presented, sometimes more, and sometimes less. This indicates that the subjects sometimes missed a target and sometimes invented one. There was even the possibility that some subjects were responding completely at random. The subjects knew roughly the range of the number of targets that might be presented.

THE MODEL

At first, it seemed unlikely that it would be possible to devise a model that would adequately describe

a situation that included both failure to detect the signal and invention of false signals. The model proposed was that each target had a fixed probability of being identified and each standard tone had a different but fixed probability of being identified as a target. Since there were many standard tones and the probability of identifying them as targets was small, a Poisson distribution was used for inventions.

The observed data are (n_i, x_i) for $i = 1, 20$, where n_i is the number of targets presented and x_i is the number identified in each 10-sec interval. Each x_i could consist of zero targets and x_i inventions or one target and $x_i - 1$ inventions, etc., up to r_i targets and $x_i - r_i$ inventions, where r_i is the minimum of x_i and n_i . The probability of observing x_i is therefore:

$$P(x_i | n_i, \pi, \mu) = \sum_{j=0}^{r_i} \binom{n_i}{j} \pi^j (1-\pi)^{n_i-j} e^{-\mu} \frac{\mu^{x_i-j}}{(x_i-j)!},$$

where $r_i = \min(n_i, x_i)$, π is the probability of correctly identifying a target, and μ is the expectation of the number of inventions in a 10-sec interval. The log likelihood function is given by:

$$\text{Log } L(\pi, \mu | n, x) = \sum_i \text{Log } P(x_i | n_i, \pi, \mu).$$

The function is complex. Its maximum can be obtained only by numerical methods. Because it was believed that, given the nature of the data, it might be difficult to estimate the parameters, a plot of the log likelihood function was first obtained for various data

Figure 1. Contours of constant likelihood for data of Table 1.

sets. This is relatively easy to do, since π must lie between 0 and 1 and μ must be greater than 0 and is not likely to be greater than 4. The log likelihood surface is shown in Figure 1 for the four sets of data of Table 1. The log likelihood is a function of the two variables π and μ , since n and x are known. Figure 1 plots this function. Since it is a function of two variables, it is a surface, and the contours of this surface are shown.

The interpretation of a model of this kind is that high values of π combined with low values of μ indicate good signal detection. Specifically, $\pi = 1$ and $\mu = 0$ would result from a perfect match of the target and responses. At the other end of the scale, the value $\pi = 0$ and some value for μ would indicate a perfectly random response unrelated to the targets. Values of the log likelihood function give plausibility regions for the values of the parameters. Specifically, the maximum of the log likelihood gives the most plausible value and the 100 α % confidence regions can be obtained from the asymptotic relationship:

$$\text{Log } L(\pi, \mu | n, x) = \text{Log } L(\hat{\pi}, \hat{\mu} | n, x) - \frac{1}{2} \chi_{2, \alpha}^2,$$

where $\hat{\pi}$ and $\hat{\mu}$ are the maximum likelihood values for π and μ and $\chi_{2, \alpha}^2$ is the 100 α percentile of the chi-square distribution with two degrees of freedom (Cox & Hinkley, 1974, pp. 313-314). The confidence contours represent a composite hypothesis test. Any hypothetical value of π and μ inside the confidence contour would not be rejected by a hypothesis test, and any values outside the contour would be rejected. The 95% contours are shown in Figure 1.

After observing a number of plots such as the ones shown in this paper, it became evident that, in many

cases, there were clearly defined maxima for the log likelihood function, and it was decided to attempt an iterative procedure to calculate the maximum. The Newton-Raphson method was used, with starting values of $\pi = .5$ and $\mu = 1$. This method calculates corrections to the parameter vector by the formula:

$$\begin{pmatrix} \delta\mu \\ \delta\pi \end{pmatrix} = -(\mathbf{H})^{-1}(\mathbf{G}),$$

where $\mathbf{H}$ is the Hessian matrix

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \text{Log } L}{\partial \pi^2} & \frac{\partial^2 \text{Log } L}{\partial \pi \partial \mu} \\ \frac{\partial^2 \text{Log } L}{\partial \pi \partial \mu} & \frac{\partial^2 \text{Log } L}{\partial \mu^2} \end{pmatrix}$$

and $\mathbf{G}$ is the gradient vector

$$\mathbf{G} = \begin{pmatrix} \frac{\partial \text{Log } L}{\partial \pi} \\ \frac{\partial \text{Log } L}{\partial \mu} \end{pmatrix}.$$

The iterations are continued until $\delta\pi$ and $\delta\mu$ are both less than .001 in absolute value. Because the log likelihood function was so complex, numerical derivatives were used. A detailed description of this procedure, including a worked numerical example, is given in Kalbfleisch and Prentice (1980, pp. 55-58).

DISCUSSION

The methods described here have been used on over 100 subjects. It has been found that the estimated π and μ values are helpful in reducing the data from each run to two meaningful statistics. These parameter estimates have been used as response measures in various experiments, which were then analyzed using standard statistical techniques (see Brooker, 1980, Donald, 1983, and Donald & Little, 1981). The examples shown in Figure 1 have been selected from the many that were run, and represent both good and bad likelihood functions. The case with 800 Hz and attending the right earphone shows a relatively sharp maximum, and the 95% confidence region is entirely within the admissible region. In the other cases, the 95% region includes $\pi = 0$, and it is possible that the subject was only guessing. This example was for a relatively young subject (11 years) and was reasonably typical of such cases. Older subjects, and particularly adults, were much more consistent, whereas the youngest subjects tended to be erratic.

It was found that, for adult subjects, the Newton-

Raphson procedure converged quickly. Estimates of the confidence limits for the parameters could also be obtained, since the variance of the parameter estimates are given by the diagonals of H^{-1} . For the younger subjects, the iterative method did not always converge, and therefore an estimate for π and μ had to be obtained from the likelihood function. In these cases, the confidence limits tended to be wide, and the data confirmed the poor detection ability of the subject. The technique does assume a fairly constant strategy on the part of the subject; perhaps this is why the π and μ parameters failed to converge regularly in a sample of 6- to 10-year-old children. In a total of over 100 runs with adult subjects, it was possible to obtain convergence in all but two cases, and in those cases π and μ values could be estimated by a grid search on the likelihood function. To overcome this difficulty, we propose to incorporate some other optimization procedure, such as the quasi-Newton procedure by Fletcher (1980), which is in the IMSL library.

Although there have been applications of the maximum-likelihood estimation technique in conventional signal-detection paradigms that demand a yes-no response to every stimulus (Dorfman & Alf, 1968, 1969; Grey & Morgan, 1972; Ogilvie & Greelman, 1968), there have been, to our knowledge, no previous applications of this technique in situations in which responses cannot be made to every stimulus. The relation between the π and μ estimates derived with this technique, and signal-detection terminology for somewhat similar functions, is problematic. The value π corresponds reasonably well to the concept of hit probability, but μ is not precisely equivalent to false-alarm probability. The μ parameter, at least in the context of the auditory task used in this study, includes not only false alarms, but also errors in counting. However, even the concept of "false alarms" is somewhat ambiguous when stimulus intensity is well above threshold, encompassing errors of memory and impulsive motor responses, as well as genuine misperceptions of stimuli. The complementarity of π and μ corresponds to the complementarity of hits and false alarms; in both cases, all positive responses must fall into one or the other category.

CONCLUSION

We conclude that the maximum-likelihood estimation technique can be useful in assessing discrimination performance when a response is not possible to every stimulus. Although the application was developed for use with a high-speed selective-listening task, in principle it can be extended to any paradigm requiring a simple repetitive discrimination, provided that a delayed counting response is possible.

REFERENCES

- BROOKER, B. *The development of selective attention to learning disabled and normal boys*. Unpublished doctoral dissertation, Queen's University, Kingston, Ontario, Canada, 1980.
- COX, D. R., & HINKLEY, D. V. *Theoretical statistics*. London: Chapman and Hall, 1974.
- DONALD, M. W. Neural selectivity in auditory attention: Sketch of a theory. In A. Gaillard & W. Ritter (Eds.), *Tutorials in event-related potential research: Endogenous components*. Amsterdam: North-Holland, 1983.
- DONALD, M. W., & LITTLE, R. The analysis of stimulus probability inside and outside the focus of attention, as reflected by the auditory N_1 and P_3 components. *Canadian Journal of Psychology*, 1981, **35**, 101-113.
- DORFMAN, D. D., & ALF, E. Maximum likelihood estimation of parameters of signal detection theory—A direct solution. *Psychometrika*, 1968, **33**, 17-124.
- 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.
- FLETCHER, R. *Practical methods of optimization: Unconstrained optimization* (Vol. 1). New York: Wiley, 1980.
- GREY, D. R., & MORGAN, B. J. Some aspects of ROC curves—Fitting normal and logistic models. *Journal of Mathematical Psychology*, 1972, **9**, 128-139.
- KALBFLEISCH, J. D., & PRENTICE, R. L. *The statistical analysis of failure time data*. New York: Wiley, 1980.
- KENDAL, M. G., & STUART, A. *The advanced theory of statistics* (Vol. 2). London: Charles Griffin, 1967.
- OGILVIE, J. C., & GREELMAN, C. D. Maximum likelihood estimations of ROC curve parameters. *Journal of Mathematical Psychology*, 1968, **5**, 377-391.
- SPROTT, D. A., & KALBFLEISCH, J. G. Use of the likelihood function in inference. *Psychological Bulletin*, 1965, **64**, 15-22.

(Manuscript received December 30, 1982;
revision accepted for publication July 22, 1983.)