were reinforced when the light was on, and not reinforced when the light was off. Lever B responses in both groups were not reinforced in either stimulus condition.

High discrimination ratios in Fig. 1 mean that the animal seldom presses lever A during $S^{\Delta}$ (when the light is off and lever A responses are not reinforced). The experimental subjects that were reinforced for pressing lever B during $S^{\Delta}$ made fewer lever A responses during this period than did the control subjects (Fig. 1). The experimental group was up to a ratio of .65 within the first 3 days, whereas the control group was still around .50. At the end of this phase the experimental subjects had stabilized at about .90, whereas the control group had stabilized at about .70. These differences were statistically significant ($t = 7.67$, $P < .01$).

The most dramatic results, however, took place in the third phase of the experiment. When reinforcement for the competing behavior was withdrawn, the experimental subjects resumed pressing lever A when the light was off, even though such behavior was still not being reinforced in the presence of this stimulus. The discrimination ratios deteriorated drastically for the experimental subjects and there was only a gradual recovery over the next 15 days. In other words, lever A responses for experimental subjects had not been extinguished during $S^{\Delta}$ conditions in the previous phase; they had been temporarily supplanted by lever B responses. When this competing behavior was no longer reinforced the experimental animals still had to learn what the control animals had already learned. No savings in total number of errors (lever A responses during nonreinforced periods) was indicated. In fact, the trend was in the opposite direction; the mean number of errors made by the experimental subjects in phases 2 and 3 combined was greater than that made by the control subjects (978 compared to 682). This difference, however, was not statistically significant ($t = 1.40$, $P > .05$).

An almost identical finding was obtained in a second experiment where, instead of studying the same phenomenon within the context of discrimination training, we used a simple extinction procedure. Twenty-four male hooded rats maintained at 80 percent of their weight when given free access to food served as subjects. In phase 1 only lever A was available and all subjects were reinforced for pressing this lever on a variable-interval, 30-second food schedule. In phase 2, lever A responses were no longer reinforced and lever B was available. During this phase the experimental subjects ($n = 12$) were reinforced after every tenth lever B response, whereas the control subjects ($n = 12$) underwent the typical extinction procedure, that is, neither lever A nor lever B was reinforced. In phase 3 experimental subjects were no longer reinforced for lever B responses. They were now in the same situation as the control subjects; neither response was reinforced (Fig. 2).

In phase 2 when the extinction procedure for experimental subjects was supplemented by reinforcement of competing behavior, lever A responses declined more rapidly and more substantially than was the case for the control subjects undergoing a conventional extinction procedure ($t = 8.12$, $P < .01$) (Fig. 2). When reinforcement of competing behavior was discontinued in phase 3, however, the experimental subjects resumed pressing lever A and exhibited an extinction curve similar to that of the control subjects in the preceding phase. Again, there was very little overall savings from the previous reinforcement of competing behavior. Although the experimental group responded on lever A less often than the control group in phase 2, they responded significantly more often in phase 3 ($t = 2.97$, $P < .01$). When the two extinction phases are combined, there is no significant difference in the number of lever A responses made by the two groups ($t = 1.42$, $P > .05$).

These findings correspond to the reports of Skinner (*8*) and Estes (*9*). They found that mild punishment did not hasten the course of extinction; likewise our results indicate that extinction may not be hastened by the reinforcement of competing behavior. In fact both punishment and reinforcement of competing behavior suppress the behavior to be extinguished and thus may prevent extinction from taking place. When punishment is stopped or reinforcement of competing behavior is stopped, the extinction procedure still needs to be carried out. This is not true, however, in the case of more intense punishment (*10*) and it may not be true if competing behavior is reinforced for a longer period before being terminated, or if a different schedule of reinforcement for the competing behavior is used, or if reinforcement for competing behavior is discontinued more gradually, or if the competing response has a topology different from the response being extinguished.

HAROLD LEITENBERG
RICHARD A. RAWSON
KENT BATH

*Department of Psychology,*
*University of Vermont, Burlington*

**References and Notes**

1. G. H. Kimble, *Hillgard and Marquis' Conditioning and Learning* (Appleton-Century-Crofts, New York, 1961).
2. A. C. Catania, in *Operant Behavior: Areas of Research and Application*, W. K. Honig, Ed. (Appleton-Century-Crofts, New York, 1966), p. 213; A. C. Catania, *J. Exp. Anal. Behav.* **12**, 731 (1969).
3. J. A. Dinsmoor, *Psychol. Rev.* **62**, 96 (1955).
4. A. Amsel, *Psychol. Bull.* **55**, 102 (1958); ———, *Psychol. Rev.* **69**, 306 (1962).
5. J. Wolpe, *Psychotherapy by Reciprocal Inhibition* (Stanford Univ. Press, Palo Alto, Calif., 1958).
6. J. W. M. Whiting and O. H. Mowrer, *J. Comp. Psychol.* **36**, 229 (1943); W. C. Holz, N. H. Azrin, T. Ayllon, *J. Exp. Anal. Behav.* **6**, 407 (1963); R. L. Herman and N. H. Azrin, *ibid.* **7**, 185 (1964); E. O. Timmons, *J. Gen. Psychol.* **67**, 155 (1962).
7. E. E. Boe, *Can. J. Psychol.* **18**, 328 (1964).
8. B. F. Skinner, *The Behavior of Organisms* (Appleton-Century, New York, 1938).
9. W. K. Estes, *Psychol. Monogr.* **57**, No. 3 (1944).
10. E. E. Boe and R. M. Church, *J. Comp. Physiol. Psychol.* **63**, 486 (1967).
11. Supported in part by PHS grant MH 12222. We thank G. Bertsch and R. Coughlin for their help in planning this study.

1 June 1970

## Means and Variances of Average-Response Wave Forms

John *et al.* (*1*) reported that, during generalized responding, the stimulus for such generalization releases a neural process representing the previous experience of the animal—that is, the engram associated with the original conditioning. I am not questioning this conclusion—in fact it would appear to be a likely hypothesis. However, I do question certain aspects of the methodology employed in reaching the conclusion.

My first point concerns the comparison of average evoked responses to the generalization stimulus; evoked responses to generalization stimuli eliciting the instrumental response for food were compared to evoked responses to generalization stimuli eliciting avoidance. A number of *t*-tests (the number was not specified) were used to compare amplitudes at similar latency points of individual average-response wave forms. The variance estimates for these *t*-tests

were variances calculated from the raw electroencephalographic (EEG) input used to obtain the average responses. Such variance estimates are inappropriate. Average-response procedures are used in recognition of the fact that such signals are imbedded in a background of noise; they are an attempt to improve the ratio of the signal to the noise. Variance estimates calculated on input data probably are best interpreted as simply indicating the relative amounts of noise in the recordings. Generally such "noise" consists of the normal spontaneous EEG activity of the subject, but the example where the noise also includes artifacts such as 60-hz interference should be considered. As an example, suppose average responses are recorded in two conditions. In one case the subject is ungrounded or inadequately shielded from 60-hz interference, and in the other such is not the case. In the former case, it may still be possible to obtain a reasonable average-response wave form. Mean amplitudes of such a wave form may correspond very well with the means obtained in the better recording condition, but such will not be the case for the variance computations. Therefore, it would be more reasonable to treat individual average wave forms as single determinations. Also in recording human evoked responses the individual responses are often undetectable; computation of the variance of such an input seems meaningless. Thus, the basic generalization tests of John *et al.* should have been repeated a sufficient number of times to estimate the variance of the averages. In other words, in carrying out average-response computations, the basic datum ought to be the individual average-response wave form rather than the individual EEG trace.

It might appear that the above criticism is carping and that the procedure employed by John *et al.* was in fact a conservative test of their hypothesis. The logic of such an argument would be based on the fact that variances were estimated from signal plus noise rather than signal alone. As such they should be larger than variances of the signals alone. Thus, the standard errors for the *t*-tests would be expected to be inflated, and any statistical significance would actually be underestimated. Such an argument has its immediate appeal but is seen to be fallacious once it is realized that variability within a single average-response wave form implies nothing about the variance of several such wave forms. It is possible that averages computed at two different times may show small internal variance but relatively large differences in amplitude. I know of no studies on this point and, in the absence of such studies, it would be unwise to assume these two variance estimates are related in any specific manner.

To make the above argument more meaningful, I would like to illustrate it from the data of John *et al.* (compare figures 1 and 2). The electrode sites and recording conditions are comparable in the two figures (average responses for cats 2 to 5), but there are a number of instances where there are marked discrepancies between average-response wave forms computed from different subsamples of exactly the same data. This variability of the average-response wave forms is the relevant variability to be considered in these studies. It is especially prominent if the averages for $V_3CAR$ (generalized response to frequency 3) and $V_2CAR$ (correct response to frequency 2) for cat 3 are compared between figures 1 and 2.

Finally, two other points must be considered. (i) John *et al.* were troubled by the fact that ". . . wave shapes which were obviously different both by visual inspection and by *t*-test yielded high correlation coefficients." Two wave forms may be different in amplitude by *t*-tests at all corresponding points and still yield a correlation of 1.00; they would then be parallel wave forms and have identical shape. (ii) A substantial part of the discussion by John *et al.* is devoted to the fact that selected samples of their data conform to their hypothesis even better than do their objectively derived computations. I am never surprised by the results that one can obtain by selecting data—that selection is possible, however, is attributable to the variability possible in average-response wave forms.

MARVIN SCHWARTZ
*Department of Psychology,*
*University of Cincinnati,*
*Cincinnati, Ohio 45221*

**Reference**

1. E. R. John, M. Shimokochi, F. Bartlett, *Science* **164**, 1534 (1969).

9 October 1969; revised 19 February 1970

Schwartz (*1*) makes a number of comments about the usage of averaging, variance measures, *t*-tests, and correlation coefficients in the analysis of evoked potentials (EP's), none of which, as he states, has any bearing on our conclusions (*2*). Nevertheless, his comments show a commonly held, potentially misleading attitude. This attitude consists of attributing the variance of an average evoked potential (AEP) only to noise—that is, variation caused by an array of unknown factors whose effects are usually assumed to be random and not time-locked to the stimulus.

Electroencephalographic (EEG) activity may not be homogeneous (*3*), and recently it has become apparent that there can be nonhomogeneities in collections of EP's (*4*). Perhaps the need for significant computer resources has hindered the development of methods for dealing with nonhomogeneities in collections of EP's. However, perhaps it is also true that experimenters have not always recognized nonhomogeneities in their data—averages give no indication, and standard deviations give, if any, an equivocal indication.

When initially examining the data obtained in the generalization experiments, we often found on visual inspection what appeared to be nonhomogeneities in the EP sequences recorded during decision-making trials. In some structures, this took the form of rather abrupt changes in the shape of the EP's; in other structures, it was apparent that a harmonic of the stimulus frequency appeared or was lost.

At that time, the available methods for computer detection of nonhomogeneities and sorting of EP's into appropriate classes were inadequate for our data. We resorted to the ad hoc methods described in our paper. It was from this attempt at pattern recognition by visual inspection that our figure 2 was derived. This figure consists not of selected averages, but of averages of homogeneous (to visual inspection) sets of EP's selected from every available trial of the appropriate type during the day's recording session. Since we were aware of the possible pitfalls of introducing any data selection procedure, no matter how objective the rules, averages of the data were also taken from a fixed interval before the animal's lever press (our figure 1). For some animals more than others, it appeared that these latter averages were sets of nonhomogeneous EP's; hence the sometimes striking differences between our figures 1 and 2.

The *t*-test results illustrated in our figures 1 and 2 were computed on the digitized EP's (500 samples per second), providing a separate assessment for time-points every 2 msec along the analysis epoch. The means and variances obtained at the corresponding latencies were used for these computations, rather than the variance of the raw EEG as

erroneously inferred by Schwartz. Inspection of the figures shows that the percentage of these $t$-tests which indicated $P < .01$ varied from 10 to 70 percent, depending on the cat. Thus, certain components of the AEP obtained when a neutral stimulus elicited generalized approach differed significantly from those obtained when presentation of the same stimulus resulted in generalized avoidance.

Ruchkin (*5*) has written a computer-sorting program that can take a collection of EP's and compute amplitude histograms at selected time points. A chi-squared measure of the approximation of each amplitude histogram to a normal distribution is computed, and a mode count is made (the probability of data drawn from a population with a unimodal, normal density function falsely indicating a multimodality was computed). With the assumption that the noise is either (i) normally distributed or (ii) distributed with one mode, we now have statistical criteria for establishing nonhomogeneity. A boundary can be defined between two or more modes in a histogram. The computer will sort the EP's into their respective classes and calculate the AEP and standard deviation for each class.

This program has been applied to the data that we previously reported. It typically finds what we found by visual inspection, but works faster and can detect nonhomogeneities that are not readily apparent to visual inspection. For example, the data of cat No. 3 were found to be bimodal for each of the four types of stimulus-decision trails. Computerized sorting was performed on those EP time points which were indicated to be maximally nonhomogeneous by the above method. The probabilities that the amplitude distributions at these latencies were derived from populations distributed normally ranged from $P < .03$ to $P << .0001$.

With such methods, where statistics are computed for a number of points, these statistics should be used as indicators. The points may not be independent, making it difficult to state the true overall significance of a computed statistic. Other facts—such as internal consistency, reproducibility, presence of correlated components, and ordering of the types of EP's during the decision-making trials—have a bearing on the significance of the indicated nonhomogeneities. When such observations can be quantified, the significance of the nonhomogeneity often becomes very high. For example, runs tests computed on the four bodies of data from cat No. 3 give $Z$ scores ranging from $-3.09$ to $-10.07$ ($P = .001$ to $<< .0001$). Thus the sequence of the different types of EP wave shapes in the trials was not randomly distributed. In the long run, the final justification for our approach must be its heuristic value in clarifying phenomena present in the EP.

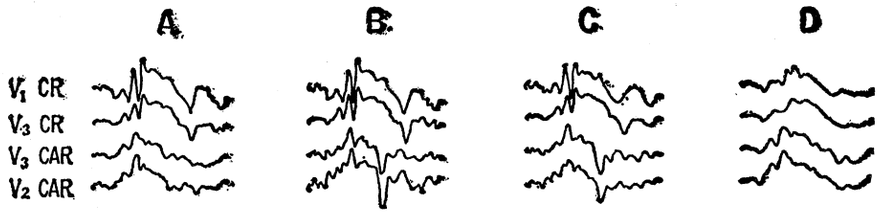


Fig. 1. Average evoked response wave shapes (124-msec epoch) recorded from the left lateral geniculate body (bipolar derivation, two electrodes 1 mm apart) of cat No. 3. The evoked potentials were recorded during trials resulting in correct performance of a lever-press for food during presentation of flicker at frequency 1 ($V_1CR$), correct performance of a lever-press to avoid shock during presentation of flicker at frequency 2 ($V_2CAR$), and during presentations of flicker at an intermediate frequency 3 resulting in generalized approach ($V_3CR$) or avoidance ($V_3CAR$). The averages are from subsets of the data determined as follows: (A) evoked potentials occurring in the arbitrary interval of 4 seconds preceding the animal's lever-press; (B) evoked potentials selected by a visual inspection procedure; (C) one type of evoked potential—primarily occurring late in the decision-making trials—identified and selected by a computer; and (D) the other type evoked potential—primarily occurring throughout the earlier portion of the trials—identified and selected by a computer.

The average of the EP's from the arbitrary time interval beginning 4 seconds before the behavioral response and the average of homogeneous sets of EP's which were selected by our visual inspection procedure are shown in Fig. 1, columns A and B. Comparison of these columns shows an apparent marked attenuation in the "last 4 seconds" averages of the striking components seen in the "selected" averages. The remainder of this figure shows that this was due to the confounding of the two modes of activity present during the trials and not to the data selection process which was suggested by Schwartz. The averages of the two modes of activity (types of wave shape) distinguished by the computer are shown in Fig. 1, columns C and D. It is clear from the similarity of columns B and C that our attempt to be objective in the selection of EP's for averaging was successful. Inspection of column D, the other type of wave shape selected by the computer, indicates that the arguments which were presented concerning neural readout from memory (*2*) are also supported by modes of activity that we neglected at that time. A film showing these phenomena is now available (*6*).

Schwartz's main suggestion, the assessment of the reliability of EP averages by way of multiple replication, is, of course, important where there are neither resources nor criteria available for establishing the validity of averaging. However, reproducibility should not be confused with validity. It is possible to obtain reproducible AEP's from nonhomogeneous data. Our experience indicates that assessment of the validity of particular AEP's can be productive in understanding the data. Reliance on AEP's from arbitrarily set time intervals or of arbitrarily set numbers of EP's can lead to obfuscation of important phenomena.

FRANK BARTLETT
E. ROY JOHN

*Brain Research Laboratories, Department of Psychiatry, New York Medical College, New York 10029*

**References and Notes**

1. M. Schwartz, *Science*, this issue.
2. E. R. John, M. Shimokochi, F. Bartlett, *ibid.* **164**, 1534 (1969).
3. H. Berger, *Arch. Psychiat. Nervenkr.* **87**, 527 (1929).
4. R. Hernandez-Peon, H. Scherrer, M. Jouvet, *Science* **123**, 331 (1956); E. R. John and K. F. Killam, *J. Nerv. Mental Dis.* **131**, 183 (1960); S. K. Burns and R. Melzack, *Electroencephalogr. Clin. Neurophysiol.* **20**, 407 (1966); G. C. Galbraith, *IEEE (Inst. Elec. Electron Eng.) Trans. Bio-Med. Eng.* **14**, 223 (1967); G. R. Gullickson, B. Rosenberg, C. W. Darrow, *Electroencephalogr. Clin. Neurophysiol.* **22**, 188 (1967); D. S. Ruchkin, *Exp. Neurol.* **20**, 275 (1968); E. Donchin, *Electroencephalogr. Clin. Neurophysiol.* **27**, 311 (1969); H. Fruhstorfer and R. M. Bergström, *ibid.*, p. 346.
5. D. S. Ruchkin, in preparation.
6. We have produced a 30-minute teaching film which presents the rationale of the generalization experiment, discusses the results, and shows animals working at the tasks we used, but which consists largely of the actual sequences of EP's obtained during a number of decision-making trials. The individual EP's were photographed directly from an oscilloscope, and all of the phenomena illustrated in Fig. 1 can clearly be seen in the raw data. Interested readers may make arrangements to view this film by contacting us.
7. Supported by PHS grant No. MH-08579. Dr. Shimokochi has returned to Japan and was not available for consultation or comments.

29 April 1970