Evoked Potential Correlates of Psychophysical Judgments: The Threshold Problem. A New Reply to Clark, Butler, and Rosner

DINA D. PAUL

Department of Psychiatry, Columbia University, New York, New York 10032

and

SAMUEL SUTTON

Biometrics Research, New York State Department of Mental Hygiene; and Department of Psychiatry, Columbia University, New York, New York 10032

Additional data are offered bearing on the controversy initiated by the report of Clark, Butler, and Rosner [(1969) J. Comp. Physiol. Psychol. 68, 315-319] of dissociation between psychological responses and evoked potential data when subjects are performing psychophysical tasks in a semi-anesthetized state.

Following a critique by Donchin and Sutton [(1970) Commun. Behav. Biol. 5, 111-114], Clark, Butler, and Rosner [(1970) Commun. Behav. Biol. 5, 105-110] repeated their experiment with a few modifications and reported essentially the same results. The present authors point out the inadequacy of the repeated design in controlling for attentive factors and cite their own experimental evidence for the crucial role of focused attention in psychophysical tasks.

A series of experiments is described which demonstrate the close correspondence between the $P_3$ component of the average evoked response and perceptual behavior in psychophysical tasks involving near-threshold stimuli. However, this relationship appears only when the behavioral measures do not pool or confuse a subject's sensitivity to physical parameters of stimulation and the response criterion he adopts in making his decisions. Whereas Clark, Butler and Rosner relied on the traditional

1Supported in part by Grants MH 14580 and MH 07997 from the National Institute of Mental Health, U.S. Public Health Service. We are indebted to Emanuel Donchin for extensive comments on this manuscript. We thank Crawford Clark, Marion Hartung, David Jenness, Robert Laupheimer, Ray Simon, and Joseph Zubin for assistance and suggestions and Dr. Morton B. Wallach, director of Brooklyn State Hospital, for providing facilities for carrying out this research. Address reprint requests to: Dr. Samuel Sutton, Biometrics Research, 722 West 168th Street, New York, 10032.

Copyright © 1973 by Academic Press, Inc.
All rights of reproduction in any form reserved.
measurement of the threshold, which is a global indicator of correct responses but fails to take into account differences in response bias, the present study was based on Signal Detection Theory, which allows the quantitative assessment of nonstimulus variables. It is shown that the $P_3$ component of the average evoked potential bears a linear relationship to both sensitivity and criterion measures, whereas the same data, plotted as percentage of correct responses, do not relate in any meaningful fashion to $P_3$ amplitude.

INTRODUCTION

A spirited controversy was triggered a few years ago by a report entitled, “Dissociation of sensation and evoked responses by a general anesthetic in man” (Clark, Butler and Rosner, 1969). The authors described an experiment in which the administration of an anesthetic drug greatly diminished or abolished evoked potentials to electrical stimulation of the ulnar nerve, while behavioral measurements of absolute thresholds for these stimuli seemed relatively unaffected. On the basis of these negative results, the authors expressed doubts about “the psychological significance of evoked potentials.” Their report was punctuated by a phrase of such evocative quality that it stood out sharply against the otherwise sober texture of the article: “An extreme form of this argument which may arise in some quarters is that cerebral evoked responses are ‘full of sound and fury, signifying nothing’.” (Clark, Butler and Rosner, 1969, p. 318.) This oblique but devastating conclusion could hardly be ignored, and the report was promptly challenged by Donchin and Sutton (1970), who pointed out serious methodological defects in the study. Clark, Butler and Rosner subsequently repeated their experiment with a few modifications and reported essentially the same results (Clark, Butler and Rosner, 1970).

At the time we received the Clark, Butler and Rosner (1969) paper, a series of experiments was in progress in our laboratory, focusing on the relationship between evoked potentials and behavioral measures based on Signal Detection Theory. At the conclusion of these experiments, it became clear that our results were relevant to the sweeping statements made by Clark, Butler and Rosner. Although not directly comparable to the Clark, Butler and Rosner study, which involved threshold determinations and physiological recordings before and during drug-induced anesthesia, our experiments were basically addressed to the same question: are evoked potentials sensitive to changes in the psychological state of the subject, and do they reflect perceptual or cognitive differences in the subject’s performance when the outside world has not been altered physically? The answer, according to our results, is an emphatic “yes!”: a late positive component of the evoked potential, $P_3$, is clearly related to the subject’s behavioral performance in
psychophysical tasks, provided the behavioral measures do not pool or confuse a subject's sensitivity to physical parameters of stimulation and the response criterion he adopts in making his decisions. Before describing these experiments, it will be helpful to place the "sound and fury" controversy in its proper context by briefly reviewing some pertinent data accumulated in the last decade.

The Controversy in Perspective

The concluding sentence of the report by Clark, Butler and Rosner (1969), which implicitly denies the validity of evoked responses as possible clues to brain-behavior relationships, is startling in view of growing evidence that evoked potentials are related to a number of "complex psychological variables" (Sutton, 1969). It seems increasingly clear that a late positive component identified as P3 (Sutton, 1965) is related to subjective experiences of a perceptual or cognitive nature and does not merely mimic changes in the physical aspects of stimulation. Correlation between various components of evoked responses and perceptual reports have been found in audiometric tests (Rapin, 1964; Davis, 1965; Teas, 1965); in visual discrimination tasks (Bartlett and White, 1965; Vaughan, Costa and Gilden, 1966; White and Eason, 1966; De Voe, Ripps and Vaughan, 1968); in heterochromatic flicker photometry (Siegfried et al., 1965); for discriminations based on cutaneous stimulation (Uttal and Cook, 1964; Rosner and Goff, 1967), and in relation to judgments of loudness (Davis, Bowers and Hirsh, 1968). However, many investigators report that correlations seem to break down when intensity is reduced and stimuli are difficult to discriminate. In many instances, evoked response amplitude increases with increased task difficulty (Davis, 1964; Desmedt, Debecker and Manil, 1965; Spong, 1966; Mast and Watson, 1968; Nielsen, Teas and Idzikowski, 1970), and inconsistencies are especially notable in situations involving threshold or near-threshold intensities. Some investigators find excellent agreement between evoked responses and intensity thresholds (White and Eason, 1966; Siegfried et al., 1965; De Voe, Ripps and Vaughan, 1968; Hillyard et al., 1971), others find disagreement between physiological and behavioral measures (Rapin, 1964; Diamond, 1964; Regan, 1968), and the most discordant note was voiced by Clark, Butler and Rosner (1969), who claimed "dissociation of sensation and evoked responses."

Sources of Disagreement

A number of disagreements may be traced to obvious differences in experimental conditions, such as physical differences in stimuli, rate of presentation, modalities explored, electrode placement, and method of stimulation—all of which are known to influence evoked potentials (MacKay, 1969). It is also true that levels of analyses and desired precision of
measurements vary widely among researchers. Donchin and Sutton (1970), in
their critique of Clark, Butler and Rosner's report, pointed out the fallacy of
comparing evoked potential data and threshold estimates when the behavioral
and physiological measures are based on different sets of trials. The
"decoupled" relationship reported by Clark, Butler and Rosner was based
upon two sets of measures: behavioral data obtained with the method of
limits, and physiological data obtained while no behavioral response was
required. Aside from the fact that psychological conditions under the two
procedures are not comparable, there is evidence that the mere fact that a
response is—or is not—required affects evoked potential waveforms (Hirsh,
1971). But one of the most important objections to the Clark, Butler and
Rosner experimental design is the lack of control for attentional variables. The
effects of attentive factors on evoked potential amplitude was demonstrated
(Sutton and Paul, 1973) in an experiment designed within the framework of
Signal Detection Theory. The subject's task was to detect the presence of a
barely audible click presented in a background of Gaussian noise. This task
click, the signal, was presented randomly with 50% probability. On the other
hand, a clearly audible click termed the "reference" click, occurred on every
single trial and was irrelevant to the task. Evoked potentials to the barely
audible signal were three times larger than evoked potentials to the
"reference" click, although the latter was 30 dB more intense than the signal.
This experiment clearly demonstrates the role of attention in evoked potential
recordings and underlines the necessity for controlling attentive factors when
attempts are made to relate psychological thresholds and physiological
responses.

In a repetition of their experiment, Clark, Butler and Rosner (1970)
attempted to equalize the psychological factor of attentiveness in the procedures
before and during anesthesia by asking the subjects to count stimuli in sets of
eight while physiological recordings were made. This easy task—merely
counting stimuli—is hardly a guarantee that the subjects are behaving at a
comparable attentional level as when they are engaged in difficult psycho-
physical tasks. In our opinion, the only really satisfactory procedure is to
compare behavioral and physiological measures when they are obtained from
the same set of trials. Even when this is done, extreme care must be used not
to confuse "cabbages and kings" by pooling trials arising from different
behavioral categories. Thus as early as 1964, Haider, Spong and Lindsley had
noted that evoked potentials are different for correct, as compared to
incorrect discriminations. More recently both Hillyard et al. (1971) and Paul
and Sutton (1972) have reported that monotonic relationships between
behavioral data and evoked potential amplitude are obtained only for trials in
which the subject was not only correct, but in which stimuli were actually
presented. On the other hand, for correct detections of an absence of a
stimulus, and for incorrect detections, no relationships have been found
between behavioral and physiological data—in fact, it is usually difficult to discern any evoked potentials at the lowest stimulus intensities in these classes of trials.

The Threshold Problem

Perhaps our most serious source of disagreement with Clark, Butler and Rosner arises from what we consider to be their inadequate measurement of the behavioral response, resulting in fallacious comparisons which in turn lead to false conclusions. The psychophysical procedure employed by Clark, Butler and Rosner fails to take into consideration the role of the subject’s response criterion (Sutton, 1970; Hillyard et al., 1971).

It has long been recognized that in psychophysical experiments different methods produce different results (Hake and Rodwan, 1966). Since the method itself is an uncontrolled source of variance, results cannot always be compared across experiments using different procedures. Moreover, and perhaps most important, is the fact that classical psychophysical measures—method of limits, adjustment, constant-stimulus—cannot be relied on to yield a pure measure of sensitivity.

In traditional psychophysics, a stimulus is presented in every, or almost every, observation interval and only two kinds of responses are analyzed, a positive and negative report of detection; the threshold is a global indicator of the observer’s hit rate. In these procedures, reluctance to report a stimulus is called an error, and the highly trained subject is exhorted to maintain a constant criterion. In the “Yes-No” procedure of Signal Detection Theory, the stimulus may or may not be present, resulting in four stimulus-response contingencies: Hits, misses, false affirmatives, and correct rejections. Analysis of the hit rate and the false-affirmative rate makes it possible to obtain a numerical estimate of the response criterion which is independent of the sensitivity index. The observer’s responses are characterized by two independent measures, a sensitivity index, affected primarily by changes in stimulus parameters, and a criterion index, determined by motivation, expectation, evaluation, and other inferred psychological variables which are said to influence decision-making (Swets, 1964).

Because differences in response bias, not sensitivity, often result in differences in performance, a Signal Detection Theory approach to behavioral measurements is often superior to the methods of traditional psychophysics (Clark, Brown and Rutschmann, 1967; Clark, Mehl and Rutter, 1970). If evoked potentials are sensitive to both stimulus and nonstimulus variables, it is essential to adopt a procedure which allows evaluation of both. Our experiments illustrate this important distinction.

---

2W. Crawford Clark, not to be confused with the investigators in the Clark, Butler, and Rosner studies.
RESULTS

Sensitivity, Criterion and $P_3$

In a preliminary experiment (Paul, 1971), we studied the relationship between the evoked potential and a subject's sensitivity to signal intensity using only low-detectability signals. When the criterion level was stable across experimental conditions, the $P_3$ component of the evoked potential increased monotonically with $A'$, a nonparametric analogue of $d'$ (Pollack and Norman, 1964) which is the usual Signal Detection index of sensitivity. This relationship is shown in Fig. 1 for hit trials. It will be noted that the functions are linear, except for the last two points on the curve for subject PM. These points which appear deviant, are in fact properly displaced from the straight line: the amplitude of $P_3$ for these points appears disproportionately high because they reflect not only higher detectability but also a higher criterion level. At all intensity levels, even when detection performance was barely above chance level, $P_3$ was always larger for detected signals (hits) than for undetected signals (misses), as shown in Fig. 2. In the limited range of intensities (signal-to-noise ratios 15 to 19 dB, yielding accuracy levels ranging from .55 to .85 correct decisions) employed in our study, these results are similar to those reported by Hillyard et al. (1971).

In another set of experiments, physical parameters of stimulation were kept invariant and the response criterion was manipulated by two methods: varying a priori probability of signal, and changing the relative payoffs attached to decision outcomes (Paul and Sutton, 1972). Both methods yielded essentially the same results, behaviorally and physiologically. The amplitude of $P_3$ for hit trials increased monotonically with "Per-Cent Bias." This relationship is illustrated in Fig. 3. Since $P_3$ varies as a function of both sensitivity and response criterion, changes in $P_3$ should not be attributed to altered sensitivity unless the observer's criterion is unchanged. For example, even the "deviant" points mentioned earlier (Fig. 1) support this generalization. In that experiment, the signal probability was .50, the payoff matrix was symmetrical, and maximum gain could only be achieved if the subject adopted an "unbiased" criterion, responding "yes" or "no" equally often when uncertain. This "ideal" criterion was adopted by both subjects in all but two instances: observer PM, when performing at the two highest signal-to-noise ratios, adopted a high criterion for reporting a signal; and this resulted in a larger

---

3This point is discussed below, in connection with the experimental manipulation of the response criterion.

4On the other hand, there was little evidence of $P_3$ for correct rejections or false-affirmatives, a fact which suggests that two conditions are necessary for $P_3$ to emerge: the signal must be presented and it must be detected.

5"Per-Cent Bias" is a nonparametric analogue of $L_X$, the signal detection measure of response criterion (see Hodos, 1970, p. 351).
Fig. 1. Relationship between $A'$ and amplitude of $P_3$ for hits for two subjects. One subject was run under two conditions, feedback (f) with respect to accuracy on every trial and no feedback (nf) with respect to accuracy (from Sutton and Paul, 1973).

evoked response than would be expected as a result of increased sensitivity alone.

The Classical Threshold and $P_3$

The "absolute threshold" of classical psychophysics may be defined in terms of a certain statistical rule, for example a stimulus intensity which results in 50% correct responses; it is generally measured under optimal experimental conditions, with highly trained subjects. The validity of this measure, when two or more experimental conditions are compared, depends on the assumption that the subject has maintained a constant criterion. If this is not the case, two identical threshold measures do not represent the same performance in terms of sensitivity, or criterion, or both, and evoked potentials will reflect these differences.

Reanalysis of our experiments in terms of percentage of correct responses yields entirely different relationships than those illustrated in Fig. 3. The relationship between $P_3$ and correctness of both decisions ("Yes" and "No") is shown in Fig. 4, for each subject, for each experimental procedure; varying signal probability (upper graph), and varying values and costs of
Fig. 2. Comparisons between average evoked responses for detected and undetected signals as intensity of signal is increased. Traces for hit trials are superimposed on traces for misses. Signal-to-noise ratio (dB) appears in the right column. The arrow shows the occurrence of the signal. (A) observer $PM_1(t)$ with feedback, at five levels of signal intensity; (B) observer $RL_1(t)$ performing with feedback at four intensity levels; and (C) $RL_1(nf)$ performing without feedback at three levels of signal intensity. In all traces, positivity is down (from Paul, 1971).

outcomes (lower graph). The abscissa is percentage correct responses, pooling hits and correct rejections; the ordinate is the amplitude of $P_3$ for correct responses, measured from the waveforms obtained by computer-averaging together hits and correct rejections. There is no detectable, simple relationship to be derived from these graphs. On the other hand, percentage correct is monotonically related to $P_3$ of waveforms for correct responses when the subject’s criterion is unbiased, as shown in the left graph of Fig. 5. The abscissa is average percent correct responses (hits and correct rejections) for only those cases where experimental procedure called for a “Per-Cent Bias” of zero and the subject’s performance was indeed unbiased. The ordinate is the average amplitude of $P_3$ for waveforms corresponding to correct responses. The graph shown on the right side of Fig. 5, with the same coordinates as the left graph, plots the relationship for those cases where “Per-Cent Bias” was different from zero, as a result of experimental manipulation of criterion, or when the subject adopted a particular response bias unjustified in terms of the experimental situation. These results clearly show that thresholds given in terms of percentage correct responses do not relate to $P_3$ amplitude in any simple fashion because $P_3$ for correct responses is a composite of two response categories that fails to segregate the effects of sensitivity and criterion.
Fig. 3. Relationship between percentage bias and amplitude of $P_3$ for hits. Upper panel, a priori probability (25, 50, or 75%) of signal occurrence is the independent variable. Lower panel, the values and costs of decision outcomes were the independent variable. Abbreviations refer to subjects (from Sutton and Paul, 1973).

DISCUSSION

The relevance of these analyses for the “sound and fury” controversy is that the disappearance or reduction of late components in the evoked potential waveform when a subject is in a semi-anesthetized state may be accounted for on the basis of changes in response criterion. If a subject adopts a “liberal” criterion in the drug-induced state, his overall performance (percentage correct) may not seem to have deteriorated, but $P_3$ will be greatly reduced. When the subject performs in the “normal” state, he is presumably unbiased with regard to responding “yes” or “no.” Under increasing doses of cyclopropane, his subjective criterion may tend to be relaxed; many stimuli will be treated as if they were accepted as signals, and the “absolute threshold” will not appear to have changed considerably. But late components of the average evoked potential—which we have shown to be sensitive to the criterion—will, however, greatly diminish. Clark, Butler and Rosner’s report that at high concentrations of cyclopropane their subjects described a sensation of “floating,” of being “detached” from their environment and “unconcerned” with the passage of time tends to support our suspicion that

6 It may even be difficult to assume a normal state when the subject is anticipating a combination of electrical shocks and anesthesia; apprehension might conceivably lead to a more cautious criterion.
Fig. 4. Relationship between percent correct responses (hits and correct rejections) and $P_3$ for correct responses (computer-averaged evoked responses for the same two perceptual categories).

Fig. 5. Relationship between average amplitude of $P_3$ for correct responses (computer-averaged hits and correct rejections) and average percentage correct (pooling the same two perceptual categories) (a) for all cases when the subject was unbiased (left graph), (b) for all cases when the subject was biased (right graph).
criterion changes did occur in the drug-induced state. The crucial effect of a probable criterion change during anesthesia was not controlled, nor even suspected, and therefore conclusions about "dissociation" between physiological and behavioral data are unjustified.

A fruitful sequel to this prolonged controversy would be to utilize the best features of Clark, Butler and Rosner's physiological techniques and concurrently take behavioral measures based on Signal Detection Theory. Presentation of signals and nonsignals on a probabilistic basis would make it possible to obtain two independent measures, one of signal detectability (replacing the "threshold") and one of response criterion. Simultaneous recording of physiological and behavioral data, and segregation of evoked potentials according to both stimulus conditions and behavioral responses would make it possible to ascertain whether application of cyclopropane does in fact leave performance unchanged, and whether evoked potentials still show correlations with sensitivity, criterion, or both.

REFERENCES


