ON THE RATE OF GAIN OF INFORMATION

BY

W. E. HICK

(*Medical Research Council Applied Psychology Research Unit, Cambridge*)

The analytical methods of information theory are applied to the data obtained in certain choice-reaction-time experiments. Two types of experiment were performed: (*a*) a conventional choice-reaction experiment, with various numbers of alternatives up to ten, and with a negligible proportion of errors, and (*b*) a ten-choice experiment in which the subjects deliberately reduced their reaction time by allowing themselves various proportions of errors.

The principal finding is that the rate of gain of information is, on the average, constant with respect to time, within the duration of one perceptual-motor act, and has a value of the order of five "bits" per second.

The distribution of reaction times among the ten stimuli in the second experiment is shown to be related to the objective uncertainty as to which response will be given to each stimulus. The distribution of reaction times among the responses is also related to the same uncertainty. This is further evidence that information is intimately concerned with reaction time.

Some possible conceptual models of the process are considered, but tests against the data are inconclusive.

I

INTRODUCTION

THE work described in this paper was suggested by the observation that the values of choice-reaction times obtained by Merkel (1885), when plotted against the number of alternative stimuli, appeared to lie very close to a smooth uninflected curve. Merkel himself was chiefly interested in the supposed divisibility of reaction time into "cognition time" and "choice time," and does not even give the raw data. However, they are tabulated by Woodworth (1938). Other psychologists of what may be called the "reaction-time era" discussed the increase in reaction time with number of alternatives, attributing it to such causes as the division of attention or a reduction in the effective intensity of the stimulus; but no quantitative theory seems to have emerged. Indeed, as far as the writer is aware, the only reference to a mathematical relation between reaction time and number of alternatives comes later, when Blank (1934) mentions a logarithmic relation, without suggesting any explanation.

As Merkel's data provide important supplementary evidence for the theory put forward here, his method will be briefly described. (The original paper is not very accessible, and the writer is indebted to Mr. A. Leonard for obtaining it and translating the relevant parts.) The display was provided by a kind of tachistoscope in which the numbers 1–5 (Arabic) and I-V (Roman) were printed on a disc. The subject waited with his fingers on ten keys, and, on the illumination of one of these numerals, he released the appropriate key. An interesting piece of experimental technique is the use of a Geissler tube in order to ensure the sudden onset of the illumination. The Geissler tube was an early form of gas-discharge tube, the forerunner of the modern fluorescent lighting. The illumination would be rather weak, but doubtless quite adequate if the subject was moderately dark-adapted. Some of the other archaisms, however, are less pleasing. For example, there is no indication of the order in which the stimuli were given—i.e. whether it was according to the experimenter's whim or determined by some system. The very large practice improvement, even with ten alternatives, suggests that the sequence was easily learnt, since the present writer, using an irregular sequence, found very little improvement with practice. But according to Merkel's point of view, the predictability of the next stimulus in the sequence was very likely irrelevant. One might also criticize his presentation of the different degrees of choice, from one to ten, in ascending order only, although we possibly have that to thank for the remarkable consistency of the results. Figure 1 shows the reaction times he obtained; each point is the mean of about fifty readings from each of nine subjects.

Consideration of these results led the writer to formulate and test a hypothesis of a new kind, and one which would have been impossible in the early days of reaction-time work, because the theoretical framework did not then exist. To put it succinctly, the hypothesis is that the rate of gain of information is, on the average, constant with respect to time, at least within the duration of a single perception. Further qualification may prove to be necessary, but the evidence presented here shows that the hypothesis is true for the conditions employed.

II
Theoretical Background

More detailed discussion and interpretation will be given later, but some points must be explained here in order to make the experimental approach and results comprehensible.

First of all, the definition of "quantity of information" is that which was originated by communication engineers and has been greatly developed in recent years by Shannon (1949) and Wiener (1948). Briefly, the amount of information given by an event whose probability is p is $-\log p$. It may seem strange that information should depend solely on a probability, but in fact if we introduced any other attribute of the physical—or, for that matter, the psychological—world, the definition would no longer be of information *in general*, but of some particular kind of information. Moreover, it must depend on probability because that is a measure of prior expectation, and information should be that which changes expectation into certainty. By using the negative logarithm, it is ensured that information is always positive (p being a fraction), that it is zero for an event which is absolutely certain to happen and infinite for one which is certain not to happen, and that contributions from independent sources can be added (the probabilities being multiplied together to give the joint probability).

We shall, however, be dealing, not with particular pieces of information, but with average or expected information. We shall average over events of the same kind in order to estimate the probabilities required, and we shall average $-\log p$ over all relevant alternatives to find the expected information. Where the probabilities of the possible alternatives are $p_1, p_2, \ldots p_n$, the expected information (entropy) is the sum of the contributions from each, multiplied by their chances of occurring; thus

$$\mathrm{H} = -\sum_{1}^{n} p_i \log p_i$$

This is essentially the same entropy as appears in statistical mechanics; it is a measure of our uncertainty as to *what* will happen (as distinct from our doubt as to whether a particular x will happen, which is given by $-\log p_x$, as above). Following Shannon (loc. cit.) we write $\mathrm{H}(x)$ for the entropy of the distribution of a variate x.

Shannon shows that if we have a source of messages, signals, or stimuli—whatever we choose to call them—such that their entropy is $\mathrm{H}(x)$, and a destination at which signals having the entropy $\mathrm{H}(y)$ arrive, the average information actually transmitted is

$$\mathrm{R} = \mathrm{H}(x) - \mathrm{H}_y(x).$$

$\mathrm{H}_y(x)$ is the conditional entropy of x when y is known; Shannon calls it the "equivocation." It measures our remaining uncertainty about x even when we know y, on the assumption that the signals are subject to some degree of mutilation in transit by a statistically-definable disturbing influence. If there is no such interference, $\mathrm{H}_y(x)$ is, of course, zero, and the information transmitted is the total generated by the source, namely $\mathrm{H}(x)$.

In a choice-reaction-time experiment, we have a display which is capable of generating any one of a set of n alternative signals. If they are generated in completely random order, the probability of any particular signal is $1/n$ and

$$H(x) = -n.\frac{1}{n}\log\frac{1}{n} = \log n$$

If the subject makes no mistakes—i.e. if the "equivocation" is zero—his response entropy $H(y)$ is also $\log n$. He is then extracting all the relevant information from the display, because $R = H(x)$. (The fact that $H(y)$ is $\log n$ does not *alone* indicate that any information is being extracted, because the responses might be entirely random.)

Now the hypothesis that the rate of gain of information is constant apparently requires that the reaction time should be proportional to $\log n$. But when $n = 1$ (simple reaction) $\log n = 0$, indicating zero reaction time for this case. Evidently $\log n$ does not include the whole of the display entropy, and obviously the missing portion is due to the possibility of "no stimulus" at any instant during the waiting period. In other words, $\log n$ only measures the uncertainty as to what the stimulus will be, and in the simple reaction there is no uncertainty on this point. But there *is* doubt as to when it will occur; and when it does occur, it must be distinguished from the mass of other, irrelevant, information continually pouring in, so as to be recognized as that which was awaited.

At this stage, the writer made a guess that the possibility of "no stimulus" was treated by the subject as if it had the same probability as any particular stimulus. As it seems to have been a successful guess, as far as the present experimental work can show, discussion of its theoretical significance will be postponed. The result is to make the information gained, assuming no mistakes, equal to $\log(n + 1)$. It can be seen from Figure 1 that the equation

$$\text{RT (seconds)} = 0{\cdot}626 \log_{10}(n + 1)$$

does fit Merkel's data very well. (As a matter of interest, the function $A + B \log n$ was tried by Miss V. R. Cane, but proved to give a slightly worse fit; not so much worse, however, as to put it out of court altogether, if some reason for preferring it should arise.) The slight discrepancy when nine and ten stimuli were used may be due to a higher frequency of mistakes, but Merkel gives no data on this. He excluded wrong responses from his results, but that does not rule out this explanation. It is worth noting that Kraepelin (1894) remarks significantly that the more the choice reaction approaches in character the simple reaction, the more errors tend to occur.

However, it seemed that a prima facie case had now been established for a further extension of this approach to the problem of choice and time.

III

Apparatus

As the apparatus is rather complicated, only a general description will be given here. The device for delivering a predetermined irregular sequence of stimuli has been described previously (Hick, 1951). It operates on the punched-tape principle and is driven at a constant rate of about five seconds per stimulus. Electrical signals in binary code pass to the main part of the apparatus, where they can either activate display elements, such as lamps, in the same code, or be first decoded so that only one out of fifteen elements operates at a time. At the same time, four pens record the stimulus on moving paper, also in the same code.

For the present experiments, ten elements of the decoded display were used. They took the form of ten pea lamps arranged in a somewhat irregular circle. The objects of this arrangement were (*a*) to place them sufficiently close together to obviate the need for eye movements, yet not so crowded as to form a confusing pattern, and (*b*) to avoid any very obvious grouping. It is doubtful whether the latter has any relevance at all, since the subject is bound to invent a system of grouping if it is not given to him, but it might have been argued of an externally-imposed system that it really determined the results. The lamps were supplied through a resistance-capacitance network designed to make them light up almost instantaneously. The time between the filament becoming visibly incandescent and the corresponding pens recording was carefully measured by means of a shutter device.

The response was to press the correct one of ten Morse keys on which the subject's fingers rested. But the apparatus can be used with any other form of response which has the effect of selecting one contact out of fifteen (or one combination of not more than four contacts). Thus either the stimulus or the response or both can be given as a pattern or "in clear." The facility for pattern representation has not yet been used.

Since the same pens record the response in the same binary code, we have a fairly complete picture of the events in a run. The chief purposes of using a coding system were to avoid having to keep ten or fifteen pens in working order and to simplify the form of the punched-tape device. There is, however, a disadvantage in that if the subject presses several keys at once, as he is naturally tempted to do, it is not always possible to tell which keys they were. Certain safeguards were incorporated, and these, in conjunction with suitable training, make it safe to say that not more than 0·5 per cent. of the recorded responses can have been wrongly interpreted.

The stimulus sequences are from 100 to 200 stimuli in length, and combine near-equality of frequencies with elimination of the first-order auto-correlation. The sequences were checked by inspection for obvious regularities. Two such sequences were made for each degree of choice from 2 to 10 inclusive, but omitting 7 and 9. Since it was impracticable to make a large number of different sequences, the use of a table of random numbers would not have been a safe expedient.

IV

Experiment I

The first experiment was carried out mainly to confirm the fitness of the function log $(n + 1)$. It served that purpose, and as the results are involved in later arguments, it will be briefly described.

The experimenter acted as subject, in order to gain an idea of the amount of practice likely to be needed for later variations. The rule adopted was to achieve one errorless run before doing each test run; it seemed necessary to have some such incentive in this excessively tedious task. The total amount of practice given is not exactly known, but cannot be less than 8,000 reactions, since the recorded reactions in this experiment total over 2,400. However, much of this practice was needed to make up for long periods away from the task and minor changes in the display.

The degrees of choice were taken first in ascending order from two to ten, and then in descending order, followed by an irregular order. Provided practice at the same degree of choice was given before a test run, there appeared to be no appreciable carry-over of "set" from the previous degree of choice, as far as could be seen from the reaction times. In other words, it is not enough to know that one is going to do a 5-choice run, say; one must have just done perhaps two or three 5-choice runs, if the effect of having previously done, let us say, a 10-choice run is to be virtually abolished.

Since only two fingers (the left little and ring fingers) were used in the 2-choice task, the data for these two fingers were extracted from the results obtained at other degrees of choice. These are the mean reaction times plotted against the degree of choice (n) in Figure 1. In order to obtain a comparable value for the simple reaction time, the 2-choice punched tape was used, but only one of the stimuli was responded to, the other being ignored. Some such method had to be used with this apparatus, on account of the fixed interval between stimuli.

The curve represents 0·518 log $(n + 1)$. It is worth noting that the origin was *not* one of the points to which it was fitted; the fact that it does so—i.e. the fact that the additive constant necessary to give the best fit is negligibly small (less than one millisecond) is more of a coincidence than a sign of precision. The reaction-time measurements themselves may well have a constant error of several milliseconds. Moreover, no allowance has been made for the few milliseconds occupied by peripheral conduction.

Incorrect reactions were omitted from the calculations; as they only amounted to about four per cent., the omission is not thought to be important. However, errors are taken strict account of in the subsequent experiments.

It may be added that the function A + B log n was again tried, and again gave a slightly worse fit. The difference is too small to mean anything by itself, but as it is the second case in which log $(n + 1)$ gives the better representation, and as a third will be cited below, it is worth noting.

V

Experiment II

It was suggested by Mr. J. D. North that if the proposed law has general validity, it ought to apply to the case of partial extraction of information from the stimulus. For example, if the subject can be persuaded to react more quickly, at the cost of a proportion of mistakes, there will be a residual entropy which should vary directly with the reduction in the average reaction time.

Two subjects performed this task. The only special difficulty encountered lay in making enough mistakes to give some points near the origin of the graph, without abandoning altogether the attempt to make the correct responses.

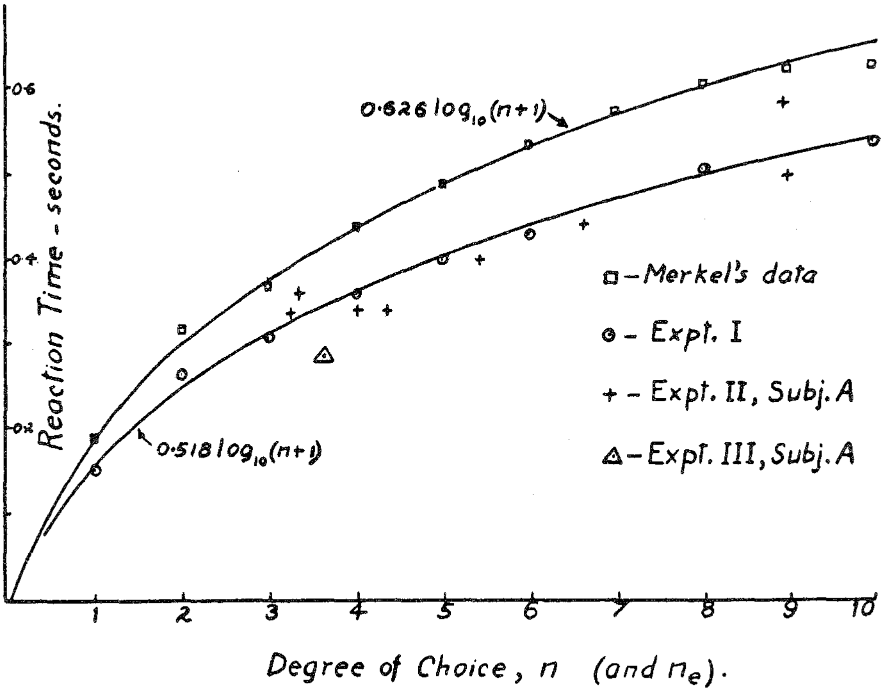

FIG. 1.—Relations between reaction times and numbers of different stimuli. Some results from fast reactions with errors are also plotted to the same scale, on the basis of the calculated equivalent number of stimuli.

The 10-choice sequences were used, and the results are exhibited as reaction times plotted against the equivalent degree of choice (n_e). For example, if there were no mistakes it would mean that all the information was being extracted, and n_e would be 10. The method of calculation is given in Appendix I; it is enough to say here that n_e is the antilogarithm of the amount of information gained (apart from the component due to the possibility of "no stimulus").

One of the subjects was the experimenter, and his results are shown by the crosses on Figure 1. Each cross represents one run of about 100 reactions, and it will be observed that they are distributed reasonably closely about the theoretical curve (which is also the curve fitted to the data of Experiment I).

The performance of the other subject—a research worker—is shown in Figure 2. This subject (labelled B) was trained practically entirely on the 10-choice sequences, and this may have been partly responsible for the curious separation of the reaction times into two groups. The upper group (circles) represents test runs done while the subject was still learning the code and trying to minimize errors. As there was no sign of improvement, he was then asked to try reacting quickly, with as many errors as he liked. This produced ten of the lower group (crosses). In the hope that he had now acquired the technique, he was again asked to do an accurate run, whereupon he reverted to the middle of the upper group. It was only by going back to the high-speed "set" and then very gradually reducing errors in successive runs that it was possible to extend the readings towards the higher values of n_e. This subject, however, managed to achieve a run containing 70 per cent. of errors, without losing control of the situation.

The curve fitted to the lower group represents the function $-0{\cdot}042 + 0{\cdot}519 \log (n_e + 1)$. This means, in effect, that extrapolation backwards by the $\log (n + 1)$ formula misses the origin by 42 milliseconds in this case. Since the points are somewhat scattered and none of them is very near the origin, the discrepancy could reasonably be a sampling error. Any curve which purports to represent the information extracted should, of course, pass through the origin, since where there is no choice there is no information. A curve defined by $A + B \log n_e$ was also tried; it gives a very slightly worse fit, and of course has the theoretical disadvantage of going to negative infinity at the origin.

VI

Experiment III

Since the same two 10-choice sequences were used a considerable number of times by the same subjects in Experiment II, the question of learning arises. The input entropy $H(x)$ was calculated on the assumption that there was no learning; in other words, that the only thing the subject knew about the sequence presented to him was that the frequencies of the different stimuli were equal.

Of course it was not imagined that the subject would learn the actual frequencies of the stimuli, and at the same time refrain from learning any other properties of the sequence. The supposition was rather the negative one that, provided the sequence had no obvious and striking regularities, it would tend to be treated as if random; that is, at any stage during a test run, the next stimulus would be treated, as far as the recognition process was concerned, as if it had an equal chance of being any one of the ten possibilities.

Neither subject was conscious of any learning of the two sequences, or indeed of any thought that it would be worth attempting, but that is no proof that it did not occur. There were exceptions to this statement in two respects. The first stimulus of both sequences happened to be the same, and no stimulus occurred more than twice in succession. Both subjects were aware of these features. The former resulted in a slightly shorter reaction time to the first stimulus, though still of the order of twice a simple reaction time. The latter should have reduced the reaction times to stimuli immediately following such pairs by at least 0·02 of a second, if the knowledge had been consistently acted upon, since it reduces n by one. In reality, the reduction was found to have the quite negligible value of about 0·0006 of a second. But as this has a standard error of the order of 0·025, we cannot be sure that this piece of knowledge was not being used at all, although full advantage of it was probably not taken.

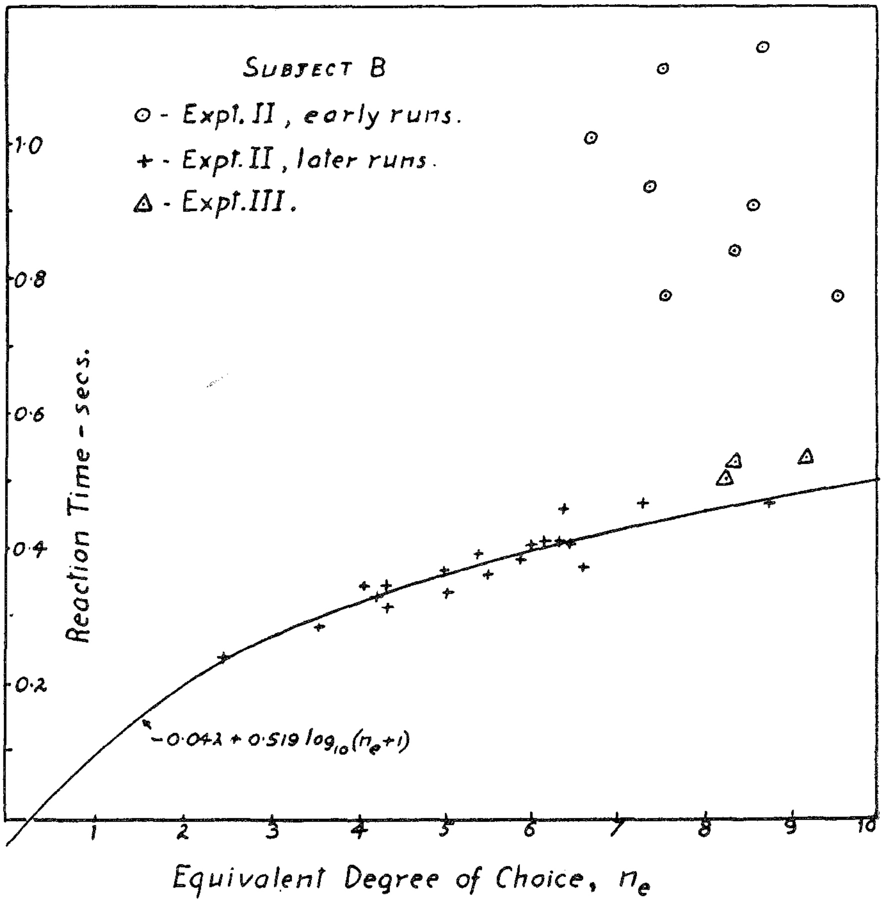

FIG. 2.—Further results from fast reactions with errors.

To find out whether learning of the sequences was introducing a serious error, an entirely new sequence was prepared. It was thought best to retain the same statistical structure with respect to frequencies and first-order serial correlation, although that meant that, again, no stimulus occurred more than twice in succession. One subject (A) did one fast run, and the other (*B*) did three runs at the other end of the "information scale"—i.e. with few errors. The results are indicated by the triangles in Figures 1 and 2. It will be seen that the reaction times are slightly increased at the higher values of n_e, but that at the low value the reaction time happens to be less than expected.

The only conclusion justified by these results is the very limited one intended—namely that learning of the sequences did not play a large part in determining the previous findings. There is also a suggestion that the effect of learning is more marked, the more information is being extracted. This is, of course, inherently likely; if we consider the extreme case of random response, where the only information gained is to the effect that *some* stimulus has occurred, anything learnt about the sequence is irrelevant.

Merely as a matter of interest, we may consider the estimated corrections to the observed reaction times, to give the times that would be expected in the case of unlearnt sequences having the same statistical structure. Naturally they are not intended to be taken very seriously, in view of the assumptions necessary and the small quantity of data available. Suppose $H'(x) = H(x) + H'$, where $H'(x)$ is the input entropy for the unlearnt type of sequence, and $H(x)$ is the effective input entropy of the present sequences, i.e. $H(x)$ is less than $H'(x)$ by the amount of the learnt information H'. The simplest acceptable hypothesis for H' is that it is proportional to the information gained. In terms of effective degrees of choice, the outcome is that the degree appropriate to random sequences is equal to n_e^k. The value of k which best fits the unlearnt-sequence data in Figure 2 is about 1·1, indicating that the amount of learnt information utilized was of the order of ten per cent. of the total extracted. It may also be remarked that, for this subject and this statistical class of sequence, the rate of gain of information from a stimulus has the average value of 5·6 "bits" per second (one "bit" being the information conveyed by an event whose probability is 0·5).

VII

Differences Related to Particular Stimuli and Responses

Before discussing more general questions, a particular finding should be mentioned. In Table I will be found the data for the second subject (B) of Experiment II, whose performance appears in Figure 2. The figures for the aberrant early runs have been excluded, as also have those obtained with the unfamiliar sequence. The Table shows the pooled frequency with which each response was evoked by each stimulus, and the lower entry in each cell is the corresponding mean reaction time. The inclusion of response categories numbered 11 and 14 is to accommodate the few cases of more than one key being pressed simultaneously; 11 and 14 are simply the code numbers of the responses as recorded.

Regarding the Table as a matrix, we see that the responses are grouped about the leading diagonal. Had there been no errors, all the responses would, of course, have been on this diagonal. It can also be seen that the reaction times for particular stimuli, averaged over all responses, are far from equal, and the same is true of the times for particular responses, averaged over all stimuli. It is further noticeable that the probability of the correct response is not the same for all stimuli; stimulus No. 8, for example, seems to have been especially difficult. Do these differences reveal, as we might expect them to, a relation between reaction time and some appropriate measure of information?

The basic hypothesis must be, as before, that reaction time is proportional to information extracted. Unfortunately we have no independent measure of the latter, with respect to individual stimulus categories, because it is not now permissible to calculate the stimulus probabilities according to the frequencies with which they were actually given; it is the probability as seen by the subject that determines the input information. All we can do is to estimate the two sets of residual entropies, which may be called $H_i(y)$ and $H_j(x)$. $H_i(y)$ is the remaining uncertainty about the response, given the ith stimulus, and is assessed directly from the frequencies in the ith column and its marginal total. Similarly $H_j(x)$ is the residual uncertainty, given the jth response, about which stimulus evoked it. These are the two measures of information which we might hope would show some relation to the corresponding reaction times. In fact, the following correlations were obtained:

Between RT_i and $\mathrm{H}_i(y)$: $r = 0{\cdot}798$, $\mathrm{P} < 0{\cdot}01$
,, RT_j ,, ,, $(i = j)$: $r = 0{\cdot}947$, $\mathrm{P} < 0{\cdot}001$.

RT_i is the mean reaction time to the ith stimulus and RT_j is that for the jth response. The correlations with $\mathrm{H}_j(x)$ were insignificant. The scatter diagrams corresponding to the two significant correlations suggested that two points, which were some distance from the remaining eight, were largely responsible for the correlations. However, even when these two points were neglected, the larger correlation was still significant at better than the 5 per cent. level.

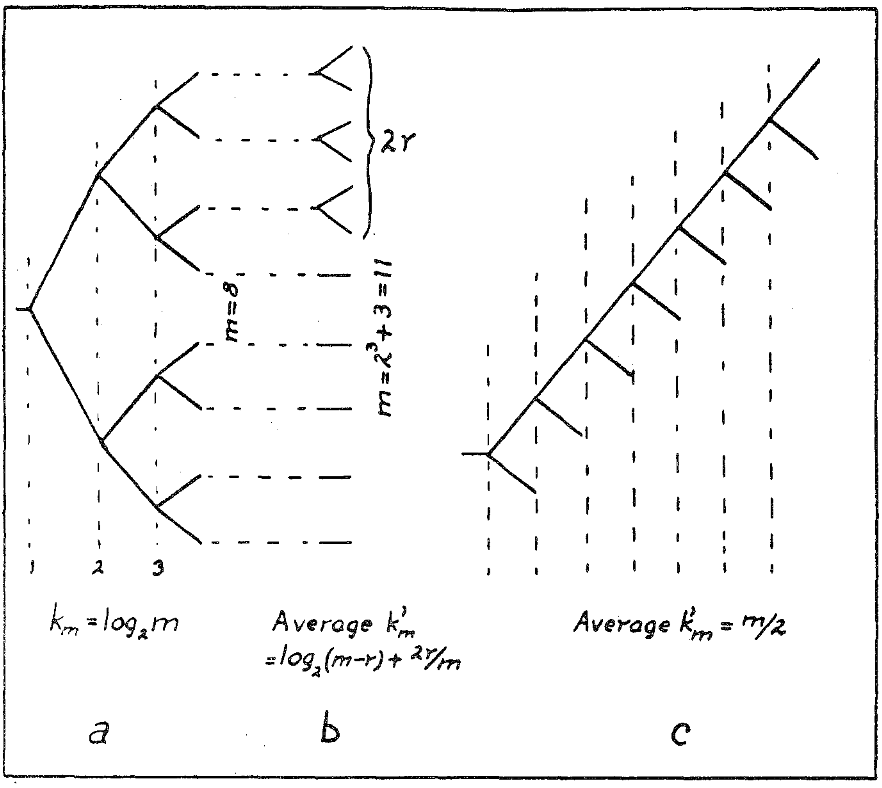

FIG. 3.—Diagram of the progressive classification process. The symmetrical "tree" (*a*) has the required logarithmic property. The minimum asymmetry (*b*), necessary when *m* is not a power of 2, gives a close approximation. Maximum asymmetry (*c*) depicts the systematic search process.

It can be said, therefore, that there is almost certainly some relation between the reaction time of a particular response or to a particular stimulus and the corresponding uncertainty (as seen by the outside observer) as to which response will be evoked by the particular stimulus. How the relation comes about cannot be inferred from the present data, which would apparently agree with several plausible hypotheses. The only point it is desired to make is that this is further evidence pointing to the dependence of reaction time upon information, in the technical sense of that term.

VIII

Conceptual Models

With regard to the mechanism responsible for these results, speculation about neural networks is outside the present scope. There is no objection to trying to depict schematically the component operations, but it must be admitted that what analysis of the data has been carried out does little more than draw attention to the difficulties involved in finding any simple scheme.

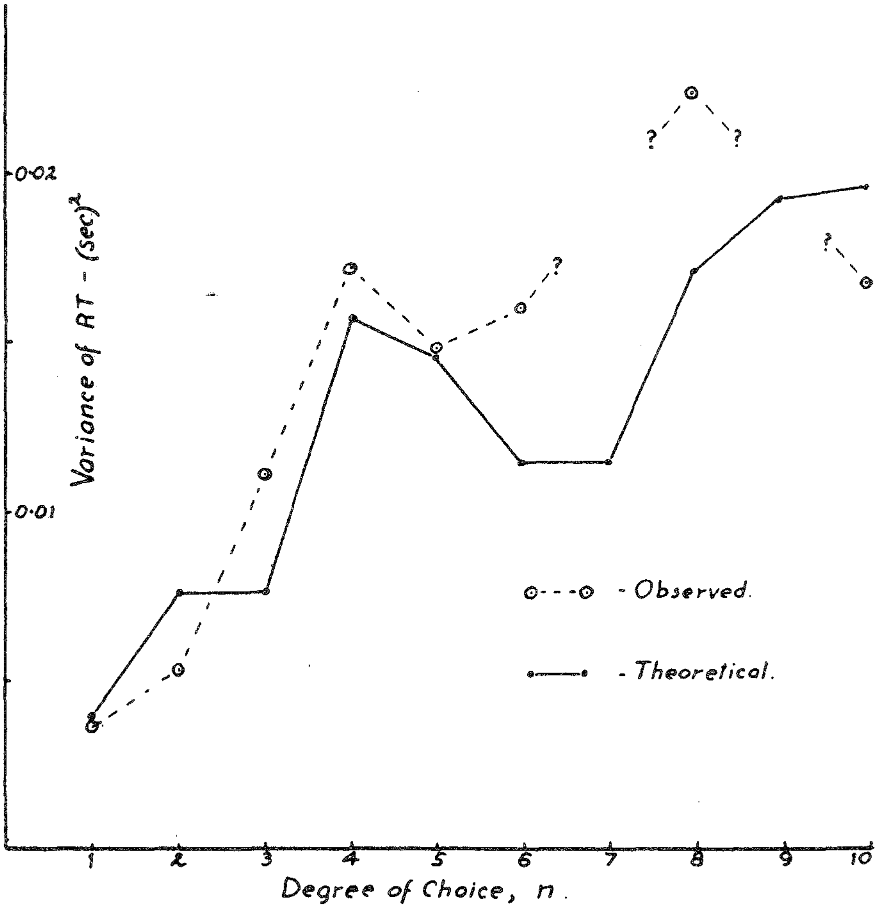

Fig. 4.—Variances of choice reaction times, compared with theoretical variances.

If we consider the process of recognition or identification operationally, we can liken it to matching a given object to the correct one of m gauges or templets. That is to say, the object or event to be identified must be compared with a standard or set of standards. When the matching standard (or combination of standards) has been found, the object may be said to have been identified, as exactly as is possible with that whole set of standards. There are only four fundamental modes of procedure; any actual mode may be regarded as either one of the four or as having a mixed or intermediate character.

Given that an object can be matched with one, and only one, of m standards provided, the problem is to find that one. One of the four possible methods is to produce m replicas of the object and try them on the standards simultaneously. If we wish to try to picture this replication in neural form, we have only to remember that the nervous system must make a replica of the stimulus in any case, and it only requires a leash of fibres diverging from the original afferents to provide as many replicas as may be needed. The real point, however, is that the time occupied in matching is clearly independent of m. Since Reaction-Time increases with m, according to this scheme it can only be because it takes longer to produce a large number of replicas than a few.

If we examine the process of replication as we know it in other systems, we can distinguish three types. Firstly, there is simultaneous replication, in which any number within the capacity of the system can be produced in the same time as it takes to produce one. This together with inversion of the equation $\mathrm{RT} = \mathrm{K} \log (\mathrm{n} + 1)$, leads to a conception of reaction time as a period of continuous accumulation of evidence, which has both advantages and disadvantages over the models discussed here. It is hoped to publish the argument when it has been more fully developed. Secondly, replication may be serial, so that the time taken is proportional to the number produced, in a manner analogous to "manufacture without expansion." This also must be rejected, since reaction time is not linearly related to m. Thirdly, there is self-replication—the "chain-reaction" type of process—in which the number increases in geometrical progression with the passage of time; in other words, the time taken is proportional to the logarithm of the number of replicas required. This at least satisfies one condition. Moreover, self-replication or "increase at compound interest" is very easily provided by a proper arrangement of relays, such as neurones.

In the absence of replication, identification can be effected either by searching or by a system of classification. We can distinguish two extreme types of searching—the purely random and the systematic. In both, the templets—to retain that analogy—are tried one by one, but in the former, any templet may be tried at any stage, and therefore may be tried repeatedly. In systematic searching, no templet is tried more than once, and the correct one must obviously be found in m trials at the most (or $m - 1$ if it is not necessary to try the last one). However, the average number of trials required is, for the random case, m, and for the systematic case, $(m + 1)/2$. That is to say, in both cases the average number is a linear function of the number of alternatives.

Now all the trials are operations of the same kind, and therefore might be expected to take about the same time. Hence if reaction time were due to a search process of this simple sort, it would probably vary linearly with the number of alternatives. It is possible, of course, that a more complicated form of search process would give an approximation to the logarithmic relation required. The complication might take the form of a progressive reduction in the times taken by the elementary operations in the course of a single reaction time; but there seems to be no reason why that should happen, and we can only bear it in mind as a possibility. Alternatively, the probabilities of trying the different templets might be influenced by the choices made earlier in the search. In random searching there is no such influence. In systematic searching, the influence has the simple effect of prohibiting repetitions of the same choice. This, of course, is the best possible method, with the information given; if each trial merely answers the question: "right or wrong?" the average number of trials cannot be less than $(m + 1)/2$. But we wish to manipulate the probabilities so as to make the average number equal to $k \log m$. Unfortunately, this is strictly impossible, because, as m increases, even $(m + 1)/2$ will eventually exceed $k \log m$, no matter how large k is. It is conceivable that a fair approximation to $k \log m$ up to a limited value of m could be obtained, but that has not been attempted.

It may be useful, at this stage, to recapitulate the basic modes of procedure so far considered. They are (*a*) replication with simultaneous trial, (*b*) random searching, and (*c*) systematic searching. Of the three types of (*a*), self-replication was shown to be the only promising one. Neither (*b*) nor (*c*) had anything to recommend them, and we come now to a further reason for regarding them with disfavour—a reason which brings us also to the fourth of the basic modes.

We have seen that to identify one out of m objects requires the extraction of log m units of information. If logarithms to the base 2 are used the units are called "bits." A "bit" is the information we get from applying a test which must turn out in one of two equiprobable ways. Therefore we need only $\log_2 m$ such tests to effect the identification. Now each trial in a search process is certainly a dichotomising test, but the probabilities of "right" and "wrong" are not equal; that is why we need $(m + 1)/2$ trials, which is considerably more than $\log_2 m$.

Is it possible to devise a procedure which is ideally efficient, in terms of the average number of dichotomising tests required? The answer is that in principle a near approximation is possible, and that it involves what may be called progressive classification. The process can be represented by the well-known "tree" diagram, shown in Figure 3. The first test places the object—the stimulus, in the present case —in the correct one of two equiprobable classes. According to the result of this first test, the next one is chosen so as to make a similar cut, and so on until the stimulus has been classified with the degree of precision required. The number of tests or stages of classification applied will evidently be $\log_2 m$, where m is the numbei of terminal sub-classes allowed foi, *provided* that m is an integral power of 2. If m is not an integral power of 2, it is necessary to stipulate that the stimulus probabilities be specially adjusted so that each dichotomy has a probability of 0·5 associated with it. Otherwise, if the stimulus probabilities are held equal, the average number of stages is slightly greater than the expected number $\log_2 m$; but it appears that the maximum excess need not be more than about 0·086 of a stage. It can be seen from Figure 3 that the excess is associated with asymmetry of the "tree," and that the asymmetry is least when we merely add terminal twigs to the symmetrical "tree" which is just too small (Figure 3 (*b*)). The extreme of asymmetry would be a single main trunk with a succession of twigs along its length (Figure 3 (*c*)); this represents the systematic search process referred to previously.

If we can again assume that the component operations—the dichotomising tests—are of like kind and therefore will probably take about the same time, this process of progressive classification agrees closely with the logarithmic relation between reaction time and degree of choice. The only other simple scheme found to satisfy this condition is the self-replication process. Of course there is no real evidence to make us prefer either to the other, but perhaps the classification process may seem a more appropriate model. At any rate, it will not be out of place to consider some of the difficulties in reconciling it with other aspects of the experimental data.

In the first place, although the model accounts for the *average* reaction times lying on a nearly smooth curve, when plotted against the number of alternatives, each individual reaction time is represented as the sum of an exact whole number of stage times. In fact, apart from peripheral delays, each choice reaction time should be an exact multiple of the simple reaction time. Although inevitably blurred by random variations, we might expect some sign of this to show as a periodicity roughly equal to the simple reaction time in the frequency distiibution of a large number of choice reaction times. A group of 773 ten-choice times by the same subject were examined for such an effect, but the only periodicity found was much too short, and was eventually traced to a slight tendency to avoid estimating fractions of a scale interval in measuring the times from the paper record. Of course, if the stage-time variations were highly correlated a periodicity of the kind sought would readily be obscured, and might only appear in a very much larger sample. But it is pointed out below that a high positive correlation would make it difficult to explain the observed variances.

The relation between the variability of the reaction time and the degree of choice might have thrown some light on the problem, but unfortunately it could not be ascertained with much confidence owing to the large scatter. The reason for the scatter is simply that, for the purpose of curve-fitting, we are interested in relative rather than absolute deviations. By the usual approximations for large numbers,

$$\sigma_{\bar{x}} = \sigma/\sqrt{\mathrm{N}} \text{ and } \sigma_{\sigma} = \sigma/\sqrt{2\mathrm{N}}$$

where $\sigma_{\bar{x}}$ is the standard deviation of the mean and σ_{σ} is that of the standard deviation (σ) itself. Reduced to comparable scales, they are σ_{σ}/σ and $\sigma_{\bar{x}}/\bar{x}$ and these, respectively, are in the ratio $\bar{x} : \sigma/\sqrt{2}$. This ratio is of the order of 4 for the present data, so that the standard deviations, when plotted, are bound to—and do—appear that much more scattered than the means.

Nevertheless, it is possible to compare a few hypotheses. According to the classification scheme, $\mathrm{RT}_m = k_m\mathrm{RT}_2$, where RT_m ($= \mathrm{RT}_{n+1}$) is the reaction time for n alternative stimuli and RT_2 is the simple reaction time; k_m is therefore the number of stages of classification. We assume for the moment that k_m is an integer, i.e. that m is an integral power of 2. Hence $k_m = \log m/\log 2$. Now if the stage times vary independently, but with equal variances, the variance of RT_m is

$$\mathrm{V}(\mathrm{RT}_m) = k_m\mathrm{V}(\mathrm{RT}_2) = \mathrm{V}(\mathrm{RT}_2) \log m/\log 2$$

That is to say (granting the hypothesis) that $\mathrm{V}(\mathrm{RT}_m)$ is calculable from the variance of the simple reaction and the degree of choice; and the converse is equally true. More generally, if s is any power of 2,

$$\mathrm{V}(\mathrm{RT}_m) = \mathrm{V}(\mathrm{RT}_s) \log m/\log s$$

Now, the data of Experiment I provide estimates of $\mathrm{V}(\mathrm{RT}_s)$ for $s = 2$ and $s = 4$, but not, unfortunately, for any higher power of 2, because the case of seven stimuli (corresponding to an effective degree of choice of eight, by the main hypothesis) was not tried.

In order to make use of all the data, we may calculate the average number of stages, k'_m, on the assumption that a known proportion of the m signals are given one more stage of analysis than the remainder. In other words, the "tree" is made asymmetrical by the addition of a sufficient number of terminal twigs. The total variance is now increased by the fact that the number of stages varies about the average number. In the absence of correlations, this effect is additive; in fact,

$$\mathrm{V}(\mathrm{RT}_m) = k'_m\mathrm{V}(\mathrm{RT}_2) + \overline{\mathrm{RT}}_2^2\mathrm{V}(k)$$

where $\overline{\mathrm{RT}}_2$ is the mean simple reaction time (assumed to be the same as the mean stage time) and $\mathrm{V}(k)$ is the variance of the number of stages. We must also allow for the fact that the observed variances are those of reaction times to actual stimuli—not to the supposititious "no stimulus." If the latter is a real signal in the present sense, it seems reasonable to assume that it receives the smaller number of stages of analysis; perhaps the best justification for this is that it slightly improves the fit of the calculated variances.

The outcome is shown in Figure 4. The observed values are for stimuli Nos. 1 and 2 only, as before. The lines joining the points are merely inserted for clarity; they have, of course, no significance. The calculated $\mathrm{V}(\mathrm{RT}_2)$, which will be seen to differ slightly from the observed value, is the weighted mean derived, according to the hypothesis, from the observed variances. The calculated values do show some tendency to follow the fluctuations of the observed values, but not to any convincing degree. In fact, a rough test suggests that the fit is quite unacceptable, and if the main hypothesis is to be retained, we must postulate some further source of variance not so far accounted for.

TABLE I

Pooled Reaction Times (seconds) and Response Frequencies for Second Subject in Experiment II

Response Categories 11 and 14 accommodate the few cases where more than one key was pressed

Resp. (*j*)	*Stimulus* (*i*)										*Totals and Mean RT's*
	1	2	3	4	5	6	7	8	9	10	
1	165 0·36	2 0·35	4 0·31	1 0·30	0 –	1 0·17	0 –	0 –	0 –	1 0·37	174 0·36
2	16 0·30	167 0·35	27 0·36	2 0·28	0 –	2 0·18	1 0·13	0 –	2 0·25	1 0·23	218 0·34
3	2 0·23	24 0·32	145 0·47	32 0·44	5 0·32	1 0·27	0 –	1 0·20	0 –	1 0·33	211 0·44
4	0 –	1 0·27	12 0·44	144 0·45	38 0·45	2 0·42	0 –	0 –	1 0·30	2 0·37	200 0·45
5	1 0·43	0 –	2 0·67	29 0·40	139 0·45	32 0·39	1 0·30	1 0·43	0 –	1 0·33	206 0·43
6	0 –	0 –	1 0·30	0 –	6 0·39	133 0·46	38 0·46	0 –	1 0·33	0 –	179 0·46
7	1 0·33	0 –	0 –	2 0·42	6 0·37	21 0·40	126 0·51	48 0·45	2 0·40	1 0·40	217 0·48
8	0 –	0 –	1 0·23	0 –	0 –	0 –	12 0·46	95 0·53	24 0·47	3 0·38	135 0·51
9	1 0·27	0 –	0 –	0 –	0 –	0 –	3 0·44	40 0·43	148 0·48	33 0·41	225 0·46
10	1 0·40	0 –	0 –	0 –	0 –	0 –	1 0·27	2 0·35	12 0·40	143 0·47	159 0·46
11	0 –	0 –	0 –	0 –	0 –	0 –	0 –	0 –	1 0·47	4 0·40	5 0·41
14	0 –	0 –	0 –	0 –	0 –	0 –	0 –	1 0·33	0 –	0 –	1 0·33
Totals Mean RT's	187 0·35	194 0·35	192 0·45	210 0·44	194 0·44	192 0·44	192 0·49	188 0·48	191 0·47	190 0·45	1930 0·436

However, it is at least possible to say that complete positive correlation between stage-time variations would make the fit very much worse. More generally, if the whole process can be broken down into a sequence of operations of equal average durations, and if their number increases not less rapidly than log m, it seems unlikely that their durations could have large positive correlations. The present data provide no basis for considering combinations of positive and negative correlations.

All that can be added with regard to variances is that those of Experiment II are of similar magnitude to the ones just discussed, and show the same decelerating upward trend with increasing degree of choice (which, it will be remembered, is the theoretical equivalent degree, in this case).

IX

Discussion

Turning now to more general topics, we may consider first some of the implications of the provisional conclusion that information is gained at a constant rate. Perhaps the most important, even though the most obvious, point is that it has proved valid to estimate the stimulus probabilities *ab extra*. Although the differences with respect to individual stimuli and responses suggest that the subjective—or perhaps one should say the psychologically effective—probabilities do not exactly correspond to the objective frequencies, it will be an enormous practical advantage if, for the purpose of estimating average effects, it can be assumed that they do. To discover the limits within which that assumption is justifiable will require a great deal of experimentation. It may be conjectured, for instance, that the effective probabilities are very little affected by increasing inequality in the stimulus frequencies until something like a threshold is reached. Certainly this matter would have to be examined before anything more than the most tentative application of information theory in real-life situations could be made, for it must be seldom that all the relevant possibilities are equiprobable, either subjectively or objectively. However, it may be found both practicable and valid, in some cases, to estimate subjective probabilities by some form of "guessing" technique.

Perhaps the whole matter can be best summed up in the following way. Fairly strong evidence has been obtained that the amount of information extracted is proportional to the time taken to extract it, on the average. Reasons have been adduced which seem to make this proposition inherently likely. But the simplest scheme of operations which fits the general proposition has been found to lead to hypotheses which other aspects of the data largely fail to confirm, although they do not definitely contradict it. At present, therefore, it is impossible to venture beyond the general statement in terms of information theory. This, indeed, may be adequate for practical applications; but it inevitably leaves the details vague; and so they must remain, until more evidence or better reasoning is brought to the problem.

Appendix I

The method of estimating the average information per stimulus which is gained in each run was as follows. It has already been mentioned that the information gained (R) is

$$R = H(x) - H_y(x)$$

An alternative and more convenient formula is

$$R = H(x) + H(y) - H(x,y)$$

where $H(x,y)$ is the joint entropy based on the joint probabilities of particular stimulus-response pairs. The response frequencies were entered in a table similar to Table I, and R was computed from the above formula, with the appropriate frequency ratios substituted for the probabilities: thus

$$R = \log N - \{\sum_i f_i \log f_i + \sum_j f_j \log f_j - \sum_{ij} f_{ij} \log f_{ij}\}/N$$

where f_i is the marginal total of the i^{th} column, f_j is that of the j^{th} row, f_{ij} is the number in the cell ij, and N is the grand total.

Since R is a quantity of information, it can be formally expressed as the logarithm of a number of equiprobable alternatives; thus

$$R = \log n_e$$

whence the effective degree of choice (n_e) can be obtained. The expression is only formal because, of course, n_e may not happen to be an integer.

Now R is only the information with respect to *which* stimulus occurred; we have still to consider the component due to the occurrence of *some* stimulus. Let $p(s,i)$ be the joint probability of a stimulus occurring and of its being the ith stimulus. Then the total input entropy is

$$H(X) = -\sum_i p(s,i) \log p(s,i) - q(s) \log q(s)$$

where $q(s)$ is the probability of no stimulus. After some manipulation this becomes:

$$H(X) = -p(s)\sum_i p_s(i) \log p_s(i) - p(s) \log p(s) - q(s) \log q(s)$$

where $p_s(i)$ is the conditional probability of the ith stimulus, given that some stimulus must occur, and $p(s) = 1 - q(s)$. This can be written briefly as

$$H(X) = p(s)H(x) + H(s)$$

in which $H(x)$ is the same as the $H(x)$ in the formula for R given above. $H(s)$ may be regarded as the uncertainty as to when the stimulus will occur, or the information to be gained from the fact that *some* stimulus has occurred.

Now, if there are no superfluous responses and no failures to respond, we can say that the $H(s)$ component suffers no loss in transmission, the only loss being that sustained by $H(x)$ in its degeneration into R. Therefore the total information transmitted is

$$R_t = p(s)R + H(s).$$

But the immunity of $H(s)$ from depreciation does not necessarily imply that it is independent of R. The temporary capacity of the organism for extracting information from the display is, in a limited sense, indicated by R; in fact, we have expressed this capacity as n_e, the number of equiprobable categories into which the n actual stimuli are, in effect, divided. The further information needed for the selection of one particular response must be drawn from some independent source—an irrational preference or an appeal to chance or something of that kind. In other words, if the stimuli are grouped into n_e categories, the display is being interpreted as if it could generate only n_e different stimuli. It is therefore reasonable to expect that the possibility of no stimulus gives us, altogether, $n_e + 1$ equiprobable signals, by analogy with what we have found to apply to ordinary choice reaction times.

Adopting this assumption, we write $p(s) = n_e/(n_e + 1)$, and the total information transmitted takes the simple form:

$$R_t = \log (n_e + 1)$$

As we have seen (Figures 1 and 2), this function gives a reasonably close fit to the corresponding reaction times, as the main hypothesis requires.

References

1. Blank, G. (1934). Brauchbarkeit optischer reactionsmessungen. *Indust. Psychotech.*, **11**, 140–150.
2. Hick, W. E. (1951). A simple stimulus generator. *Quart. J. Exp. Psychol.*, **3**, 94–95.
3. Kraepelin, E. (1894). Beobachtungen bei zusammengesetzen reaktionen. *Philos. Stud.*, **10**, 499–506.
4. Merkel, J. (1885). Die zeitlichen verhältnisse der willensthätigkeit. *Philos. Stud.*, **2**, 73–127.
5. Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana.
6. Wiener, N. (1948). *Cybernetics.* New York.
7. Woodworth, R. S. (1938). *Experimental Psychology.* New York.