INFORMATION AND CONTROL 2, 315–332 (1959)

The Information Content of Nonsequential Messages

B. R. JUDD* AND N. S. SUTHERLAND

The Clarendon Laboratory, University of Oxford, and the Institute of Experimental Psychology, University of Oxford, Oxford, England

In psychological experiments on recognition and recall, the subject clearly conveys information. If a technique could be developed for measuring the information transmitted in these situations, meaningful comparisons could be made between subjects' performances under different conditions. With this end in view, a mathematical technique is developed for measuring the information in a message where *no* information is transmitted by the order in which the symbols composing the message are received. The theory is presented in four stages. At each stage assumptions or, as at stage IV, approximations are made which enable the information transmitted by a subject to be estimated by performing fewer experiments than at the previous stage. A crucial assumption at stage III involves minimizing the information subject to certain parameters being held constant. The assumptions are discussed, and the theory is illustrated by calculating the information transmitted in recall and recognition tasks by a set of subjects in an actual experiment. Further applications are briefly discussed.

1. INTRODUCTION

Stated in general terms, the problem which this paper attempts to solve is as follows. A set G comprising g dissimilar symbols is selected from a set K of total number k and is sent by a source. These symbols are transmitted to a receiver which selects a set Q consisting of q dissimilar symbols, where $Q \subset K$. We attempt to discover how much information is transmitted over the channel between source and receiver, given that there is no correlation between the *order* in which the symbols are sent and the order in which they are received.

This problem has arisen in the course of psychological experiments on recognition and recall. In experiments on recall a subject is presented

* Present Address: The Radiation Laboratory, University of California, Berkeley.

with g dissimilar symbols (the set G), and is later asked to recall as many as he can: for example, supposing g is 15, a given subject may recall 15 of which 10 are contained in the set G, and five are not contained in G; whereas another subject may recall only eight but all eight are contained in G. Because of the difficulty of comparing the performance of subjects who recall a different number q of symbols in this way, many experimenters have restricted the number which subjects are to recall—e.g., in the present example a subject might always be instructed to recall 15; if this results in the subject making guesses, this will tend to reduce his level of performance. In the past, psychologists have made very crude attempts to cope with the problem of differential numbers of errors occurring when the number of symbols to be selected by the subject is not specified in advance; e.g., a score may be computed based on giving points for every symbol correctly recalled and taking away an arbitrary number of points for every symbol incorrectly recalled (cf. Hunter, 1957, p. 18). At present there is no rational technique for comparing the performance of subjects who perform in these different ways, nor for comparing their performances under conditions where the numbers g and k are changed. This problem becomes particularly acute if we wish to compare recognition with recall. In recognition experiments a subject is first shown a set G of g symbols which is selected from a set M, and is later presented with the set K, where $G \subset K \subset M$, from which he makes his actual choice. Under these conditions it has frequently been found (e.g. Postman *et al.*, 1948; cf. further references there cited) that recognition is superior to recall in so far as more symbols are correctly recognized by the subject than are correctly recalled. However, subjects are not necessarily more efficient at recognition in the sense that they convey more information when recognizing than when recalling. Thus they may make more correct responses in recognition only at the expense of making more errors; and, more importantly, in recognition experiments as previously conducted, they have been selecting from a set K which is smaller than the total set M from which symbols might have been transmitted; whereas in recall, selection has to be made from the total set M. If the amount of information transmitted in recall and recognition can be measured, then we have a rational method of comparing the efficiency of performance on the two tasks, and of different performances on the same task.

Several authors (e.g. Brown, 1959) have stressed the desirability of applying information theory to the results of experiments on memory, but the appropriate mathematical techniques have not so far been developed. The aim of this paper is to begin the task of supplying this need: as will be shown below, the development of the mathematical techniques opens up a new field of experimentation, because it becomes possible to compare meaningfully subjects' performances under widely different conditions.

2. ENTROPY AND INFORMATION

We commence by introducing the symbols which will be used in the subsequent account. The operation of sending the set G and receiving the set Q will be said to constitute an *experiment*. The intersection of G and Q defines a set R of r dissimilar symbols, and we shall say that the receiver selects r correct symbols. In view of the relations $K \supset G \supset R$ and $K \supset Q \supset R$ we have $k \geqq g \geqq r$ and $k \geqq q \geqq r$. The number q may be greater than, equal to, or less than g. For reasons which will be discussed later it is convenient to introduce three further symbols. The quantity s is defined as the average value of r over an infinitely long series of experiments; p_r is defined as the probability of the receiver selecting r correct symbols for a given k, g, and q. Thirdly, an "S" symbol is defined as any *one* combination of symbols which can be sent in one experiment in G, or which can be received in one experiment in Q. Where the word "symbol" alone is used it will refer to the individual symbols in K, G, or Q: each different selection of g symbols from k constitutes an "S" symbol, and each different selection of q symbols from k constitutes an "S" symbol.

The problem to be solved is to find an expression for I, the average amount of information transmitted per experiment. In what follows we develop successively four expressions for evaluating I. The results at each stage will be used in the succeeding stages. Each succeeding stage enables us to evaluate I with fewer experiments; but in order to do this, extra assumptions have to be made at each stage. Assumptions will be introduced at the stage at which they become necessary, but the discussion of the validity of the assumptions in the context of a psychological experiment will be postponed until section 3.

Stage I

Assumptions

(1) The numbers k, g, and q are the same for all experiments of a series.

(2) In one experiment no symbol occurs more than once in G and no symbol occurs more than once in Q.

(3) There is no correlation between the order in which the symbols of G are sent and the order in which the symbols are received in Q.

(4) The series of experiments constitutes an ergodic sequence: i.e. the probabilities of sending a given "S" symbol and of receiving a given "S" symbol do not change over the series of experiments. This implies that p_r for each value of r does not change over the series of experiments.

Derivation

If the above assumptions hold, it is possible to derive an expression for I by a direct application of Shannon's (1948) formulae to the "S" symbols. Thus

$$I = H_A + H_B - H_{AB} \tag{1}$$

where H_A is the entropy of the source, H_B is the entropy of the receiver, and H_{AB} is the joint entropy of the source plus the receiver (i.e. the entropy of the whole system). In addition, the entropy of a system is

$$-\sum_i p_i \log p_i \tag{2}$$

where p_i is the probably of the state i. The amount of information conveyed per experiment is therefore

$$-\sum_i p_i \log p_i - \sum_j p_j \log p_j + \sum_{i,j} p_{ij} \log p_{ij} \tag{3}$$

where p_i is the probability of the ith "S" symbol being sent, p_j is the probability of the jth "S" symbol being received, and p_{ij} is the probability of the ith "S" symbol being sent and the jth "S" symbol being received.

Discussion

This result is trivial: obtaining it depends merely on the decision to consider "S" symbols as the basic information symbols rather than the individual symbols in K. However, derivation of further formulas depends on taking this step. Unfortunately there will be very many more "S" symbols than there are symbols in K. Thus the possible number of "S" symbols sent is ${}^kC_g$ and the possible number of "S" symbols received is ${}^kC_q$; hence there will be ${}^kC_g\ {}^kC_q$ possible states of the whole system. For $k = 90$, $g = 15$, and $q = 12$, this number is about 10^{31}. In order to apply Eq. (3) we must evaluate the probability of occurrence of each of these states. It would clearly be impossible to perform the number of experiments necessary to obtain sufficient data for the evaluation of these probabilities. We proceed to consider whether by introducing further assumptions we can obtain an expression for I which can be evaluated on the basis of fewer experiments: the principle behind our method is to attempt to group different "S" symbols into classes.

STAGE II

Assumptions

(5) The *a priori* probability of a given symbol being sent or received in any experiment is the same for all symbols in K (namely g/k for symbols sent, and q/k for symbols received).

(6) There are no correlations between the occurrence of given symbols either in the set of symbols sent or in the set of symbols received; i.e., the probability of any symbol G_i occurring in G is independent of whether or not any other symbol G_j is also contained in G; similarly the probability of any symbol Q_i occurring in Q is independent of whether or not any other symbol Q_j is also contained in Q.

(7) Given that a symbol G_h occurs in G, there is an equal probability of any two symbols Q_i and Q_j occurring in Q if G_h, Q_i, and Q_j are different symbols; i.e.

$$P_{Q_i}(G_h) = P_{Q_j}(G_h)$$

In addition, for any symbols Q_i and Q_j,

$$P_{Q_i}(G_i) = P_{Q_j}(G_j)$$

where Q_i and G_i are the same symbol and Q_j and G_j are the same symbol.

Derivation

From assumptions (5) and (6) above it follows that every way of selecting g symbols from K is equi-probable. Applying equation (2), the entropy of the source is

$$H_A = -\log ({}^kC_g)^{-1} = \log {}^kC_g$$

In an analogous way the entropy of the receiver is

$$H_B = \log {}^kC_q$$

From assumption (7) it follows that every state in which r symbols of Q are contained in G is equi-probable. The receiver may select r symbols from g in ${}^gC_r$ ways, and there are ${}^{k-g}C_{q-r}$ ways in which it may select the remaining $q - r$ symbols from those excluded from G. The total number of states of the source plus receiver is therefore ${}^kC_g\, {}^gC_r\, {}^{k-g}C_{q-r}$. The entropy of the combined system of source plus receiver is hence given by

$$H_{AB} = -\sum_r p_r \log\,({}^kC_g\, {}^gC_r\, {}^{k-g}C_{q-r})^{-1}\, p_r$$

We find (1) evaluates to

$$I = \sum_r p_r \log \Omega_r p_r \tag{4}$$

where

$$\Omega_r = \frac{k!r!(g-r)!(q-r)!(k-g-q+r)!}{q!(k-q)!g!(k-g)!} \tag{5}$$

If the sets R, Q, and G coincide, then $q = g$, and every p_r is zero except $p_q = 1$. In this case it is easy to see that

$$I = \log {}^kC_g = H_A = H_B$$

This is an example of noiseless transmission, every state of the source determining uniquely a state of the receiver. If, on the other hand, the set Q are chosen at random (subject to the condition that they are dissimilar), then, since the probability p_r of r of these symbols being contained in the set G is ${}^gC_r\, {}^{k-g}C_{q-r}/{}^kC_q$, we immediately obtain the expected result that $I = 0$ by substituting this expression for p_r in Eq. (4).

Discussion

Making use of Eq. (4), we can now estimate I if we know the probabilities p_r. In the case formerly discussed, where $q = 12$, r can take 13 values, and in order to estimate I we need only to perform enough experiments to obtain an estimate of the probability of occurrence of each of these 13 values. This could clearly be achieved to within a good approximation in little more than 100 experiments. Thus it is practicable to apply formula (4). We now ask whether by introducing any further assumptions we can obtain an expression for I which could be evaluated on the basis of fewer experiments. The principle behind the next stage is that we assume that for a given s, the values of p_r are distributed in a random way. The sense in which the word "random" is used will be discussed in Section 3.

Stage III

Assumption (8)

For given values of k, g, and q, the quantities p_r that characterize the receiver possess values which minimize I subject to a prescribed value for s.

Derivation

We have to minimize

$$I = \sum_r p_r \log \Omega_r p_r$$

subject to the two conditions

$$s = \sum_r r p_r \tag{6}$$

which follows immediately from the definition of s, and

$$1 = \sum_r p_r \tag{7}$$

The technique for doing this involves the use of Lagrange's Undetermined Multipliers, and the solution is

$$p_r = A e^{\alpha r}/\Omega_r \tag{8}$$

Substituting into (4), and using (6) and (7), we find

$$I = \log A + \alpha s \tag{9}$$

The constants A and α can in principle be determined from (6) and (7). However, it is numerically much easier to select a value of α, calculate the corresponding value of A from the equation

$$A^{-1} = \sum_r e^{\alpha r}/\Omega_r \tag{10}$$

and then find s from

$$sA^{-1} = \sum_r r e^{\alpha r}/\Omega_r \tag{11}$$

Asking for the minimum information is very similar to the demand made in statistical mechanics that the physical entropy S should be a maximum, the analogue of s being the total energy E of the system. A similar mathematical procedure has been used to minimize information in a different context (Mandelbrot, 1953; Good, 1957).

At this point it is convenient to turn aside from our program for developing formulas for I to discuss a number of topics which the minimizing procedure raises. The following section, which is largely mathemati-

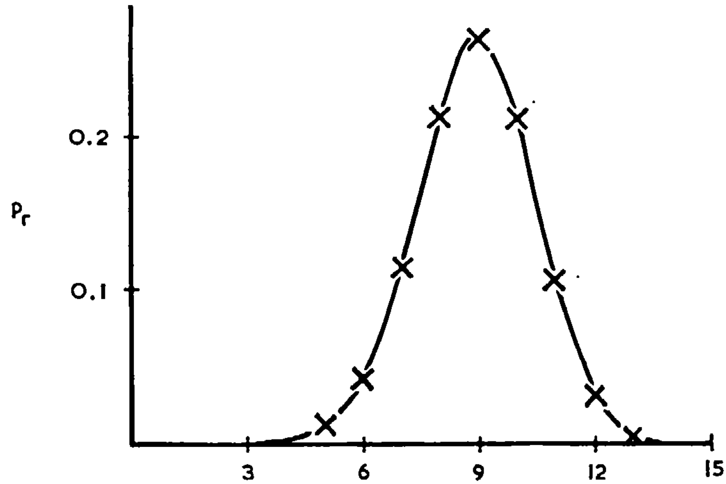

FIG. 1. The crosses show the values of p_r against r for $k = 90$, $q = g = 15$. The full line has been drawn in to link up adjacent points. It can be obtained in detail by replacing factorials by Γ-functions in the expression for Ω_r.

cal, may be omitted by readers interested mainly in the relation of the theory to problems in psychology.

THE PROPERTIES OF α, s, AND I

We note that the values of p_r given in (8) are defined only for integral values of r, though it is of course possible to construct a continuous function by replacing the factorials which occur in Ω_r by Γ-functions. This is illustrated in Fig. 1 for the special case of $k = 90$, $g = q = 15$. The value of α is 2.7535, and with the aid of (9), (10), and (11) we find $s = 8.9159$ and $I = 9.1567$. The maximum of the curve of p_r against r may of course be obtained by differentiation. The derivatives of Γ-functions are not particularly simple, however, and instead of finding the maximum of p_r for a given α, it is much more convenient to determine the value of α which makes the p_r associated with a particular value of r, say t, a maximum. From (8),

$$\frac{dp_t}{d\alpha} = \frac{dA}{d\alpha} e^{\alpha t}/\Omega_t + Ate^{\alpha t}/\Omega_t$$

It is easy to show from (10) and (11) that

$$(dA/d\alpha) = -As \tag{12}$$

and hence

$$(dp_t/d\alpha) = A(t - s)e^{\alpha t}/\Omega_t \tag{13}$$

This is zero when $s = t$. It is clear that the two ways of finding the maximum of p_r are very similar, and that the maximum of the p_r curve for a fixed α will be expected to occur at a value of r close to s. This is so for the example illustrated in Fig. 1.

From a computational standpoint it is highly desirable to know what value of α to select which, when substituted in (10) and (11), leads to a prescribed value for s. There is no solution to this problem which is both exact and simple. However, we may use Stirling's approximation (see Whittaker and Watson, 1952) to replace $x!$ by $(x/e)^x$ for the factorials which occur in Ω_r, and it is now a simple matter to differentiate p_r with respect to r. We find

$$\frac{1}{p_r}\frac{dp_r}{dr} = \alpha - \log\frac{r(k - g - q + r)}{(g - r)(q - r)}$$

Setting dp_r/dr equal to zero, we obtain

$$\alpha = \log\frac{r(k - g - q + r)}{(g - r)(q - r)} \tag{14}$$

In view of the penultimate sentence of the previous paragraph, we may expect that by substituting $r = t$ in (14), a value of α will be obtained, which, when inserted in (10) and (11), will lead to a value of s very close to t. A Ferranti Mercury digital computer was programmed to run over integral values of t and find sequences of values of α, A, s, and I. Table I contains the results for s and I in the case where $k = 90$, $q = g = 15$. The quantity t runs from 1 to 14, and it can be seen that the values of s are quite close to integers, thus demonstrating the effectiveness of (14). The I, s curve so obtained is drawn in Fig. 2.

Every value of α defines uniquely a value of s (for a given k, g, and q) and *vice versa*. When $s = 0$, corresponding to $\alpha = -\infty$, the receiver selects q symbols, none of which are included in G. The random distribution corresponds to $\alpha = 0$. As α becomes infinitely large and positive, s is equal to the smaller of q and g, and the receiver selects as many correct symbols as is possible for the given value of q.

Under a variation in α, Eq. (9) becomes

TABLE I

DATA FOR FIG. 2

t	s	I	I in bits
1	0.9916	0.778	1.122
2	1.9952	0.076	0.110
3	3.0063	0.070	0.101
4	4.0222	0.595	0.858
5	5.0409	1.580	2.279
6	6.0615	2.992	4.317
7	7.0832	4.820	6.953
8	8.1054	7.067	10.195
9	9.1280	9.751	14.067
10	10.1509	12.907	18.621
11	11.1742	16.594	23.940
12	12.1990	20.911	30.168
13	13.2289	26.051	37.583
14	14.2802	32.496	46.882

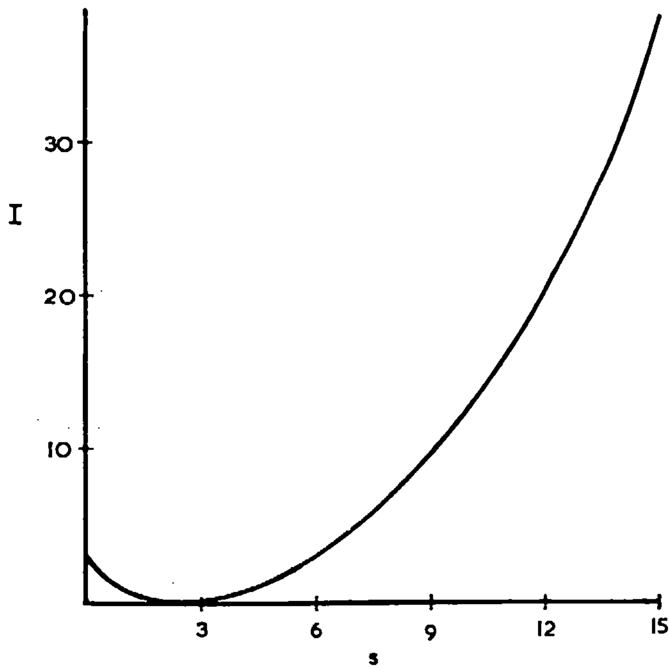

FIG. 2. Values of I are plotted against s for $k = 90$, $q = g = 15$. The random case corresponds to $s = 2.5$, for which $I = 0$. The units in which I is measured can be converted to bits by multiplying I by $\log_2 e$.

$$\begin{aligned}\delta I &= A^{-1}\delta A + \alpha\delta s + s\delta\alpha\\ &= \alpha\delta s\end{aligned}$$

from (12). Hence

$$(dI/ds) = \alpha \tag{15}$$

and the slope of the I, s curve is simply α.

The physical analogue of (15) is $dS/dE = 1/T$, where T is the absolute temperature. It is interesting to observe that a receiver which makes fewer correct responses than a random set corresponds to a system at a negative absolute temperature. Such systems have been discussed by Purcell and Pound (1951).

To make further progress we must examine the detailed nature of Ω_r. This function is symmetrical in q and g, and all formulas derived from Ω_r must therefore be invariant with respect to their interchange. From (10),

$$\begin{aligned}A^{-1} &= \sum_r e^{\alpha r}\frac{q!(k-q)!g!(k-g)!}{r!(k-g-q+r)!(g-r)!(q-r)!k!}\\ &= \frac{(k-q)!(k-g)!}{k!(k-g-q)!}\left[1+\frac{gq}{k-g-q+1}e^{\alpha}\right.\\ &\qquad\left.+\frac{g(g-1)q(q-1)}{(k-g-q+1)(k-g-q+2)}\frac{1}{2}e^{2\alpha}+\cdots\right]\\ &= \frac{(k-q)!(k-g)!}{k!(k-g-q)!}F(-q,-g;k-g-q+1;e^{\alpha})\end{aligned}$$

where the hypergeometric function $F(\alpha, \beta; \gamma; z)$ is defined in the usual way as

$$1+\frac{\alpha\beta}{\gamma}z+\frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{1}{2}z^2+\cdots$$

From (10) and (11) it is clear that

$$sA^{-1} = (d/d\alpha)(A^{-1}) \tag{16}$$

and hence

$$\begin{aligned}sA^{-1} &= qge^{\alpha}\frac{(k-q)!(k-g)!}{k!(k-g-q+1)!}\\ &\qquad\cdot F(-q+1,-g+1;k-g-q+2;e^{\alpha})\end{aligned}$$

from the well-known properties of hypergeometric functions (see, for example, Whittaker and Watson, 1952, p. 281).

It might be hoped that the procedure for calculating I and s could be simplified by using recurrence relations between the appropriate hypergeometric functions. This would certainly be true if an extensive series of calculations are to be performed, where contiguous values of k or g as well as of q are required. It seems most convenient experimentally, however, to select only a few values of k and g for special study, and since the relevant theory would be straightforward to develop it will not be given here.

Stage IV

Interpreting the single response of a subject

In order to obtain the information I from the appropriate I, s curve, a knowledge of s is required. The precision with which s is known increases with the number of experiments a subject performs; for moderate accuracy 20 experiments would need to be carried out. The final step in our program enables us to make some estimate of I when only one experiment is carried out with a subject. We shall suppose there are a large number N of subjects and that $\phi(u)\delta u$ subjects possess values of s which lie between $s = u$ and $s = u + \delta u$. Of these subjects, a fraction p_t of them will make t correct responses. We should strictly write $p_t(u)$ to indicate that the values of p_t refer to the case of $s = u$. It is important to realize that we cannot reverse the argument; that is, we cannot estimate the probability of a subject's value of s lying between u and $u + \delta u$ on the basis of the number of correct responses he makes in one experiment.

Suppose we make the assumption that the number of correct responses is a true measure of s. Then the information which the $\phi(u)\delta u$ subjects will be adjudged to give is

$$\sum_t p_t(u)\phi(u)\delta u(\log A_t + \alpha_t t) = \phi(u)\delta u \sum_t I_t p_t(u) \quad (17)$$

where the subscripts to A, α, and I indicate that their values are appropriate to $s = t$. The information that would be ascertained per experiment if a large number of experiments with each subject were carried out is of course

$$\phi(u)\delta u(\log A_u + \alpha_u u) = \phi(u)\delta u I_u \quad (18)$$

Expanding I_t in a Taylor's series, and using Lagrange's form for the remainder, we find

$$I_t = I_u + (t - u)(dI/ds)_u + \tfrac{1}{2}(t - u)^2(d^2I/ds^2)_x$$

where the subscripts indicate the values of s at which they must be evaluated, and x lies between t and u. Thus

$$\begin{aligned}\sum_t I_t p_t(u) &= \sum_t I_u p_t(u) + \sum_t p_t(u)(t - u)\alpha_u \\ &\quad + \sum_t p_t(u)\tfrac{1}{2}(t - u)^2(d^2I/ds^2)_x \\ &= I_u + \tfrac{1}{2}(d^2I/ds^2)_x \sum_t p_t(u)(t - u)^2 \qquad (19)\end{aligned}$$

Now from (6) and (13),

$$\begin{aligned}\delta s = \sum_r r\delta p_r = \sum_r (r - s)\delta p_r &= \delta\alpha \sum_r (r - s)^2 A e^{\alpha r}/\Omega_r \\ &= \delta\alpha \sum_r (r - s)^2 p_r \qquad (20)\end{aligned}$$

while from (15) we have $d^2I/ds^2 = d\alpha/ds$. Thus

$$\sum_t I_t p_t(u) = I_u + \tfrac{1}{2}(d^2I/ds^2)_x(d^2I/ds^2)_u^{-1}$$

On the assumption that each subject makes his most probable response, we should find that the average information transmitted per subject is

$$\frac{1}{N}\int_0^q \phi(u) \sum_t I_t p_t(u)\,du \qquad (21)$$

This exceeds the true average, namely

$$\frac{1}{N}\int_0^q I_u \phi(u)\,du \qquad (22)$$

by the amount

$$\frac{1}{2N}\int_0^q \left(\frac{d^2I}{ds^2}\right)_x \left(\frac{d^2I}{ds^2}\right)_u^{-1} \phi(u)\,du \qquad (23)$$

which is always positive. $(d^2I/ds^2)_x$ may be greater than, equal to, or less than $(d^2I/ds^2)_u$ over the different ranges of the integration; if we make the approximation that they are equal, (23) reduces to $\frac{1}{2}$, or 0.72 bits. We may test the accuracy of this result by choosing special forms for $\phi(u)$ and actually working out (21) and (22) and then subtracting the two. If $\phi(u)$ is a delta function, i.e. if all subjects possess the same value of s, the integrations are trivial. For $k = 90$, $g = q = 15$, and $s =$

8.9159, we find (21) is 9.67 and (22) is 9.16. The difference is 0.51, very close to ½. It is clear that by replacing $(d^2I/ds^2)_x$ by $(d^2I/ds^2)_u$ we have assumed that (d^2I/ds^2) does not vary with the place of evaluation; in other words, we have approximated the I, s curve in the neighborhood of $s = u$ by a parabola. This approximation will be very bad when d^2I/ds^2 is infinite, that is, when the subject's responses would be *always* either completely correct or completely incorrect. Fortunately the experimental conditions can usually be arranged so that these extreme cases do not occur.

We can therefore summarize this section by stating that if the number of correct responses a subject makes is interpreted as giving his value of s, the information so obtained, J, is related to the true information by the approximate equation

$$I = J - \tfrac{1}{2} \tag{24}$$

The use of this equation will be illustrated later by an example.

3. RELATION TO EXPERIMENT

Discussion of Assumptions

Since the preceding formulas for I have been developed primarily to assist in the evaluation of results in psychological experiments on recognition and recall, the question is now asked how far the assumptions we have had to introduce at each stage are likely to hold in such experiments.

Assumption (1), that the numbers k, g, and q are the same for a series of experiments, can be made to hold in the following way: k and g are directly under the experimenter's control, and q can be held the same either by instructing a subject always to give the same number of responses or by allowing him to make different numbers of responses and performing the calculation of I separately for classes of experiments yielding different values of q. Having done this, it is then meaningful to compare the amount of information conveyed for different values of q.

Similarly, assumption (2) can be satisfied by the way in which the experimenter selects symbols to be sent and by instructing a subject not to repeat symbols in his responses or by disregarding any repetitions he makes. Assumption (3) may in practice not be satisfied in recall; but since in recognition a subject is prevented from giving order information about the list presented (G), achievement in recognition and recall can only be meaningfully compared if order information transmitted in recall is discounted as it is by the present method.

Assumption (4), that the experiments should constitute an ergodic sequence, is basic to any application of information theory and will not be further discussed here. It is possible that it would only be realized in practice if a subject was given considerable pretraining on the task: if a long series of experiments is performed it is possible to test for whether the assumption holds.

Assumptions (5), (6), and (7) concern the probabilities of receiving and sending symbols and the absence of correlations between symbols sent and symbols received. In so far as they affect the source, they can be satisfied by the method used by the experimenter to select symbols to be presented to the subject in G. It seems likely that the assumptions will only hold of the symbols selected in Q if symbols of low meaningfulness are used: a subject should not then be more likely to select one than another. It would also be necessary that intra-list similarity be low; otherwise a subject might tend to recall a symbol similar to one of the symbols sent, and this would render assumption (7) invalid; e.g. if a symbol BA were sent, a subject might be more likely to recall AB or BE than (say) RE. Two further points are worth making on these assumptions: (1) It is possible to test whether or not they hold; (2) If we are interested in comparing relative amounts of information conveyed under different conditions rather than in computing the absolute amount of information, deriving I in the way outlined above may still give a good method of doing this even if these three assumptions are not fully satisfied. Thus failure of assumption (7) would only give a false result about *relative* amounts of information conveyed under different experimental conditions if under different experimental conditions the ratio of "near misses" to correct responses made by a subject were to alter.

Assumption (8) enables us to evaluate I without first computing the values of p_r from a long series of experiments. This assumption is likely to be valid for the following reasons. First, it is essential that any general expression for the probabilities p_r should include both the random case, where $p_r = 1/\Omega_r$, and the case where the subject makes as many correct responses as possible. Secondly, it is highly desirable that the values of p_r for other cases should correspond to imposing as few constraints upon the system as possible. A demand, for example, that all the p_r values should be zero except one not only fails to include the random case but also presupposes that the receiver makes r correct responses and $q - r$

incorrect responses with certainty. In an actual experiment the subject would never make responses which he is sure are incorrect. The most unrestrictive constraint which we can impose is that the average value of r over a large number N of experiments, namely s, should have a prescribed value. In other respects we assume that the subject's responses are random, i.e. all ways in which he can achieve Ns correct responses are equi-probable. Since randomness is directly related to a lack of information, this condition corresponds to finding values of p_r which minimize I. The assumption that the set Q is random subject to s having a prescribed value seems reasonable in the absence of further information concerning the behavior of the subject. The following argument demonstrates that this assumption is in fact equivalent to assumption (8). The number of ways a subject can make Ns correct responses in a series of N experiments in which he gets r correct n_r times is

$$W = \frac{N!}{n_0!n_1! \cdots n_q!} \prod_r \left[\frac{g!(k-g)!}{r!(g-r)!(q-r)!(k-g-q+r)!}\right]^{n_r}$$

On the assumption that all ways are equi-probable, we can find the most probable distribution for the quantities n_r by maximizing W subject to the two conditions

$$\sum_r rn_r = sN$$

and

$$\sum_r n_r = N$$

For very large N the result is

$$n_r = NAe^{\alpha r}/\Omega_r$$

Since $p_r = n_r/N$, we again obtain (8), showing that the principle of minimizing I is equivalent to assuming that all ways a subject can achieve Ns correct responses are equi-probable.

An Example

A brief example of the application of the formulas here derived to a problem of psychology will now be presented. Some of the results of an unpublished experiment on recognition and recall will be quoted—a detailed account of the experiment is in preparation (Davis, Sutherland,

and Judd). Subjects were presented with a list of 15 symbols, selected out of 90 possible symbols, and were asked either to recall or recognize the symbols. There were four conditions—recall (out of 90 symbols), recognition out of a list of 90, recognition out of a list of 60, and recognition out of a list of 30. Thus in this experiment, $k = 90, 60$, or $30; g = 15$. The number of responses made by subjects was not restricted, and q varied between 5 and 17. Since there was only one experiment on each subject, it was necessary to apply Eq. (24) separately to the result of each subject to calculate information. For example, in the recognition condition out of 90, two subjects made 13 responses and got 9 correct. Using the method set out on page 17, we can plot the curve of I against s for the values $k = 90$, $g = 15$, and $q = 13$. We find that for $s = 9$, $I = 16.1$ bits. Applying the correction given in Eq. (24), the best estimate of I is $16.1 - 0.7$ bits $= 15.4$ bits per subject. By successive applications of this method for each subject's score, the average information transmitted per subject can be estimated for each condition. It transpired that subjects had transmitted an average of 12.5 bits in recall out of 90, 12.0 bits in recognition out of 90, 10.1 in recognition out of 60, and 7.0 in recognition out of 30. Under the four conditions the raw scores of the subjects were respectively 149/207, 181/303, 201/295, and 250/317, where the first figure indicates the total number of correct responses and the second the total number of responses for each condition. It is clearly impossible to make meaningful comparisons between the raw scores: the formulas here derived enable us to make a meaningful comparison between results obtained under different conditions in terms of the average amount of information transmitted per subject.

Further Applications

We consider now two further applications of the mathematical techniques: it will be obvious that some hypotheses which can be tested by the application of the techniques are themselves suggested by the mathematical development.

(1) In the experiment outlined above, when a subject was recognizing symbols out of a list of 30 presented to him, he only knew that his selection would be made from 30 and not 90 *after* the initial list (G) of 15 had been presented to him. It is of considerable interest to ask whether subjects would perform better in terms of information transmitted, if they knew that the symbols in G could only be selected from a list K of

30 *before* the list G was presented. Such an approach should throw light on how symbols are coded in the memory store.

(2) In recall a subject's probability of being right in responding with a given symbol is inversely related to the rank order of that symbol in the sequence of responses made in any one experiment. We can ask the question whether a subject is able to stop responding at the point which will maximize the information he conveys. Thus where subjects made q responses, we can consider the probability of each successive response being correct, and from a curve so obtained evaluate whether or not the subject would have conveyed more information if he had stopped responding either before or after the point at which he customarily does. Clearly the point at which a subject stops responding may be influenced by both experimental and personality variables which would be of considerable interest.

Acknowledgement

We wish to thank the director of the Oxford University Computing Laboratory for providing us with computing facilities. One of the authors (N.S.S.) also wishes to thank the Nuffield Foundation for their generous financial support: his share in the work was undertaken as part of a project on stimulus-analyzing mechanisms which is financed by the Nuffield Foundation.

Received: February 24, 1959; revised May 27, 1959.

References

Brown, J. (1959). Information, redundancy and decay of the memory trace. *In* "Symposium on the Mechanization of Thought Processes." H.M.S.O., London.

Davis, R. D., Sutherland, N. S., and Judd, B. R. (In preparation).

Good, I. J. (1957). Distribution of word frequencies. *Nature* **179,** 595.

Hunter, I. M. L. (1957). "Memory: Facts and Fallacies." Penguin, London.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. "*Communication Theory.*" W. Jackson, (ed.), pp. 486–502. Academic Press, New York.

Postman, L., Jenkins, W. O., and Postman, P. L. (1948). An experimental comparison of active recall and recognition. *Am. J. Psychol.* **61,** 511–519.

Purcell, E. M., and Pound, P. V. (1951). A nuclear spin system at negative temperature. *Phys. Rev.* **81,** 279.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27,** 379–423.

Whittaker, E. T., and Watson, G. N. (1952). "Modern Analysis." Cambridge Univ. Press, London and New York.