PART ONE

The Theory of Selective Information and Some of Its Behavioral Applications

By R. DUNCAN LUCE

UNIVERSITY OF PENNSYLVANIA

Contents

SECTION I. THE DISCRETE THEORY

SECTION II. APPLICATIONS TO BEHAVIORAL PROBLEMS

SECTION ONE

THE DISCRETE THEORY

1. INTRODUCTION(*)(†)

THERE IS a widespread belief – most forcefully articulated by Norbert Wiener [1948] – that we are undergoing a new scientific revolution, one comparable in scope and scientific significance to that of the last century; but where the dominant concepts in the previous development were energy, power, and efficiency, the central notions are now information, communication, and feedback. Many current problems stem from attempts to transmit information and to exercise effective control rather than to achieve an efficient use of energy; little more than chaos would result, for example, were the design of a high-speed computer approached from the energy standpoint. "Information is information, not matter or energy. No materialism which does not admit this can survive at the present day." (Wiener [1948], p. 155).

What then is information? How is it measured? What scientific statements can be made using the term?

Several schools of thought have developed, each formulating and restricting these questions in its own way and offering answers to the resulting and more specific questions. In this essay I shall examine the formulation and the answers of one of these schools and describe some of the impact it has had for certain problems of psychology.(**) But before we turn to this, a certain amount of background material on the history, orientation, and relation of information theory to other theories is appropriate.

(*) Most often the title "information theory" is used without the prefix "selective"; however, some feel that the simpler title is misleading, especially since there exists a theory of structural information and one of semantic information. Indeed, Bar-Hillel [1955] feels the title should be the "theory of signal transmission" for, as he argues, the seductive word "information" has led to considerable confusion; however, it is probably now too late to make such a crucial psychological change.

(†) I wish to express my appreciation to Professors A. H. Hastorf, W. E. Hick, B. Mandelbrot, F. Mosteller, H. Quastler, and H. Raiffa for reading and commenting on the original version of this essay. Many of their suggestions have been incorporated into the present version.

(**) A number of summaries of this theory have been given: Gabor [1953b], Hockett [1953], McMillan [1954], Miller [1953], Osgood [1954], Slepian [1954] and Weaver [1952].

It is clear that if Wiener and others are correct in their views, the intuitive concept "information" must be given at least one precise meaning and maybe more. Considering the variety and vagueness of its meanings in everyday usage, it is *a priori* certain that objections will be raised against any particular formulation, which will surely ignore some of these meanings. This problem — if it be such — has been met many times in science; we need only think of words and concepts such as force, energy, work, etc. It is doubtful that a formal definition ever stands or falls because of such debates; it is rather the power and depth of the resulting theory that determines its ultimate fate.

Within the last two decades two distinct attempts have been made to deal with the notion of information, one in Europe, and one in America; these have been complementary rather than competitive. Both theories seem to have arisen from much the same class of applied problems: communication involving electrical signals. The European school, in which the names of Cherry, Gabor, and MacKay are the most important, has been concerned with the problem of the information contained in a representation of a physical situation. As seems intuitively reasonable, the concepts of size and dimensionality are important here. In America, because of work by Wiener and Shannon, a theory of information transmission has been developed in which the dominant concepts are selection, statistical possibilities, and noise.

In this essay I shall not undertake an examination of the notions of structural and metrical information (the European school). This theory has had, so far as I can determine, almost no effect on behavioral applications. Of interest to the behavioral scientist, however, is the apparently overlooked fact that one basic concept of structural information theory is identical with the central assumption of factor analysis. Both theories are concerned with the number of independent dimensions that are required to represent a certain class of data, and the geometrical model of any particular situation is as a point in an Euclidean n-space. If this observation is correct, it is interesting that basically the same concept has been independently arrived at by both physicists and psychologists, and it may be unfortunate that each is unaware of the work of the other.

There are, of course, marked differences of emphasis which reflect the different origins and problems. For example, the European information theorists have examined the basic natural units in which the several dimensions can be scaled. Whether this theory of metrical information, as it is called, is related to any of the scaling work in the behavioral sciences is not

immediately obvious and appears not to have been investigated. There is at least a superficial parallel to psychophysical scaling based upon imperfect discrimination. On the other hand, factor analysts have developed an elaborate matrix machinery suited to determining the approximate dimensionality of the Euclidean space representation of certain types of data. A comparable machinery does not appear to exist in structural information theory, though, of course, the close relation of the structural model to matrix theory is apparent.

Our concern, however, is with selective information theory. The central observation of this theory is that for a great many purposes — in particular, in the design of communication equipment — one is never concerned with the particular message that is sent but rather with the class of all messages that might have been sent and with the probability of the occurrence of each. "We are scarcely ever interested in the performance of a communication-engineering machine for a single input. To function adequately it must give a satisfactory performance for a whole class of inputs, and this means a statistically satisfactory performance for the class of inputs which it is statistically expected to receive." (Wiener [1948], p. 55) From this point of view, information is *transmitted* by a selection from certain alternatives. The contention is that selection of an *a priori* rare event conveys more information to the receiver than does the selection of one that is more probable. This use of "information" obviously ignores all questions of meaning. "It is important to emphasize, at the start, that we are not concerned with the meaning or the truth of messages; semantics lies outside the scope of mathematical information theory."(*) (Cherry [1951], p. 383) The failures to adhere to this position, and the consequent difficulties, are discussed in detail by Bar-Hillel [1955]; they have led to considerable confusion and not a little empty debate.

It may be useful to introduce at this point three common-sense observations which will be given precise meanings in the presentation of the theory of selective information — precise to the point where numbers can be attached to them.

1. A person communicating over a noisy telephone line can get less "across" in a given period of time than he can over a perfectly clear line.

(*) Carnap and Bar-Hillel [1952] and Bar-Hillel and Carnap [1953] have presented a theory of semantic information which is based on Carnap's work in inductive logic. Since their approach is different from that of selective information theory, and since, as far as I know, there have been no behavioral applications of it, I have elected not to summarize it here. It may, however, become important, and should therefore not be neglected by the serious student of this area. Also, see Hockett [1952].

2. Not every letter,(*) nor indeed every word, of a message in any natural language is as important as every other one in getting the sense of the message. For example, the missing letter in "q_iet" or the missing word in "many happy _____ of the day" can be filled in, with a high probability of being correct, by anyone knowing English, and therefore in the above context they do not carry much important information.

3. Every person seems to have a limited capacity to assimilate information, and if it is presented to him too rapidly and without adequate repetition, this capacity will be exceeded and communication will break down.

As they stand, it is not immediately obvious that these statements are not concerned with semantics, or, for that matter, that the whole problem of information transmission is not almost wholly semantic. One major contribution of selective information theory is in showing that much of what is implied or suggested in these examples and others like them can be given a precise and useful meaning by a purely statistical treatment.

We shall delve into this more deeply in the following sections; but first, let me discuss briefly some of the origins of the theory and of the developing interest of behavioral scientists in it.(†) Electrical communication engineers gradually had been gaining experience in the handling and transmission of information since the early days of the telegraph, telephone, and radio, and during the 1920's this experience began to be formalized as a theory. One of the most important early papers was by Hartley [1928]; in it the logarithmic measure so characteristic of modern information theory was employed in a simple form and much of the terminology was introduced. The maturation of the theory, however, resulted from the work of two men: Norbert Wiener of MIT and his former student C. E. Shannon of the Bell Telephone Laboratories. Shannon's papers of 1948 (reprinted in book form, Shannon and Weaver [1949]) are now the classic formulation of the theory.

(*) Here, and elsewhere, I shall speak as if the letter is the carrier of information: there will be presented calculations of the number of "bits of information transmitted" per letter, etc. The linguist may quite properly raise objections to this usage, for presumably it is the spoken, not the written, language that determines the information bearing units. Much effort has been expended in recent years to isolate and to understand the natural unit of spoken language — the phoneme — and it is in terms of this unit that we probably should deal. For a survey of this work and an extensive list of references see Osgood [1954]. Yet, for reasons of convenience — both because letters are more familiar to me and to many readers and because many of the existing information theory calculations are in terms of letters (the exceptions being Cherry, Halle, and Jakobson [1953] and Black [1954]) — I shall ignore this basic proposition of modern linguistics. Of course, this is not intended as a scientific position on the matter.

(†) A much more complete history of both the American and European schools has been given by Cherry [1951, 1953].

The more mathematically inclined reader will find McMillan's later [1953] presentation of the central theorems more satisfactory. Also, see the formulations of Feinstein [1954] and Watanabe [1954].

The implications of the theory and of several related concepts — of which feedback(*) is one of the most important — were quickly recognized to extend beyond improved electrical communication. Shannon and Wiener realized this, and the latter in his book *Cybernetics* both outlined the extent of the new discipline and offered a generic title for its somewhat nebulous components. From 1941 on, these concepts and theories have been examined and debated in a series of conferences and seminars.(†) For the most part, these meetings have been held in the East, many of them in Cambridge, and as a consequence the impact of information theory, which has been so strong along the Eastern seaboard, has been less marked in the West.

Many of the empirical sciences dealing with human behavior — psychology, linguistics, physiology, biology, psychophysics, social psychology, neurology, medicine, anthropology — have had representatives at these seminars; indeed, scientists from these fields have organized and dominated many of the meetings. From them emerged a small group of analytically inclined behavioral scientists who believe that information theory is, or can be, a useful tool in handling some problems in various disciplines. I shall try to indicate some of the uses, and the usefulness, of the theory in Part II of this essay.

Our material is organized into two parts. In the first, I shall try to present a synopsis, which draws heavily upon one's everyday experience with communication systems, of the discrete theory of selective information. The presentation is most deeply influenced by Shannon's. I was strongly

(*) "Feedback" has become such a familiar term that it is probably not necessary to define it, especially since it will not be a central notion in this survey. Still, a few suggestive words may do no harm. Many systems are designed, or behave as if they were designed, to respond to a certain class of inputs in such a manner as to achieve a particular goal. For example, an ideal amplifier attempts to reproduce the exact form of the input while changing the amplitude scale. A device designed to do this will, because of variability of its components, etc., fail to respond perfectly, and the problem arises how to improve the performance. One way is this: build into the system certain appropriate adjustable parameters, whose values are determined at any particular time so as to reduce the discrepancy between the desired output and the actual output. This is effected by feeding back a fraction of the output signal and comparing it with the input to determine the discrepancy. Under certain conditions, the resulting system will be stable and large errors will not occur. In a more general context, feedback is taken to mean any messages a system receives informing it as to what *its* response has been, and usually this information is used to modify its behavior to reach a specified goal.

(†) In his introduction to *Cybernetics*, Wiener presents a detailed history of the early meetings.

tempted to depart from his organization completely and to formulate the theory along the lines of a multivariate statistical analysis, following McGill's work, for it is this aspect of information theory that seems most pertinent in behavioral science applications. However, such a course would have reduced the number of familiar signposts available, left much of the language of the theory unmotivated, and rendered some of the applications close to incomprehensible. With some regret, I have elected the well trodden path. In the second part I shall be concerned entirely with applications of the theory to problems in psychology. Again, two modes of organization are possible: either by the conventional categories used in psychology or by the structure of the theory. While the latter more strongly appeals to me, it tends not to seem appropriate to most psychologists — the substantive rather than the methodological boundaries are held sacred — so again I have conformed. I hope one day to see a monograph on the use of information theory in psychology which follows the two courses not used here.

An appendix giving a short summary of Shannon's theory of continuous communication systems concludes the essay. While this theory is of great importance in electrical applications, it has so far been of minor significance for traditional problems of the behavioral sciences. However, it is interwoven into certain new work; see Licklider [1960].

2. GENERAL CONCEPTS

Communication Systems. Information transmission always occurs within a certain physical framework which may be termed a communication system. Basically such a system consists of three central parts: a *source* of messages, a *channel* over which the messages flow, and a *destination* for the messages. The source, which very often is a human being, generates messages (and so information, see below) by making a series of decisions among certain alternatives. It is the sequence of such decisions that we call a message in a discrete system. These messages are then sent over the channel, which is nothing more than an appropriate medium which establishes a connection having certain physical characteristics between the source and the destination. Mechanically, this picture is incomplete, since the decisions made by the source must be put into a form which is suitable for transmission over the channel, and the signals coming from the channel must be transformed at the destination into stimuli acceptable to it. Thus, between the source and the channel a *transmitter* is introduced to "match" the channel

to the source, and between the channel and the destination a *receiver* is introduced to "match" the channel to the destination. In other words, the transmitter encodes the message for the channel and the receiver decodes it. A schematic diagram of the system is shown in Fig. 1.

FIG. 1.

It is entirely possible to have transmitters which so encode messages that it is not possible to design a receiver which can completely recover the original message. For example, if one has a transmitter which encodes all affirmative statements such as "O.K.," "yes," "all right," etc. into the same signal, then no device can be built which will translate that signal back into the particular word chosen by the source. A transmitter having this property is called *singular;* otherwise it is called *non-singular.* (These terms arise if one thinks of the transmitter as a many-many transformation or as a one-to-one transformation.) When the transmitter is non-singular it is possible to design a receiver which will completely recover the original message. In other words, there exists a receiver which is the inverse of the transmitter. Throughout our discussion we shall assume that the transmitter is non-singular and that the receiver is its inverse. In effect, this means that we can ignore them in our discussion and suppose that the source and destination are both matched to the channel.

Our abstract communication system seems fairly complete except that it does not allow for the possibility that more than one source may be using the same channel at the same time. Certainly this can happen. It occurs when, by mistake, one telephone line carries two conversations at once (crosstalk). It also happens in telephone or radio communication when there is static in addition to the desired message. In all such cases the messages from sources other than the one under consideration — which will simply be called *the* source — cause interference with messages from the source. Such interference may be minor and have no effect on the intelligibility of the message, as for example in the usual low-level telephone static, or it may be most destructive, as when another conversation is cut in. Another example which one might tend to put into the same category of interferences is the 60-cycle hum which is common to so many cheap radios. If the hum level is high enough it certainly can lower the intelligibility of speech. However, there is an important difference between the problem of inter-

ference from hum and from that due to static or to other conversations. The former is completely predictable and it is possible from a short sample to determine its exact frequency, phase, and amplitude. Thus, if there is hum, one can build into the transmitter or into the receiver a network to subtract it from the resulting signal, leaving only the message. Static, hiss, and crosstalk cannot be predicted in detail from any amount of past evidence about them. Therefore, once they enter the channel, they cannot be characterized in full and subtracted from the signal, but rather they must be accepted and compensated for in other ways.

Thus in our abstraction we must conceive a second source (which in fact may be several lumped together) also feeding signals into the channel. It has the property that (for the problem under consideration) neither *the* source nor the destination can predict in detail the messages that will emanate from it. The source or the destination may have or may obtain statistical data about the nature of this second source. For example, in an electrical communication system the average power of the second signal may be measured. Such a source is known as a *noise source* and the signal it generates will be called *noise*. Clearly, these are often relative terms and what in one context is noise may be the message in another. This, then, completes our model of a communication system, and it is shown schematically in Fig. 2.

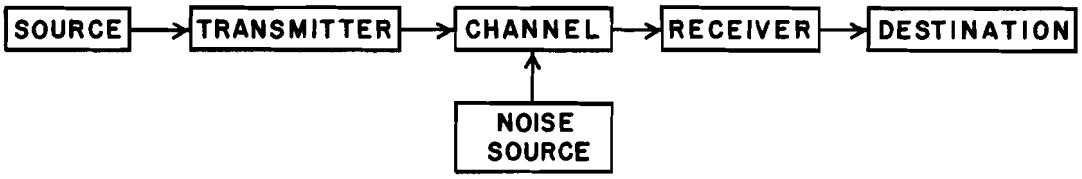

FIG. 2.

When there is a noise source in a system it is conventional to speak of the channel as being noisy, but it is well to keep in mind that this is merely an abbreviated, and slightly misleading, way of speaking. The noise signal is not an invariant of the channel, as are its physical characteristics. It is clear that one can change the amount of noise in a system while keeping the physical characteristics of the channel, the source, and the destination the same. In any given problem under consideration, the noise level will presumably remain constant and so it can be thought of as a property of the channel, but as we shall see it is a property which must be handled very differently, in the theory, from the physical characteristics of the channel.

Noiseless Systems. In one sense, no communication system is ever noiseless; there is always some noise signal. For example, in any electrical system

there must always be signals resulting from the random agitation of molecules — thermal noise. This can be a serious problem in a high-gain amplifier, but it is not in a telegraph. The point, of course, is that noise is not, in and of itself, bad, but only when it causes a significant interference in the messages sent by the source. The only pertinent feature of noise is whether it ever causes the destination to infer that a different message was sent from the one actually sent. Thus, if the noise level is low compared with the signal level, so low that it does not *significantly* alter the message as it passes along the channel, then it may be completely disregarded and the system can be treated as if there were no noise present.

Since it is assumed that by definition the effect of noise is unpredictable in advance (except statistically), all we shall be able to state about the effect of noise on messages — and all we need to state — is the probability that it changes one signal into another. If the signals sent (in a given situation) are always received correctly, then we say the system (or the channel) is *noiseless*. It must always be kept in mind that if we change the level at which the transmitter operates, or the level of the noise signal, we may change a noiseless system into a noisy one. Being noiseless is a property of the whole system and not of the channel alone!

In principle, it is not necessary to deal separately with the theory of the noiseless and noisy cases, for the former is but a special case of the latter. The presentation, however, is simpler if we bring in the complications one at a time, so we shall examine the noiseless case first (Chapter 3) and then the noisy one (Chapter 4).

The Bit — a Unit of Information Transmitted. To carry out the program mentioned in the Introduction, namely, to make precise and measurable some features of the transmission of information, it is necessary to introduce a unit in terms of which amounts of information transmitted may be measured. The central observation which is needed before one can arrive at an appropriate unit is that a message conveys information in the sense of reducing uncertainty only by its relation to all the other messages that might have been received. Suppose a person is asked whether he smokes. If we have no prior information other than population statistics on smoking, then all we know is the probability that he, as a random selection from the population, will answer "yes" or "no." When he selects one of these alternatives and transmits it, some information has been conveyed. But if it is known *a priori* that he does smoke, e.g., from previous conversations or from seeing him smoke, then with probability 1 the answer will be "yes" and

the receipt of "yes" from him cannot convey any (new) information. In effect, our prior knowledge reduced the set of possible messages to a single element, and so far as we are concerned there was no choice to be made. Thus, no information could be transmitted.

The minimum condition, therefore, under which information can be transmitted is that of a choice between two alternatives. The maximum uncertainty in such a choice exists when the two alternatives are equally probable. Hence the maximum information is conveyed by a choice between two alternatives when they are equally likely. We take such a choice to be one unit of information. That is, whenever a choice is made between two *a priori* equally likely alternatives (no matter what they are) we shall say that one unit of information has been transmitted by the choice. According to Shannon, Tukey proposed that the unit be called a *bit* — a shortened form of binary digit — and that term is commonly used. Goldman [1953] prefers the term "binit" in order to avoid such expressions as "a bit of information" which, unfortunately, has quite another everyday meaning, but I shall conform to common usage. All statements about information transmission, therefore, will be given in this unit; we shall speak of so many "bits in a message," or the "bits transmitted per second," or the "bits per English letter," etc.

With this as the definition of the unit, the next problem is to say just how many such units are transmitted when a selection is made from an arbitrary finite set with an arbitrary *a priori* probability distribution over it; and just how many units are transmitted when several selections are made. Certainly, one wants to require at least this: if two independent choices are made between *a priori* equally likely alternatives, then a total of two bits of information are transmitted. More generally, we shall impose the condition that whenever two statistically independent selections occur, the total information transmitted is the sum of the amounts transmitted by each of the selections, i.e., the measure of information transmitted shall be *additive.*

As an example of how the additivity condition and the bit may be used, consider a set of elements (think of them as letters of an alphabet or phonemes in a phonemic system) in which each element is equally likely to be selected. (This, of course, does not hold for any natural language.) Further, suppose that the number n of elements is of the form 2^N, where N is an integer. Question: when an element is chosen from this set, how many bits of information are conveyed? The answer is N bits per selection. We can easily show that there are no more than N bits. Let any element be selected

and divide the set of elements into half, each half being composed of 2^{N-1} elements. The element selected is in one half or the other, and the information transmitted as to which half it is in is a decision between two equally likely alternatives (since each element has the same probability of being chosen). So, that conveys one bit of information. Take that set and divide it in half, each half now consisting of 2^{N-2} elements. Again, the decision as to which of the two sets contains the selected element is between two *a priori* equally likely alternatives, and so another bit of information is transmitted in isolating it. Continuing the process until the element is isolated clearly requires N steps, and, assuming additivity, N bits of information are transmitted. The fact that all the elements were assumed to be equally likely should suggest that no scheme can be devised to isolate the element in fewer than N binary decisions. This can be proved to be the case. I shall not prove it, for the conclusion that there are N bits per selection in this situation will follow from much stronger and deeper results to be presented later.

The English alphabet consists of 26 letters which with punctuation marks comes to about $32 = 2^5$ symbols. Were we to suppose them to be chosen independently and with equal probabilities (both patently false assumptions) then each letter of a message would yield five bits of information. Clearly, this is not an accurate estimate of the bits per letter in English prose. However, it does stand as an upper bound to this number. Later (Chapter 7) more precise estimates will be given which show that it is actually somewhere between 1 and 2 bits per letter.

Continuing with the example, observe that when $n = 2^N$, then $N = \log_2 n$ by definition of the logarithm, and so we may say that in this situation there are $\log_2 n$ bits of information per element. We will find that our subsequent discussion of information transmission results in logarithmic measures slightly more complicated than this.

3. THE DISCRETE NOISELESS SYSTEM

IN THIS CHAPTER I shall discuss what is known as the discrete noiseless communication system. The definition of a noiseless system was given in the last section, and it may be summarized by saying that in such a system there is never any confusion at the destination as to which signal (of a known class of signals) was emitted by the source. This, of course, does not mean that the signal received is necessarily physically identical to the signal sent, but only that no confusion can arise as to what signal was sent.

The word 'discrete' refers to the nature of the information source. It describes a source which generates messages by temporally ordered sequences of selections from a finite set of possible choices. Thus, the discrete case includes a vast amount of familiar communication, such as the selections made from a phonemic system to generate words and sentences. But the theory of this section does not include sources, such as a musical instrument, which *can* select from a continuum of continuous functions; that theory is outlined in the appendix.

Channel Capacity. In any communication system the transmitter is so chosen as to match the source to the channel. Signals emanating from the transmitter, which are assumed to be in one-to-one correspondence with the selections made by the source, are propagated along the channel. As far as this communication process is concerned, the relevant effect of the physical characteristics of the channel is to determine how many different signals can be transmitted over it in a given space of time. Roughly, this is what we mean by the capacity of the channel. Formally, let $N(T)$ denote the number of different signals which satisfy the following three properties:

i. each signal can be emitted by the transmitter as a result of selections by the source,

ii. each signal is admissible on the channel, i.e., each signal is compatible with the physical characteristics of the channel,

iii. each signal is of duration T time units.

From the discussion in the last section, it is suggested (though by no means proved) that if each of these $N(T)$ signals were equally likely then there would be $\log_2 N(T)$ bits per signal of duration T time units, or

$$C(T) = \frac{\log_2 N(T)}{T}$$

bits per signal per unit time. Now, extending the discussion of the two-alternative case, it is plausible to suppose that the maximum information is transmitted when each signal is equally likely. Since we have taken $N(T)$ to be the largest number of different signals which may be transmitted over the channel in T time units, it is therefore reasonable to suppose that $C(T)$ is approximately the maximum number of bits of information per signal transmittible over the channel in one time unit. Since there can be only one signal on the channel at a time, $C(T)$ is approximately the maximum number of bits that can be handled by the channel in one unit of time. The

approximation will tend to be better the larger we take T, so we are led to define the capacity C of the channel to be:

$$C = \lim_{T\to\infty} C(T) = \lim_{T\to\infty} \frac{\log_2 \mathcal{N}(T)}{T}.$$

For any practical application of this concept the trick is to determine $\mathcal{N}(T)$ from the physical characteristics of the channel or from any theorems we may derive which involve C. In the following subsection an example of the first procedure is given, and in a later subsection a theorem is given which has been used to find approximations to C empirically.

A Special Case of Channel Capacity.(*) For the moment let us restrict ourselves to a special class of transmitter-channel combinations which, possibly, is best illustrated by the familiar dot-dash telegraphy code. Suppose that at any instant there either is or is not a signal on the wire connecting the transmitter to the receiver. A dot will be represented by one time unit of signal and one time unit of no signal, and a dash by three units of signal followed by one unit of no signal. Between letters three units and between words six units of no signal are allowed. Problem: compute the channel capacity.

For this system, let us define two different states which we shall call a_1 and a_2. The system is in state a_1 following either a letter or a word space, and it is in state a_2 following either a dot or a dash. Since a word or letter space can never follow either a word or letter space, we know that the next signal after the system is in state a_1 must be a dot or a dash, so state a_1 must be followed by state a_2; however, when the system is in state a_2 it can be followed by any of the four possibilities and so by either state a_1 or a_2. This is illustrated schematically in Fig. 3.

We are now in a position to generalize this in a natural manner to a system having m possible states $a_1, a_2, \ldots, a_m$ and n possible signals $S_1, S_2, \ldots, S_n$. When the system is in state a_i only a certain subset of the signals may arise; let S_s denote a typical one. We suppose that a_i and the admissible S_s together determine what the next state will be. Let us denote it by a_j. For all such possible triples (i,s,j), let $b_{ij}^{(s)}$ denote the time duration of the s^{th} symbol. Obviously certain of the combinations cannot arise, e.g., in the telegraphy case the triple (a_1, word space, a_2) is not admissible (see Fig. 3).

(*) This subsection is not essential to the rest of the paper, and, as it is a little more difficult, some readers may choose to omit it.

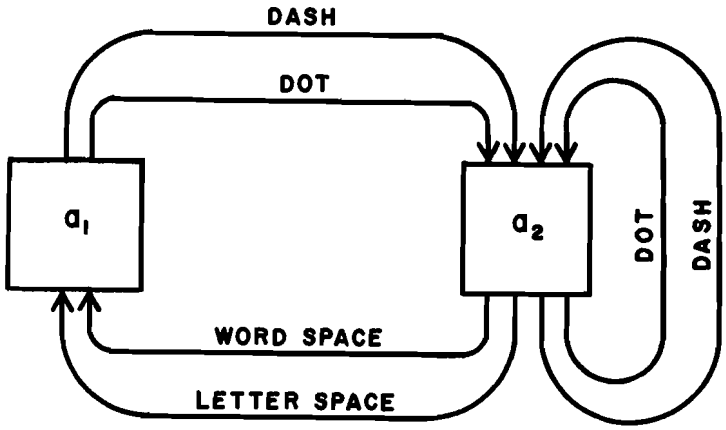

FIG. 3.

On the other hand, (a_1, dash, a_2) is admissible and its b value is four time units.

The channel capacity of this system can be shown (Shannon, [1948]) to be given by

$$C = \log_2 W_o,$$

where W_o is the largest real root of the determinantal equation

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0,$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

In the telegraphy case, the graph of Fig. 3 can be put in the following matrix form:

		Next State: a_1	Next State: a_2
Present State	a_1	—	dot or dash
	a_2	letter or word space	dot or dash

From this we see that the determinantal equation reads

$$\begin{vmatrix} -1 & W^{-2}+W^{-4} \\ W^{-3}+W^{-6} & W^{-2}+W^{-4}-1 \end{vmatrix} = 0$$

$$= \frac{1}{W^{10}}[W^{10}-W^{8}-W^{6}-W^{5}-W^{3}-W^{2}-1].$$

Solving for W_o and computing $\log_2 W_o$ one finds that $C=0.539$ bits per unit time.

More will be said about channel capacity before we are done, but first it is necessary to discuss the source and to develop a suitable measure for the average information generated by any discrete source.

The Discrete Source. As I have said, it is assumed there is a source that makes selections (with replacement) from a finite set of elements and that messages are generated by temporally ordered selections from this set. The general situation is typified by the way we form written sentences by ordered selections of letters, blanks, and punctuation marks.

A moment's reflection about English will suggest two important statistical facts about many sources:

i. There is no reason to suppose that the probability that one symbol will be selected is the same as that for another symbol: the letter "z" is much less frequently used in English than is "e".

ii. In general, the choice of one symbol in the middle of a message will not be independent of the preceding choices: although "e" has a high *a priori* probability of being chosen, the probability is markedly reduced if the letters "automobi" have already been received and it is markedly increased if the letters "automobil" have been received.

Although most human sources produce an interdependence between symbol selections – often called intersymbol influences – there are some cases of independence, such as the transmission of random numbers or of an unconnected set of telephone numbers. In the next subsection we shall analyze the case of independent selections and later the more complicated case where there are dependencies.

To deal with these problems of symbols selected with different frequencies and of the interdependence of symbol selection, we shall obviously want to introduce probability distributions over the set of symbols. For this to make sense, we shall have to assume that the source is homogeneous in time, so that its statistical character – measured by any statistical parameter we choose – is the same at one time as at any other time. Such a source is

said to be *stationary* and the time series (of symbol selections) is called a stationary time series. This assumption is essential to the theory; it is one which seems plausible for many sources and not for others. For example, it does not hold for an individual who is learning. In most cases, however, it is quite difficult to assure oneself that a source is stationary. The problem is very closely related to the difficulty in deciding whether a particular finite set of numbers can be considered a typical sample from a random sequence generated by some probability law. The condition does serve, however, to prevent us from considering as one source the *New York Times* from time O to time T and *Izvestia* from time T to time T', for the statistical structure of messages in these two time intervals will certainly be different – indeed, some of the symbols will differ.

Assuming a stationary source S, we may now introduce a little necessary notation. We let $p(i)$ denote the probability that symbol i in S will be selected and $p(i,j)$ the probability that symbols i and j in S will be selected in the order i and then j. In general, $p(i,j) \neq p(j,i)$ (consider, for example, q and u in English). In general, if $i_1, i_2, \ldots, i_k$ is an ordered sequence of symbols, $p(i_1, i_2, \ldots, i_k)$ denotes the probability of its occurrence.

The selection of symbols is said to be *independent* if for every k and every possible sequence $i_1, i_2, \ldots, i_k$

$$p(i_1, i_2, \ldots, i_k) = p(i_1)\, p(i_2) \ldots p(i_k).$$

Before turning to the analysis of the case of independent selections, let me indicate how messages look when generated according to various assumed statistical dependencies. I shall present the output generated from a source which takes into account some (but not all) of the statistical structure of English. First, suppose that selections are independent but with the simple frequencies of English text. Using these frequencies and a table of random numbers, Shannon [1948] generated

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

If, however, one includes some intersymbol influences, one may, for example, generate a message in which each selection depends on the two preceding ones. Using such data for English, Shannon generated

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Neither message is English, but the second is "more" English than the first. The greater ease a typist finds in copying the second passage as against the first reflects the difference.

A Measure of Information Transmitted for Independent Selections. Let us assume for the present that messages are generated by independent selections from a discrete source. Statistically, then, the source is completely characterized by the probability distribution

$$P=[p(1),p(2),\ldots,p(n)]$$

of symbol selection over the n symbols of the source S. The problem is to assign a number to the source, i.e., to the probability distribution P, which is deemed a suitable measure of the *average* amount of information transmitted when a symbol is selected from S. There are at least four ways to get to an answer (fortunately the same answer), and since each reveals something of the structure of the problem and since the resulting statistic is of such great importance, I shall present all four.

What we want is a function which assigns a number to each probability distribution; we may denote it by

$$H=H[p(1),p(2),\ldots,p(n)].$$

The first procedure, which is heuristic and easily remembered, rests on accepting the earlier argument that, when there are $n=2^N$ equally likely alternatives, a suitable measure of the amount of information is $N=\log_2 n$. Let us extend this definition to n equally likely alternatives where n is now any integer, i.e., we shall say there are $\log_2 n$ bits per selection from among n equally likely selections. Now, if we consider any event of probability $p=1/n$, then we may treat this event as one among n equally likely alternatives and so the information involved in its selection is

$$\log_2 n=\log_2\frac{1}{p}=-\log_2 p.$$

Finally, consider an event of probability p (not necessarily the reciprocal of an integer): it is plausible to extend the above definitions further and to say that $-\log_2 p$ bits of information are transmitted by the occurrence of

this event of probability p. Thus, for the given source S, the selection of symbol i, which occurs with probability $p(i)$, transmits $-\log_2 p(i)$ bits of information. We see that this has the very reasonable property that an occurrence of a very rare event transmits a great deal of information and an event with probability near 1 transmits almost no information. On the average, however, the amount of information transmitted is the expected value of a single selection from the source, i.e.,

$$H = -\sum_{i=1}^{n} p(i) \log_2 p(i) \qquad \text{bits/selection.}$$

The above expression is without a doubt the best known aspect of information theory, and there are reasons to believe that this formula has blinded some to the content of the theory. It is, of course, nothing more nor less than a statistical parameter defined for all distributions — one which is in some ways similar to the variance. It obtains meaning and value in only two ways: first, as it is given a meaning in a theory, and second, as it becomes a conventionally accepted way of summarizing certain phenomena.

Shannon called H the entropy of the source (or more properly of the distribution characterizing the source) because the same expression (with the opposite sign) arises in statistical mechanics and is called entropy there. There has been considerable controversy as to whether this is only a formal similarity, or whether physical entropy and information are two closely related phenomena. This is a point requiring careful and sophisticated discussion and a rather deeper knowledge of physics than I want to assume here. Certain authors have been displeased with the term "entropy" and they have used terms such as the "amount of information" or simply the "information", the "specifity", and the "uncertainty" of this source. It is hard to say which term is the most common and which is the least objectionable — these two surely not being the same. Certainly, neither "amount of information" nor "information" are acceptable, since they lead much too easily to misinterpretations. Most often I shall use either "amount of information transmitted" or "entropy", the former without intending any semantic overtones and the latter without a commitment as to the identity of this statistic with physical entropy. On occasions when "uncertainty" seems the more suggestive term, I shall use it too with the tacit understanding that it really means "average uncertainty."

The second procedure to arrive at H, which many feel to be both the simplest and most elegant, amounts to a rigorous formulation of the first

one.(*) The technique is to state four conditions which intuitively seem to be met by the concept of "the information transmitted when a symbol i is selected, given that the *a priori* probability of its selection was $p(i)$." From these conditions we shall derive the entropy expression; they are:

1. *Independence of irrelevant alternatives.*(†) The amount of information transmitted by a selection of i shall be a real number which depends only upon $p(i)$ and not upon the probability distribution over the other symbols. Thus, we may denote the amount of information transmitted by the selection of i by $f[p(i)]$.

2. *Continuity.* $f[p(i)]$ shall be a continuous function of $p(i)$, since one feels that a very small change in $p(i)$ should result in only a small change in the amount of information transmitted.

3. *Additivity.* If two independent selections i and j with probabilities $p(i)$ and $p(j)$ are effected, then the amount of information transmitted in the joint selection (i,j), which has probability $p(i)p(j)$ of occurring, shall be the simple sum of amount of information transmitted by each of the selections, i.e.,

$$f[p(i)p(j)] = f[p(i)] + f[p(j)].$$

4. *Scale.* In our discussion of the bit, we said that a selection with probability 1/2 shall convey one bit, so we assume

$$f(1/2) = 1.$$

Now, observe that if n is an integer, then repeated application of the third assumption yields

$$f(p^n) = nf(p).$$

Let $q^n = p$, then from the last equation,

$$f(p^{\frac{1}{n}}) = f(q) = \frac{1}{n} f(q^n) = \frac{1}{n} f(p).$$

Thus, if m and n are integers, these two results combine to show that

$$f(p^{\frac{m}{n}}) = \frac{m}{n} f(p).$$

(*) This formalization was pointed out to me by Howard Raiffa.

(†) This term is not traditional in information theory, but it is in the closely related decision theories where it has been widely assumed and debated. There is every evidence from there that this condition, whatever it is called, must be considered much less innocent than it appears to be at first glance.

Let x be any number. We can choose integers m and n such that m/n is arbitrarily close to x, so by the continuity assumption

$$f(p^x) = xf(p).$$

Now, choose $x = -\log_2 p$, so $(\frac{1}{2})^x = p$, then

$$f(p) = f[(\tfrac{1}{2})^x] = xf(\tfrac{1}{2}) = -\log_2 p.$$

Thus, we have the form of the expression for the amount of information transmitted by the selection of any symbol with an *a priori* probability p of being selected.

The expected value of the amount of information transmitted by a source with probability distribution $p(i)$ is therefore

$$-\sum_{i=1}^{n} p(i) \log_2 p(i).$$

A third method to obtain the above expression, which is due to Shannon [1948], is similar to the last one except that it deals with the whole distribution at once. The procedure is to state five *a priori* conditions which many feel must be met by any measure of the average amount of information transmitted per selection from the source.

1. The average amount of information transmitted shall be a real-valued function of the n arguments $p(1), p(2), \ldots, p(n)$; it will be denoted by $H[p(1), p(2), \ldots, p(n)]$.

Next, it seems reasonable, as in the second method, to suppose that if the distribution is changed very slightly, then H should also change only slightly, so we require that

2. H shall be a continuous function in each of its n arguments.

Further, suppose we consider all sources for which the symbols are equally likely, i.e., $p(i) = 1/n$. As n is increased there is more information transmitted by the selection of one symbol since more messages of a given length are possible, so we require

3. When $p(i) = 1/n$ for all i, then H is a monotonically increasing function of n.

Next, we wish to require that if the calculation of the amount of information in a source is divided into a series of subcalculations, then the mode of subdivision shall not alter its value. More exactly, suppose S' is a subset of S (which by relabeling we may always take to be the elements $1,2,\ldots,s$). The set S' can, of course, be treated as a single element s' with probability of occurrence

$$p(s') = p(1) + p(2) + \ldots + p(s).$$

If the form for H is known, we can compute its value for S, for the set with elements $s',s+1,\ldots,n$, and for the set S' alone.(*) Our condition asserts that the first number shall be equal to the weighted sum of the last two, i.e.,

4. $$H[p(1),p(2),\ldots,p(n)] = H[p(s'),p(s+1),\ldots,p(n)] + p(s')\,H\left[\frac{p(1)}{p(s')},\frac{p(2)}{p(s')},\ldots,\frac{p(s)}{p(s')}\right].$$

Finally, we impose the definition of the unit:

5. $H(\frac{1}{2},\frac{1}{2}) = 1$.

From these five conditions, each of which seems to be plausible, Shannon has shown, in a manner similar to that employed in the second method, that H must be of the form

$$-\sum_{i=1}^{n} p(i)\log_2 p(i).$$

Before we discuss any of the properties of H and relate it to the other quantity — channel capacity — which we have defined, let us arrive at the expression for H from a fourth point of view. The following argument is given by Fano [1949], and it is similar to one presented by Shannon [1948]. A plausible way to compare sources is to define a recoding of any source which takes into account the probability distribution of the source and which results in one of a set of standard normal forms of sources. If we can assign a number to each of these normal forms in an intuitively acceptable way, then we have indirectly assigned a number to each source. Of course, the

(*) The analogue of the "independence of irrelevant alternatives" is implicitly assumed at this point when we suppose that choices from S' are governed by the probabilities $p(1)/p(s'),\ldots,p(s)/p(s')$. Actually, this is an extremely powerful, if seemingly plausible, assumption which is the common thread of many theories of choice behavior, as I have shown elsewhere (Luce, [1959]).

only sources to which we have associated any numbers so far are the binary equally likely ones, so it is more than reasonable that we should attempt a recoding into binary equally likely selections.

This may be done in the following manner. Form all possible messages of length r, i.e., messages consisting of r symbols, and call this set R. Since the selections are independent, the probability of each message is simply the product of the probabilities of the individual selections which make it up, hence we know the probability of each message. Thus, we have a probability distribution over R. Divide R into a subset R_1 and its complement $\overline{R}_1$ with respect to R in such a manner that the sum of the probabilities of messages in R_1 is as near 1/2 as possible. To each message in R_1 assign the digit 1 and to each in $\overline{R}_1$ the digit 0. Now, divide R_1 into a subset R_2 and its complement $\overline{R}_2$ with respect to R_1 (not R). Again the choice of R_2 is such that the probability of messages in R_2 is as nearly equal as possible to those in $\overline{R}_2$. To those messages in R_2 assign a second digit 1, so now 11 is assigned to each message in R_2. To those in $\overline{R}_2$ assign as the second digit 0, so 10 is assigned to each message of $\overline{R}_2$. Carry out a similar process in $\overline{R}_1$ leading to the numbers 01 and 00. Continue this "probability halving" until the classes contain single messages. In this manner each message will have assigned to it a sequence of binary digits, the length of the sequence being in large part determined by the probability that the message will occur — the more probable messages having fewer digits than the less probable ones.

An example may make the process clearer:

Message	Probability of occurrence	first digit	second digit	third digit	fourth digit
A	0.50	1	–	–	–
B	0.13	0	1	1	–
C	0.12	0	1	0	–
D	0.12	0	0	1	–
E	0.06	0	0	0	1
F	0.07	0	0	0	0

The first division is between {A} and {B, C, D, E, F,}. No further division of A is possible. The other set is divided into {B, C} and {D, E, F}. These in turn are divided as {B} and {C} and as {D} and {E, F}. The final division is of {E, F} into {E} and {F}.

Such a coding as this is efficient in the sense that the fewest number of binary digits are assigned to the most probable messages and the largest number to the least probable ones. Now, one can ask how many binary digits are required on the average per symbol when messages of length r

are considered. That is, for each message we multiply the number of digits required by the probability that the message occurs, sum these products over all messages, and divide the sum by the total number of symbols r in a message. Call this number H_r. In the above example $H_r = 2.13/r$ bits per symbol. The $\lim_{r\to\infty} H_r$ is a number assigned to each discrete source which both has a plausible meaning and will serve to compare different sources. Fortunately, it can be shown that

$$H = \lim_{r\to\infty} H_r = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$

Thus by four (really only three) routes we have come to the same statistic as the appropriate one to describe the average nature of the source. We can defend it in two further ways; first, by stating some of its properties and showing that they are reasonable for a measure of information transmitted, and second, by using it to make theoretical statements about the transmission of information.

Properties of *H*. A number of theorems about H may be proved (Shannon, [1948]); as we shall need them later, and as they help to give a feel for H, I shall state them.

i. $H \geq 0$, and $H = 0$ if and only if all $p(i)$ except one equal zero. In other words, the entropy of a distribution is always non-negative, and it is zero if and only if the selection of one symbol is certain. Intuitively, no information is conveyed when the selection is certain, and accordingly $H = 0$.

ii. Any averaging of the probabilities in the source increases the value of H. From this, or in other ways, it can be shown that H assumes its maximum value, which is $\log_2 n$, when and only when each of the symbols is equally likely, i.e., when each has probability $p(i) = 1/n$ of being selected.

These two properties of H have led many authors to speak of H as the uncertainty of the source: H assumes its maximum when the selections are maximally "uncertain" and its minimum when absolute certainty obtains. Without disputing the point they have made, it must be mentioned that this use of the word "uncertainty" is at variance with its use in (statistical) decision theory. There, if an *a priori* probability distribution is known, one speaks of decision making under *risk*, and uncertainty is reserved for those cases where the distribution is not completely known. Thus, if the two vocabularies were to be consistent, H should be described as an average measure of risk, not of uncertainty.

iii. Let any long message of N symbols be selected and suppose that it has probability p of occurring, then $-(\log_2 p)/N$ is an estimator of H(*). This last result is, of course, important in estimating H in practical situations, since all that can be observed generally is one message of some long duration. It must be pointed out that when this result is given in precise mathematical language, it asserts that $-(\log_2 p)/N$ almost certainly approaches H as N approaches infinity, i. e., the estimation scheme is consistent.

Non-independent Selections. So far our discussion of the source has been restricted to the independent case, which, as was pointed out, does not include most sources. But our efforts will not be lost, for fortunately we can readily carry over the results for independent sources to the non-independent case.

Consider the selection of one symbol from the set $S = \{1,2,\ldots,n\}$ followed by a second selection from the same set (possibly the next one in forming a message, but we do not need to restrict ourselves to that case). More formally, let x and y be random variables with range S. The joint distribution of x and y is assumed to be known and we shall, for convenience, denote the probability that $x = i$ and $y = j$ by $p(i,j)$. In general, of course, $p(i,j) \neq p(i)p(j)$ since the selections need not be independent. The distribution $p(i,j)$ is now defined over the product space(†) of S with itself, $S \times S$, which is of course a set and so is included among the arbitrary sources we have considered earlier. The definition of entropy can be applied without alteration to the distribution $p(i,j)$, and hence we have as the entropy of the joint distribution of x,y,

$$H(x,y) = -\sum_{i,j} p(i,j) \log_2 p(i,j).$$

Similarly, the definition can be applied to the distribution of the random variable x alone and to that of y alone, and so we have

(*) The plausibility of this can be seen as follows: In a message of length N, the expected number of times that the symbol i will occur is $p(i)N$. Thus, the expectation of the message itself is

$$p' = p(1)^{p(1)N} p(2)^{p(2)N} \ldots . \; p(n)^{p(n)N}.$$

Observe,

$$-\frac{\log_2 p'}{N} = -\sum p(i) \log_2 p(i) = H.$$

(†) The product space of two sets R and S, $R \times S$, is the set of all ordered pairs (r,s), where r is an element from R and s an element from S.

$$H(x) = -\sum_{i,j} p(i,j) \log_2 \sum_{j} p(i,j)$$

$$= -\sum_{i} p(i) \log_2 p(i)$$

and

$$H(y) = -\sum_{i,j} p(i,j) \log_2 \sum_{i} p(i,j)$$

$$= -\sum_{j} p(j) \log_2 p(j),$$

where

$$p(i) = \sum_{j} p(i,j) \text{ and } p(j) = \sum_{i} p(i,j).$$

From these definitions Shannon [1948] noted the following theorem:(*)

$$H(x,y) \leq H(x) + H(y).$$

This result simply states that the entropy (or average uncertainty or amount of information transmitted) of the joint distribution has the intuitively

(*) A simple proof, which was pointed out to me by Lee Abramson, is this: From elementary properties of the logarithm,

$$-[H(x) + H(y) - H(x,y)] = \sum_{i,j} p(i,j) \log_2 \frac{p(i)p(j)}{p(i,j)}$$

$$= \sum_{i,j} p(i,j) \log_2 a_{ij},$$

where $a_{ij} = \dfrac{p(i)p(j)}{p(i,j)}$. Since $\sum_{i,j} p(i)p(j) = \sum_{i} p(i) \sum_{j} p(j) = 1$,

$$0 = \log_2 1 = \log_2 \left\{ \sum_{i,j} p(i,j) \left[\frac{p(i)p(j)}{p(i,j)} \right] \right\}$$

$$= \log_2 \left\{ \sum_{i,j} p(i,j) a_{ij} \right\}$$

Since $\sum_{i,j} p(i,j) = 1$ and the logarithm is convex,

$$\sum_{i,j} p(i,j) \log_2 a_{ij} \leq \log_2 \left\{ \sum_{i,j} p(i,j) a_{ij} \right\},$$

so $H(x) + H(y) - H(x,y) \geq 0$.

necessary property that it is no larger than the sum of the entropies for the two distributions considered separately. In addition it is easily seen that

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

if the events x and y are independent. Thus, whenever there is any inter-symbol influence in the selections, less information is transmitted per symbol than if they had been independent.

If we introduce the conditional probabilities relating the distribution of y to that of x, further relationships of interest can be established. Let $p(j|i)$ denote the conditional probability that $y = j$ given that $x = i$, i. e.,

$$p(j|i) = \frac{p(i,j)}{\sum_j p(i,j)} \quad .$$

The conditional entropy of the random variable y given that $x = i$ is defined to be

$$H(y|x=i) = -\sum_j p(j|i) \log_2 p(j|i) .$$

Hence the expected *conditional entropy* of the random variable y given x is

$$\begin{aligned} H_x(y) &= -\sum_i p(i) \sum_j p(j|i) \log_2 p(j|i) \\ &= -\sum_{i,j} p(i,j) \log_2 p(j|i) . \end{aligned}$$

$H_x(y)$ measures the average uncertainty in the selection represented by y after the selection denoted by x is known.

Shannon has shown that(*)

$$H(x,y) = H(x) + H_x(y) ,$$

which, in words, states that the average uncertainty of the joint distribution is equal to the average uncertainty of the distribution of x added to the

(*) This result is readily proved:

$$\begin{aligned} H(x) + H_x(y) &= -\sum_{i,j} p(i,j) \log_2 \sum_j p(i,j) - \sum_{i,j} p(i,j) \log_2 p(j|i) \\ &= -\sum_{i,j} p(i,j) \log_2 [\sum_j p(i,j)] p(j|i) \\ &= -\sum_{i,j} p(i,j) \log_2 p(i,j) \\ &= H(x,y) . \end{aligned}$$

average uncertainty of the distribution of y when the value of x is known. From this and the preceding result, the following corollary is readily seen to hold:

$$H(y) \geq H_x(y),$$

i. e., the average uncertainty of the distribution of y is never increased by a knowledge of x. The two are equal if and only if the two random variables are independent.

One final concept: the ratio of the entropy of a source to the maximum entropy possible with the same set of symbols is a measure of the information transmitting efficiency of the source — Shannon called it the *relative entropy*. It is generally less than one, either because there is a non-uniform distribution over the symbols or because of the non-independence of symbol selection or, most commonly, because of both. One minus this quantity indicates the percentage of symbols which, though sent, carry no information, i. e., which are redundant. Thus we define the *redundancy* of a source to be

$$1 - \frac{H}{\max H} = 1 - \frac{H}{\log_2 n} \quad .$$

Several estimation procedures indicate that the redundancy of written English is at least 50 per cent and very likely nearer 75 per cent (see Chapter 7). The reason for such high redundancy will become apparent later.

In discussing the applicability of information theory to certain problems in psychology, Miller and Frick [1949] suggested that redundancy be called the *index of behavioral stereotypy*. The motive for this term of course is that redundancy is a quantity which is 1 when the behavior is completely stereotypic and 0 when each of the several alternatives arises with equal probability. For the most part, however, the shorter term is used.

The Fundamental Theorem of a Noiseless System. The following fundamental result, due to Shannon [1948], shows in effect that the above definitions of channel capacity and of source entropy or average uncertainty are suitable formalizations of our intuitions about the limitations on information transmission.

Theorem: *Let the entropy of a source be* H *bits per symbol and the capacity of a noiseless channel be* C *bits per second. For any positive number* ϵ *no matter how small, there exists a coding of the source, i. e., there exists a transmitter, such that it is*

possible to transmit at an average rate of (C/H) — ϵ *symbols per second. It is not possible to devise a code so as to transmit at an average rate of more than* C/H *symbols per second.*

Three points should be made about this theorem. First, it must be kept in mind that the definition of the entropy of a source rests only upon the statistical structure of the source, and it does not in any way depend upon the properties of the channel. The capacity of the channel depends only upon channel properties and not at all upon the statistical structure of the source. The theorem asserts that these definitions have, however, been so chosen that the ratio C/H is the least upper bound of the transmission rate.

Second, the code which the theorem asserts to exist is, of course, influenced by how small we take ϵ. If ϵ is near C/H then nearly any code will do, but as ϵ approaches 0 fewer and fewer codes will produce a rate of $(C/H) - \epsilon$. But the theorem asserts that there will always be at least one. A major unsolved problem of information theory is to devise a theorem which describes such a code in detail for given values of C, H, and ϵ; the above theorem only asserts that such a code exists.

Third, such optimal use of the channel as described in the theorem is not effected without paying some price. The price is delay. If one is to code a message optimally when there are intersymbol influences, then it is necessary to wait before transmission to see how that influence can be utilized in the coding, thus effecting a delay in the transmission. Similarly, at the receiver, the translation into the language of the destination must be delayed in exactly the same way, for a single received symbol will have meaning only by its relation to a number of others. In practical engineering work a compromise is reached between long delays (and hence expensive storage equipment) and nearly optimal use of the channel.

The theorem may be recast in a slightly different form, which may help clarify it and which will be useful when we study the noisy system. Let R denote the average rate at which symbols are transmitted over the channel when a given code is used. The theorem then asserts that $C/H \geq R$ and that there exist codes such that the corresponding R is arbitrarily close to C/H. If we rewrite this as $C \geq HR$ and then maximize both sides with respect to all possible codes we have

$$C = \max_{\text{codes}} C = \max_{\text{codes}} (HR).$$

It is conventional, though misleading, simply to replace HR in the above expression by H. Previously, the entropy of a source was measured in "bits

per symbol," but in this reformulation we measure the entropy of the source (and transmitter combination) in "bits per symbol" times "symbols per second," i. e., in "bits per second" transmitted. The theorem then asserts that the channel capacity is equal to the maximum number of bits per second which can be transmitted by the source-transmitter combination over the channel. In this form, and in a corresponding form for noisy systems, the fundamental theorem has been used in behavioral applications.

4. THE DISCRETE NOISY SYSTEM

AS IN THE PRECEDING chapter, I shall suppose that the source is discrete, but I shall now drop the condition of a noiseless system.

Equivocation and Channel Capacity. The significant effect of noise in a system, as was pointed out in Chapter 2, is to cause the destination sometimes to be mistaken as to which symbol was transmitted. Any other properties the noise may have are irrelevant in this theory of information transmission. Thus, if we assume that both the signal and the noise time series are stationary, and that the noise affects successive selections independently, then the noise is completely characterized by the matrix of conditional probabilities $p(j|i)$ which state the probability that symbol j is received when i was sent. Formally, this situation is identical to the case of non-independent selections: in that case we interpreted j as a selection following i; here we shall interpret j as the selection received at the destination when i was actually selected at the source.

The quantities $H(x)$, $H(y)$, $H(x,y)$ and $H_x(y)$ are defined as before. $H(x)$ is the entropy of the source distribution, $H(y)$ the entropy of the destination distribution, $H(x,y)$ the entropy of the joint distribution of x and y, $H_y(x)$ measures the average ambiguity in the signal sent given the received signal, while $H_x(y)$ measures the average ambiguity of the received signal given the signal which was sent. When we are considering noise, $H_y(x)$ is called *equivocation.*

If a system is noiseless, then $H_x(y)=0=H_y(x)$ and so $H(x)=H(y)$.

Let us suppose that all the entropies are calculated in bits/sec, rather than bits/symbol, then the effective *average rate of transmission*, R, (in bits/sec) is the average rate of information sent, $H(x)$, minus that which was lost as a result of the noise, $H_y(x)$:

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

This can easily be shown to be equal to two other expressions, the first of which states that the rate of transmission is the difference between what was received and what was received incorrectly. In symbols,

$$\begin{aligned} R &= H(y) - H_x(y) \\ &= H(x) + H(y) - H(x,y). \end{aligned}$$

The notion of rate of transmission for the noisy case is analogous to that introduced for the noiseless case in the last statement of the fundamental theorem of the noiseless case. It suggests that one way to define channel capacity in the noisy case is as follows:

$$C = \max_{\text{codes}} [H(x) - H_y(x)].$$

By the theorem of the last section, this definition reduces to that of channel capacity of a noiseless system since in that case $H_y(x) = 0$. Note, it does not reduce directly to the *definition* of channel capacity as given early in Chapter 3; however, later a theorem will be presented which shows that there is an analogous, though more complicated, definition for the noisy case.

Theorems. Consider the communication system diagrammed in Fig. 4. We assume that there is an observer who is able to perceive without error both the selections made by the source and the corresponding signals received at the destination. Let us suppose that the equivocation due to noise is $H_y(x)$. If there is a noiseless correction channel from the observer

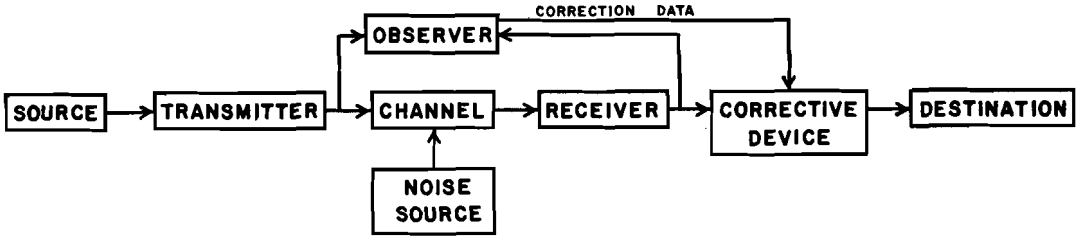

FIG. 4.

to the destination with capacity $H_y(x)$ bits/sec, it can be shown (Shannon, [1948]) that it is possible to encode correction data in such a manner as to correct all but an arbitrarily small fraction of the errors due to the noise. This is impossible if the channel capacity of the correction channel is less than $H_y(x)$. This theorem is of theoretical interest — it shows that $H_y(x)$

does in fact summarize the average effect of the noise. But it is certainly not a practical scheme to combat noise.

The following result, which is due to Feinstein [1954] and which is somewhat sharper than the original result of Shannon [1948], is a fundamental theorem for the noisy case.

Theorem. *Let the entropy of a source be* H *bits per second and the capacity of the channel* C *bits per second. Let* ϵ *be any number larger than* 0. *If* $H < C$, *then there is a number* $N(\epsilon, H)$ *such that among all messages of length* $N \geq N(\epsilon, H)$ *we can find a subset* $\{u_i\}$ *having at least* 2^{NH} *members with the properties that*

1. *we may associate a set* B_i *of messages of length* N *to each* u_i *in such a way that if* u_i *is sent the probability that a member of* B_i *is received is greater than* $1 - \epsilon$, *and*

2. *the sets* B_i *are non-overlapping.*

If $H > C$, *this cannot be done.*

Let us examine the various components of this result. As in all such theorems, ϵ is to be thought of as a very small number which represents the permitted error tolerance. We are then required to consider long messages, the length depending upon both the value of H and how small we take ϵ. Of these n^N possible messages we consider a subset $\{u_i\}$. This subset includes most of the possible messages, e. g., if the selections are equiprobable, then it includes all of them since $n^N = 2^{N \log_2 n} = 2^{NH}$. The theorem asserts that to each of these messages we can associate a subset B_i of messages such that if, whenever a member of B_i is received, we infer that u_i was sent, then we know that the probability of being wrong is less than ϵ. This last statement is not justified by 1 alone, for if the sets B_i were to overlap there would not always be a unique inference. So part of the assertion is that they do not overlap. In this way, the effect of the noise can be combatted as effectively as we choose whenever $H < C$. The more effective we require the coding process to be, the larger we are forced to take N. This means that the price of combatting noise is delay, which in practice means extensive storage equipment.

Note that the theorem also asserts that it is never possible to do this if $H > C$.

Shannon's original theorem, which is weaker than Feinstein's result, can be stated in the following way:

Theorem: *Let the entropy of a source be* H *bits per second and the capacity of the channel* C *bits per second. If* $H \leq C$, *then there exists a coding scheme such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors. If* $H > C$, *it is possible to reduce the equivocation to as near* $H - C$ *as one chooses, but it is not possible to reduce it below* $H - C$.

McMillan's comments on this result seem to be worth repeating:

"Engineering experience has been that the presence in the channel of pertubation, noise, in the engineer's language, always degrades the exactitude of transmission. [The theorem] above leads us to expect that this need not always be the case, that perfect transmission can sometimes be achieved in spite of noise. This practical conclusion runs so counter to naive experience that it has been publicly challenged on occasion. What is overlooked by the challengers is, of course, that 'perfect transmission' is here defined quantitatively in terms of the capabilities of the channel or medium, perfection can be possible only when transmission proceeds at a slow enough rate. When it is pointed out that merely by repeating each message sufficiently often one can achieve virtually perfect transmission at a very slow rate, the challenger usually withdraws. In doing so, however, he is again misled, for in most cases the device of repeating messages for accuracy does not by any means exploit the actual capacity of the channel.

"Historically, engineers have always faced the problem of *bulk* in their messages, that is, the problem of transmitting rapidly or efficiently in order to make a given facility as useful as possible. The problem of noise has also plagued them, and in many contexts it was realized that some kind of exchange was possible, for example, noise could be eliminated by slower or less 'efficient' transmission. Shannon's theorem has given a general and precise statement of the asymptotic manner in which this exchange takes place." (1953, p. 207).

He goes on to point out the similarity in the exchange between bulk and noise and the rather general exchange between sample size and power in statistical tests.

Although the simple repetition of a message is not usually an efficient way to employ the channel capacity to eliminate errors, some form of redundant transmission is required. In general it will be far more complicated than repetition, but, as with repetition, a delay in the reception of a message must result. The essential point of the theorem is that the delay need not be such as to reduce the rate of transmission to zero. The proof of the theorem is not constructive and so there is no indication what code

to use to utilize fully the channel capacity. Shannon and Weaver write, "Probably this is no accident but is related to the difficulty of giving an explicit construction for a good approximation to a random sequence." ([1948], p. 43) Much recent (engineering) work in information theory has been devoted to finding near optimal codes for certain important special cases.

Shannon's fundamental theorem of the noisy case may be recast in a form that shows the relation of the present definition of channel capacity to that given for the noiseless case. Let q be a number such that $0<q<1$. Consider all possible signals of duration T time units which might be transmitted over the channel and let R denote a typical subset of these signals. Under the assumption that each signal of R is equally probable and taking into account the statistics of the noise, let a receiver be designed to choose as the cause of the signal it receives the one in R which most likely is distorted into the one received. It is clear that in general errors will be made; let $p(R)$ denote the probability that an incorrect interpretation will be made when the subset is R. Consider now all those subsets R such that $p(R)\leq q$. Among these sets there is one which contains the most signals, let that number be denoted by $N(T,q)$. Shannon [1948] then showed that

$$C=\lim_{T\to\infty}\frac{\log_2 N(T,q)}{T},$$

which is clearly analogous to the original definition of channel capacity for the noiseless case. It is remarkable that this result is independent of the value of q. Presumably, however, the rate of convergence of the limit is not independent of q, and so in any application of the theorem one should attempt to exploit the freedom in choosing q.

Channel Capacity of a Noisy System: Independent Selections. Shannon [1948] and Fano [1950] have shown that if one assumes that the selections at the source are independent, then the capacity of the channel is given by the transcendental equation

$$\sum_j 2^{-\sum_i h(j|i)\left[C-\sum_j p(j|i)\log_2 p(j|i)\right]}=1,$$

where $h(j|i)$ is a typical element of the inverse of the noise matrix, i. e.,

$$\sum_j h(j|i)p(j|k)=\begin{matrix}1 \text{ if } i=k\\ 0 \text{ if } i\neq k\end{matrix}.$$

It is difficult, if not impossible, to see the dependence of channel capacity on the noise matrix from this equation, but, of course, in any given case one can solve for C numerically. However, if we can assume that the noise has the same disturbing effect on each symbol of the source, i. e.,

$$H_x(y) = -\sum_i p(i) \sum_j p(j|i) \log_2 p(j|i)$$

is independent of $p(i)$, then it can be shown (Fano [1950]) that

$$C = \log_2 n - H_x(y).$$

In the special case of a binary source (two elements) and noise such that the probability of an erroneous transmission is a, then the capacity is given by

$$C = 1 + a \log_2 a + (1-a) \log_2 (1-a).$$

It is easy to make interesting calculations using this last expression. For example, if the chance of an error is 1 per cent, then the channel capacity is reduced to approximately 90 per cent of its value in the absence of noise. This marked non-linearity must be kept in mind whenever thinking about the effects of noise.

5. SOME ASPECTS OF DISCRETE THEORY RELATED TO APPLICATIONS

As WE SHALL SEE in some detail in Section II, many of the applications of information theory in psychology are to problems not classically described as communication. Indeed, they are communication problems only in the sense that any experiment, or any decision, can be treated as a transmission of information. Put another way, in the attempt to analyze communication systems, a mathematical formalism has been produced to deal with the average character of certain inference problems, and this mathematics can be completely divorced from its realization as a communication system. At the same time, there are other realizations of the same mathematical system in psychology. Because of its origins, however, the information terminology is associated with the mathematics and so with its applications. Some of this vocabulary may seem peculiar in some applications, but it is probably not as misleading as it may seem initially. In this section, I propose to discuss (but divorced from the communication

model) a part of the formalism that has been particularly important in psychological applications. The topics to be considered are: a relation between the rate of information transmission and statistical inference, a generalization of the notion of transmission rate, and the statistical sampling and significance problems.

Inverse Probabilities, Bayes Theorem, Contingency Tables. The structure of very many problems in psychology and the other behavioral sciences can be reduced to the existence of two classes of possible occurrences, usually called stimuli and responses, such that an occurrence in the response class is in some degree dependent upon what stimulus occurred. It is not easy to characterize in a useful and simple way the relation between these two classes of occurrences. It is, of course, possible to present the whole matrix of joint probabilities $p(i,j)$, i. e., to give the entire contingency table, but this hardly can be called simple. Various measures of contingency have been proposed and used, but objections have been raised to each of these. Still another possibility — one that has found favor among some psychologists — is the entropy measure. The expression most often used is

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

which, when the entropies are measured in bits/sec, is called the rate of information transmission (Chapter 4). As often as not, time does not enter into psychological applications in a natural manner, and it is more appropriate to treat the stimuli and the responses as static and to measure entropies in bits. In that case the following notation is employed:

$$\begin{aligned} T(x;y) &= H(x) + H(y) - H(x,y) \\ &= H(x) - H_y(x) \\ &= H(y) - H_x(y), \end{aligned}$$

and the quantity $T(x;y)$ is simply called the *information transmitted* from the stimulus to the response. It is a quantity which is 0 when the random variables x and y are statistically independent and it is a maximum when they are in one-to-one correspondence, i. e., when a knowledge of the value of x uniquely determines the value of y and conversely. In other words, T is a measure of the contingency between x and y.

Note that in this interpretation of the formalism the role of the human being has changed: Previously, we had thought of the source and the

destination as people and the channel as a physical entity. In most psychological applications, the stimuli correspond to the source and the responses to the destination; the subject is treated as a noisy channel causing less than perfect correspondence between the stimuli and the responses.

One can also think of the relation between the two random variables x and y as a problem of inferring as well as possible the value of x from a knowledge of the value of y. This is, of course, the problem of inverse probabilities which has had a long history in statistical theory, and Bayes theorem is one of the most famous results. We may think of it in the following form: There are n possible underlying states of nature, $i = 1,2,\ldots,n$, which are known *a priori* to have probabilities $p(i)$ of occurring. We suppose an experiment is performed with possible outcomes $j = 1,2,\ldots,m$, the actual outcome depending somewhat upon which state obtains. Let x be a random variable with range the states of nature and distributed according to $p(i)$ and y a random variable with range the experimental outcomes. Further, let us assume as known the conditional probabilities, $p(j|i)$, that $y = j$ when $x = i$. The problem then is to estimate the probability $x = i$ when the outcome of the experiment is known, i. e., when $y = j$ is given.

Cherry [1953] describes the analogy to the noisy communication system as "... an observer receives the distorted output signals (the posterior data...) from which he attempts to reconstruct the input signals (the hypotheses), knowing only the language statistics (the prior data)." (p. 39).

It is well known that Bayes theorem reads,

$$p(i|j) = \frac{p(j|i)p(i)}{\sum_i p(j|i)p(i)} .$$

If one takes logarithms on both sides of this equation, multiplies the result by $p(i,j)$, and then sums on both i and j, the result is simply

$$H(x) - H_y(x) = H(y) - H_x(y),$$

i. e., the information transmitted from x to y.

Deeper connections between conventional statistics and the information statistic have been explored by Kullback and Leibler [1951] and Kullback [1952].

Modern statistical inference, stemming largely from Wald [1947] (also see Blackwell and Girshick [1954], and Savage [1954]), takes a somewhat different tack. One of the central notions is that there must be given an

evaluation of incorrect decisions, i.e., a loss function must be selected, and the problem is then to reach inferences which are optimal relative to that function. Our point of view has been tacitly to minimize information loss. However, other possible loss functions could be examined. Van Meter and Middleton [1954] have presented a theory along these lines which rests upon statistical decision theory and is therefore beyond the scope of this survey.

Multivariate Theory. An alternative way of viewing information theory — one which seems especially useful in many psychological applications — draws a close parallel between an information analysis of stimulus-response patterns and analysis of variance. It is a more general, and so a weaker, analysis than analysis of variance since it does not presuppose any metric information about the stimulus or the response sets.

Suppose we are analyzing a stimulus-response situation by information theoretical techniques, then the basic equation we have developed,

$$H(y) = H_x(y) + T(x;y),$$

decomposes an average measure of the response pattern into two parts: $T(x;y)$, which is determined by the stimulus, and $H_x(y)$, which is unexplained "random" variation — random in the sense that it is uncorrelated with the stimulus x. It may very well happen that a considerable portion of the residue $H_x(y)$ can be explained in a systematic manner, though not by the experimental stimuli that have so far been considered. For example, consider an experiment in which subjects are required to classify liminal tones into one of n categories. It may very well happen that the subject's response is determined only in small part by the tone presented, but that in large part it is predictable from a knowledge of his previous response, even if we do not know the stimulus. In such a case, it may be not only appropriate but essential to consider as the stimulus the pair of random variables (u,v), where u has the possible tones as its range and v the possible previous responses of the subject. In other words, in some cases we may be able to understand the phenomenon adequately only if we treat as the stimulus a random variable with a range which is the product space of two, or more, simpler sets. McGill [1953, 1954, 1955 a, 1955 b] has examined this problem in some detail and he has appropriately generalized the transmission concepts so as to produce a multivariate theory where, of course, Shannon's theory is the bivariate case. I shall recount this development briefly.

First of all, we may replace x by the symbol (u,v), which is equivalent to x when the range of x is the product space of the ranges of the random variables u and v, in the equation for information transmission. This yields

$$T(u,v;y) = H(u,v) + H(y) - H(u,v,y).$$

(I have systematically omitted the extra parentheses about u,v for greater clarity). It is clear that in our discussion there has not been any formal notion of direction of transmission between source and receiver, and so they may be interchanged. Formally,

$$T(u,v;y) = T(y;u,v).$$

Next, we want to introduce a measure which gives the separate dependence of y upon u and upon v. To do this, it seems appropriate to define a measure of the conditional information transmitted from the stimulus u to the response y when the stimulus v is held constant, say at the value j. With v fixed, this is simply the transmission expression we have obtained previously, namely,

$$T_j(u;y) = H_j(u) + H_j(y) - H_j(u;y).$$

Now, since we deal only with averages, we shall need

$$\begin{aligned} T_v(u;y) &= \sum_j p(j)\, T_j(u;y) \\ &= \sum_j p(j)\, H_j(u) + \sum_j p(j)\, H_j(y) - \sum_j p(j)\, H_j(u,y). \end{aligned}$$

Expand each of the three terms on the right, e. g., the first gives:

$$\begin{aligned} \sum_j p(j)\, H_j(u) &= \sum_j p(j) \sum_i p(i|j) \log_2 p(i|j) \\ &= \sum_j \sum_i p(j)\, p(i|j) \log_2 p(i|j) \\ &= \sum_j \sum_i p(i,j) \log_2 \frac{p(i,j)}{p(j)} \\ &= H(u,v) - H(v). \end{aligned}$$

The other terms are similar, and combining them we obtain

$$T_v(u;y) = H(u,v) + H(v,y) - H(v) - H(u,v,y).$$

In like manner,

$$T_u(v;y) = H(u,v) + H(u,y) - H(u) - H(u,v,y),$$
$$T_y(u;v) = H(u,y) + H(v,y) - H(y) - H(u,v,y).$$

Clearly, v will have an effect on the average transmission from u to y if and only if $T_v(u;y) \neq T(u;y)$, and the magnitude of this effect is measured by

$$A(uvy) = T_v(u;y) - T(u;y).$$

Similar quantities can be defined to measure the effect of u on the transmission from v to y and of y on the transmission from u to v. There is not, however, any need to introduce a new symbol for each of these since they can all easily be shown to be equal, i. e.,

$$\begin{aligned} A(uvy) &= T_u(v;y) - T(v;y) \\ &= T_y(u;v) - T(u;v). \end{aligned}$$

"In view of this symmetry, we may call $A(uvy)$ the $u.v.y$ interaction information. We see that $A(uvy)$ is the gain (or loss) in sample information transmitted between any two of the variables, due to additional knowledge of the third variables." [McGill, 1954, p. 101]. We shall return to the exact meaning of this term below.

With these concepts, it is now possible to express the three-dimensional average information transmitted in terms of the two-dimensional ones and the interaction information. We show that

$$\begin{aligned} T(u,v;y) &= T(u;y) + T(v;y) + A(uvy) \\ &= T_v(u;y) + T_u(v;y) - A(uvy). \end{aligned}$$

Substituting one of the expressions for A,

$$T(u;y) + T(v;y) + A(uvy) = T_v(u;y) + T(v;y).$$

Now substitute the H expressions for the two right hand terms,

$$\begin{aligned} T_v(u;y) + T(v;y) &= H(u,v) + H(v,y) - H(v) - H(u,v,y) \\ &\quad + H(v) + H(y) - H(v,y) \\ &= H(u,v) + H(y) - H(u,v,y), \end{aligned}$$

but this was previously shown to be equal to $T(u,v;y)$. The second expression follows immediately from the definition of A.

We may write this three-dimensional information transmission in another way which parallels the familiar equation $H(y) = H_x(y) + T(x;y)$, namely,

$$\begin{aligned} H(y) &= H_{uv}(y) + T(u,v;y) \\ &= H_{uv}(y) + T(u;y) + T(v;y) + A(uvy). \end{aligned}$$

The term $H_{uv}(y)$ is the residual or unexplained variability in the response y after the information about y given by u and by v and the interaction information of the three variables has been removed.

An initially unexpected feature of McGill's analysis was the possibility that the interaction term may be negative. As Miller ([1954 a], p. 411) put it, "In other words, a knowledge of the input [v] may decrease the amount of information that [y] has about [u] — communication from [u] to [y] would actually be better if no data about [v] were collected at all!" Are we then forced to think of the transmission of negative information? No, for as McGill [1955 a] has pointed out the interaction term A is composed of two effects: the interaction of the three variables plus the correlation of u and v. If the correlation is high, then there is a good chance that A will be negative. Thus, he argues, if sense is to be made of the interaction term, we must choose u and v to be independent in the experimental design. It is not obvious, however, that the organism being studied will necessarily elect to respond only to statistically independent variables. We can easily confine our analysis to such cases, but we may at the same time limit the possibility of describing the behavior simply.

One of the most important and desirable properties of the information statistic — entropy — is its additive character. This was apparent in the two-dimensional case and is even more forcibly illustrated in the three-dimensional theory. Each of the contributions — that from u, from v, from the interaction, and from the unexplained variability — is simply added to obtain the information in the response pattern. Thus, information analysis of a stimulus-response situation seems to parallel analysis of variance. McGill [1953, 1955 a] and Garner and McGill [1956] have shown that there is in fact a striking formal parallel between information analysis, analysis of variance, and correlational analysis. To be sure, there are differences: "... information transmission is made to order for contingency tables. Measures of transmitted information are zero when variables are independent in the contingency-sense (as opposed to the restriction to linear independ-

ence in analysis of variance). In addition, the analysis is designed for frequency data in discrete categories, while methods based on analysis of variance are not." (McGill [1954], p. 107). Nevertheless, "It would seem that information theory effectively corresponds to a nonparametric analysis of variance." (Miller [1954a], p. 411).

There is no reason why the above analysis cannot be extended to more dimensions than three, and McGill [1954] has carried this out in some detail. There is little reason to reproduce it here. It should be mentioned, however, that as with sequential dependencies in the source, the amount of data needed and the number of calculations required mount sharply as the number of dimensions is increased.

Statistical Tests and Estimations of Entropy. In addition to constructing models, the behavioral scientist, unlike many physical scientists, must confront the difficult statistical problem of testing and using his model when the only data available are from comparatively small samples. His use of information theory is no exception to this rule, so we turn now to that incompletely resolved problem.

Let us suppose that a distribution $p(i)$ governs the selections of the n alternatives $1,2,\ldots,n$, and let us suppose that a sample of N independent observations of selections yields $N(i)$ cases of alternative i. The true entropy is, of course,

$$H = -\sum_{i=1}^{n} p(i) \log_2 p(i),$$

while

$$\hat{H} = -\sum_{i=1}^{n} \frac{N(i)}{N} \log_2 \frac{N(i)}{N}$$

is the estimator of the entropy obtained by replacing each $p(i)$ by its maximum likelihood estimator $N(i)/N$. Miller and Madow [1954] have shown that if the $p(i)$ are not all equal, $\sqrt{N}(H-\hat{H})$ has a normal limiting distribution with mean 0 and variance

$$\sigma^2 = \sum_{i=1}^{n} p(i) [\log_2 p(i) + H]^2.$$

If, however, $p(i) = 1/n$ for every i, then $(2N/\log_2 e)(H-\hat{H})$ has a chi-square limiting distribution with $n-1$ degrees of freedom.

They point out (also see Miller [1955]) that if small samples are used to estimate the entropy there is a bias which can be corrected for by the following theorem:

$$H = E(\hat{H}) + (\log_2 e)\left[\frac{n-1}{2\mathcal{N}} - \frac{1}{12\mathcal{N}^2} + \frac{1}{12\mathcal{N}^2}\sum_{i=1}^{n}\frac{1}{p(i)}\right] + 0\left(\frac{1}{\mathcal{N}^3}\right),$$

where $E(\hat{H})$ is the expected value of H and $0(1/\mathcal{N}^3)$ denotes terms of the order of $1/\mathcal{N}^3$ or smaller. They also establish a similar expression for the variance of $\hat{H}$, but as it is fairly complex I shall not reproduce it here.

For the case of equally likely alternatives, Rogers and Green [1955] have developed an exact expression for the expected value of $\hat{H}$, namely,

$$E(\hat{H}) = \log_2 n - \sum_{i=2}^{\mathcal{N}}\binom{\mathcal{N}-1}{i-1}\frac{\sum_{j=0}^{i-2}(-1)^j\binom{i-1}{i-j-1}\log_2(i-j)}{n^{i-1}}.$$

The Miller and Madow approximation in the same case reduces to

$$E(\hat{H}) = \log_2 n - \frac{(\log_2 e)\,[n^2 + 6\mathcal{N}(n-i) - 1]}{12\mathcal{N}^2},$$

which, of course, is much simpler. Rogers and Green point out that for $\mathcal{N} \geq n$ the two give nearly the same results, but that for $\mathcal{N} < n$, "... the Miller-Madow formula ... becomes increasingly less accurate and (their formula) becomes more easily computable." (Rogers and Green [1955], p. 103). They also present a similar expression for the variance which I shall not reproduce here. Tables are given of the mean and variance in the equally likely case for small values of $\mathcal{N}$ and n (they use the symbol K for what I have called n).

Miller [1955] has also treated the problem of transmitted information, i. e., of contingency tables having r stimulus alternatives and s response alternatives. Let the three probability distributions be denoted by $p(i)$, $p(j)$, and $p(i,j)$, and let the observed sample frequencies from a sample of size $\mathcal{N}$ be denoted by $\mathcal{N}(i)$, $\mathcal{N}(j)$, and $\mathcal{N}(i,j)$. The transmitted information, T, is of course given by

$$T = -\sum_{i=1}^{r} p(i)\log_2 p(i) - \sum_{j=1}^{s} p(j)\log_2 p(j) + \sum_{i=1}^{r}\sum_{j=1}^{s} p(i,j)\log_2 p(i,j).$$

Let $\hat{T}$ be the estimator which is obtained by replacing each $p(i)$ by its maximum likelihood estimator $\mathcal{N}(i)/\mathcal{N}$. Define:

$$\lambda = \frac{\prod_{i=1}^{r}\left(\frac{\mathcal{N}(i)}{\mathcal{N}}\right)^{\mathcal{N}(i)} \prod_{j=1}^{s}\left(\frac{\mathcal{N}(j)}{\mathcal{N}}\right)^{\mathcal{N}(j)}}{\prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{\mathcal{N}(i,j)}{\mathcal{N}}\right)^{\mathcal{N}(i,j)}}.$$

It is known from Wilks' [1935] likelihood-ratio test of independence that $-2\log_e \lambda$ has the chi-square distribution with $(r-1)(s-1)$ degrees of freedom. It is not difficult to show that

$$1.3863\mathcal{N}\hat{T} = -2\log_e \lambda,$$

hence $1.3863\mathcal{N}\hat{T}$ has a chi-square distribution with $(r-1)(s-1)$ degrees of freedom under the null hypothesis $T = 0$, i. e., when the stimuli and the responses are independent.

In the same paper, Miller showed that

$$T = E(\hat{T}) - \frac{\log_2 e}{2\mathcal{N}}(r-1)(s-1) + 0\left(\frac{1}{\mathcal{N}^2}\right),$$

and so it is possible to correct for small sample bias. He suggests that $\mathcal{N}$ should be at least $5rs$ in order to make estimates of the information transmitted.

McGill [1954] has extended some of the above results to the multivariate case. First, he observes that:

$$\text{if} \begin{cases} y \text{ is independent of } (u,v) \\ y \text{ is independent of } v \\ y \text{ is independent of } v \text{ when } u \text{ is held constant} \end{cases} \text{ then } \begin{cases} T(u,v;y) = 0 \\ T(v;y) = 0 \\ T_u(v;y) = 0 \end{cases}$$

The last two conditions each imply

$$T_v(u;y) = T(u;y),$$

or, in words, v is not involved in the transmission between u and y when either of the two conditions holds.

There are, of course, analogous statements for the symbols u, v, and y.

To test the hypothesis that any of the T's are zero, McGill uses Miller's result relating independence with the likelihood-ratio test. One obtains

$$\text{if} \begin{cases} T(u,v;y) = 0 \\ T(u;y) = 0 \\ T(v;y) = 0 \\ T_y(u;v) = 0 \end{cases} \text{ then } 1.3863N \begin{cases} \hat{T}(u,v;y) \\ \hat{T}(u;y) \\ \hat{T}(v;y) \\ \hat{T}(u;v) \end{cases} \begin{matrix} \text{has approximately} \\ \text{a chi-square} \\ \text{distribution} \\ \text{with} \end{matrix}$$

$$\begin{cases} (UVY-1)-(U-1)-(V-1)-(Y-1) \\ (U-1)(Y-1) \\ (V-1)(Y-1) \\ Y(U-1)(V-1) \end{cases} \text{ degrees of freedom}$$

where U, V, and Y are the number of elements in the ranges of u, v, and y respectively, and N is the size of the sample.

He shows that if the null hypothesis

$$p(i,j,m) = p(i)p(j)p(m)$$

is true, then $T(u;y)$, $T(v;y)$ and $T_y(u;v)$ are asymptotically independent; thus, as an approximation, the corresponding $\hat{T}$'s can be tested simultaneously for significance under the null hypothesis.

McGill presents an interesting example which shows very graphically that "... we cannot decide whether an amount of transmitted information is big or small without knowing its degrees of freedom." ([1954], p. 114).

SECTION TWO

APPLICATIONS TO BEHAVIORAL PROBLEMS

6. INTRODUCTION

THE APPLICATIONS of information theory, and its indirect influences in substantive areas, are not easily summarized and evaluated. Besides the unambiguous applications which can be cited, information theory has subtly influenced the thinking of many behavioral scientists. Not only has it affected their analysis of certain kinds of data, but also their choice of experimental problems. Such influences cannot be succinctly described or tabulated.

One is, therefore, practically forced to confine his attention to the published papers where information theory has been explicitly employed. But in most of the behavioral areas these articles have been sporadic, and they hardly present a clear pattern.(*) Thus, I am more or less forced to confine my attention to the two behavioral sciences in which these publications have been especially numerous and the patterns are fairly clear: psychophysics and psychology.(†)

The importance of information theory in psychology was realized in the late forties, only a year or two after Shannon's now classic paper was published. This recognition was both symbolized and accelerated by a paper published in 1949 by Miller and Frick. They observed that "...[a] psychologist's experiments usually generate a sequence of symbols: right and wrong, conditioned and unconditioned, left and right, slow and fast, adient and abient, etc." (p. 314). Moreover, very many experiments are of the stimulus-response type where the stimuli form one sequence and the responses another. Generally, the procedures used to analyze such data ignore the sequential relations among the responses (usually, though not

(*) Biology is to some degree an exception. Much of the application to biological problems has stemmed from the interest of Quastler, who has gathered together a good deal of the work in one volume (1953).

(†) Much of the material we shall discuss here has been summarized by Miller [1954 a, 1956] in somewhat less detail. Hick [1954] has also discussed the impact of information theory in psychology, and other surveys of applications can be found in Patton [1954] and Bricker [1955].

always, sequential effects in the stimuli are experimentally eliminated by randomizing procedures). Ignoring the sequential information, they pointed out, is tantamount to assuming the independence of successive responses. They did not imply that psychologists felt that this was a reasonable assumption, but only that many standard statistical techniques are not really suited to analyze such data. An exception, of course, is the use of contingency tables to study temporally ordered pairs of responses (digrams) and the use of contingency measures to characterize the degree of association between the arguments of the table. Miller and Frick then outlined certain aspects of information theory and proposed that the information measure be employed in such situations. As Klemmer and Frick pointed out in a later paper, "The [information] measure may be applied without logical difficulty to any situation in which one is willing to identify the members of the stimulus and response classes and make some statements about their probability distributions. Whether or not the measure is useful in the analysis of human behavior remains to be proven. Early results from its application are, however, encouraging..." ([1953], p. 15).

There are difficulties, for as Miller and Frick pointed out, at least these two serious *a priori* limitations exist on the applicability of information theory:

1. Sequential responses which are generated while learning is occurring do not form a suitable sample from which to estimate the probabilities that are needed: the assumptions of learning and of a stationary response time series are incompatible.

2. The difficulty of obtaining adequate samples to estimate probabilities increases sharply with an increase in the length of dependencies in the response sequence. In fact, it is completely out of hand beyond three step dependencies.

Related computational difficulties also arise with large amounts of sequential data. Basically, however, this problem is less serious than the sampling one, since computation machines are available that are ideally suited to repetitious calculations. In addition, special equipment, such as that described by Newman [1951a], can be constructed to carry out information-type analysis. In practice, however, most computations will be done by hand, so tables of $\log_2 p$ and $p \log_2 p$ are useful to have. Several have been published: Newman [1951a], Dolanský and Dolanský [1952], and Klemmer [1955].

For the five years immediately after the publication of the Miller-Frick paper there was a steady increase in the number of psychological papers

employing information theory. In 1955 it seemed to reach a plateau provided we count the separate contributions in Quastler's *Information Theory in Psychology*. (Actually, not all should be counted, since some summarize already published studies, so the trend may have turned down somewhere in 1954 or 1955.) But even if it reached its peak in 1954, it will certainly not vanish completely in the near future; hence, any summary I attempt here is bound to be out-of-date before it can be very widely read. However, there is some pattern to the publications and a summary may serve a function, so long as it is remembered that it only covers a cross-section of an incomplete trend. Excepting the applications of information theory to psychological testing,(*) I believe this summary is fairly complete through 1954. I have not tried exhaustively to survey the contributions in 1955 and the first half of 1956 (when the final revisions were completed); however, since much of the 1955 material appears in Quastler's book, a quick and adequate view of the most recent work is readily available.

Four features of the applications seem worth noting here:

1. Few of the applications are to problems traditionally classed as communication; this was predicted by Miller and Frick.

2. The applications do not generally use the fundamental theorem relating channel capacity, the statistical structure of the source, and the transmission rate. I know of only two limited attempts to characterize directly the channel capacity of a human being — other than by observing the actual rates of transmission that can be experimentally achieved.

3. The theory has not really generated new problems to be studied in psychology. Rather it has caused re-examination and reformulation of old problems. In some cases (see Chapter 10) it has permitted several apparently disparate effects to be included within a single theoretical framework. The fact that old problems are being treated does not, unfortunately, mean that new data are not needed. A published experiment rarely fulfills the exact conditions another worker would like. More important, the isolation of sequential dependencies requires a new analysis of the raw data, and it is very rare indeed to find extensive publications of raw data.

4. Like a new mistress, information theory seemed at first elusive and full of promise. She was justification for both intensity and irresponsibility:

(*) See Cronback [1952, 1953], Glaser and Schwarz [1954], Hick [1951], Lord [1954], and Willis [1954].

a thing of perfection requiring little more than some experimentation to bear fruit. With the passing years, a more "mature" if less exciting relationship has developed. A reading of Licklider's [1954] transcription of a conference on information theory conveys this to some extent, and Cronbach [1955] has systematically presented some sobering views on the use of information theory in psychology. Although I shall try to summarize some of his points, I would suggest that his paper be read by anyone interested in applications to psychology.

Cronbach points out that many psychologists have accepted information theory, particularly the information measures, without adequate scrutiny of its underpinnings and relevance to the particular problem under consideration. Entropy is by no means the only measure one can use to summarize the relationship between two variables. And since it effects a very serious compression of the data, it is well to check that one is throwing away what one intends to and keeping what one wants. This amounts to saying that a logical rationale must be given for its use. Let me cite three general types of examples where the information model may not be as appropriate as some other models. First, some stimulus dimensions possess natural metrics, and if the subjects are thought to react, however crudely, to these metrics, then probably the entropy measure should not be used, since it completely ignores all metric information. Second, the information model for noisy systems is concerned with limiting behavior — with infinitely long messages and delays. Subjects invariably deal with finite messages and introduce comparatively short delays. It does not immediately follow that the model gives bad approximations for such cases, but it does suggest that caution is needed. Third, the model supposes that the destination is aware of and uses a good deal of the available statistical information about the source and the noise. In many actual and experimental situations, subjects have only the crudest knowledge of these probabilities, and even when they do know them, there is no *a priori* certainty that they will use this information. For a detailed discussion of these points, see Hake [1955a]. The model for statistical decision making under uncertainty (used in the statistical, not information theory, sense) may often be more appropriate than the information theory model.

With regard to using information theory for error analysis, Cronbach makes at least two points which are important. Information measures are completely insensitive to constant errors, they are only concerned with variable ones! Thus, if a subject continually interchanges responses to a pair of stimuli, information theory will treat this as error free behavior —

as indeed it is in one sense. Yet, for many purposes it is these constant errors that are of greatest interest. Second, information theory is concerned only with the existence of an error; it does not assign any value to it. When there is a metric involved, one often measures the seriousness of the error in terms of that metric — say, as the square of the distance. But even when no metric is available, different errors can be judged as having differential importance. For such problems, a notion of utility has to be introduced and the techniques of statistical decision theory, rather than information theory, seem appropriate (see Van Meter and Middleton, [1954]).

Before turning to the empirical studies themselves, a few words on how this material might have been organized. The theory introduces methods for dealing with three central concepts, and the applications could have been categorized according to which facet of the theory they employ:

1. Sequential dependencies. This would include all the applications which use information theory to deal with sequential data, as proposed by Miller and Frick. Chapters 7 and 13 are illustrative of this approach.

2. Noise. The applications, such as those of Chapters 10 and 11, which use the formalism of noisy communication to cope with problems where stimulus and response are not perfectly correlated, e. g., where there are errors of some type, would fall into this category.

3. Capacity and transmission. Those studies which employ the central theorems of information theory concerning rates of transmission and capacity would be placed in this category. Examples are Chapter 9, and, to some extent, Chapter 10.

7. THE ENTROPY OF PRINTED AND SPOKEN LANGUAGE (*)

***N*-Grams of Printed English.** A problem that has intrigued a number of authors, including Shannon, is the estimation of the entropy of printed English (or any other language, for that matter), i. e., the estimation of the average number of bits per letter in a written passage. Put another way, the problem is to characterize the average sequential dependencies in the written language. If we assume, as may be approximately true, that the English in one book or article is the typical output of a stationary source — the author — then in principle all we need do is calculate $p(j|i_1,i_2,\ldots i_N)$

(*) See Miller [1954 b] for a general discussion of the role of information theory in the study of speech.

or all letters j and for all N-tuples of letters and blanks which might precede . From this we could then compute

$$F_N = -\sum_i \sum_j p(b_i,j) \log_2 p(j|b_i),$$

where b_i denotes a typical block of $N-1$ successive letters preceding j. Were these F_N known, then we could estimate the entropy of the sample to any desired accuracy using the fact that

$$H = \lim_{N\to\infty} F_N.$$

The difficulty becomes apparent when we realize that a 27 letter alphabet yields 27^N possible N-grams. Of course, many of these are impossible in English, but even were we to assume that, say, only one per cent were possible, there would still be 1,968 cases to be examined with $N = 3$, and 53,144 for $N = 4$.

Nonetheless, F_N can be computed for very small values of N, and Shannon [1951] reports that

$$F_1 = 4.14 \text{ bits/letter}$$
$$F_2 = 3.56 \text{ bits/letter}$$
$$F_3 = 3.3 \text{ bits/letter.}$$

His calculations are based on the letter, digram, and trigram frequencies which were prepared for coding work (Pratt [1942]). Not only is it practically impossible to carry this approach much further, but Shannon suggests that F_3, and all higher F's, may be liable to some error since many of the N-grams in the sample will bridge across two words. It is clear that other approximate techniques are necessary.

Three proposals have been made. The first employs, in one way or another, the built-in knowledge of English statistics in English-speaking people. The second attempts, by an assumption, to by-pass the sampling difficulties of the direct procedure discussed above. The last utilizes the known empirical distributions of English words, though ignoring the statistical dependencies among words, to determine an upper bound on the entropy. We shall discuss the proposals in this order; however, first let us examine an analogous N-gram calculation for spoken language.

N-Grams of Spoken Language. As I indicated earlier, there is every reason to suspect the entropy of spoken language, in bits per sound, is a

more basic statistical description of language than the entropy of the corresponding written language. The former studies have lagged behind the latter — probably because they are more difficult. There are two published papers that I know of: Cherry, Halle, and Jakobson [1953] and Black [1954]. In the former a long sample of Russian prose was analyzed into 42 phonemes and the following entropy estimates were obtained:

$$F_1 = 4.78 \text{ bits/sound},$$
$$F_2 = 4.23 \text{ bits/sound},$$
$$F_3 = 3.05 \text{ bits/sound}.$$

Black estimated F_1 and F_2 for English from a sample of one and two syllable words; however the sample possessed certain peculiarities making it not statistically representative of English in general. Black felt the fact that it contained only root forms, present tenses, etc. was a serious draw back.

Were the 41 phonemes in the sample independent and equiprobable, the information transmitted per sound would be 5.35 bits. From the actual simple and digram frequencies, Black obtained:

$$F_1 = 4.15\text{—}5.04 \text{ bits/sound},$$
$$F_2 = 3.75\text{—}4.21 \text{ bits/sound}.$$

The range arises from the fact that he calculated separate estimates for each class of words having the same number of syllables and the same number of sounds per word. He noted the following trend: the sounds of the shorter words transmit more information than those of the longer ones.

For exactly the same reasons as with printed language, it is unlikely that this approach can be extended beyond trigrams, so we turn to the other attacks that have been made on printed English.

Estimates Based on Partial Deletion of Messages. In his original report, Shannon (Shannon and Weaver [1949], pp. 25–26) states that "The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters, is roughly 50 per cent." (The definition of redundancy was given in Chapter 3). In a later paper [1951] he cites his original estimate as about 2.3 bits/letter. He arrived at this figure using two techniques. First, he developed approximations to English using the published frequencies, digram, and trigram frequencies

of letters and the frequencies and digram frequencies of words to generate approximations to English. The redundancies in each case were calculated; in the last two cases some extrapolation was required, since the tables were not complete. Second, he selected some unexceptional passages of English from which he randomly deleted a certain percentage of the letters. His subjects(*) then attempted to reconstruct the original passage from the mangled one, and he found that the letters could be restored with high accuracy when 50 per cent were deleted, from which he concluded that the redundancy must be at least 50 per cent.

Chapanis [1954] carried out roughly the same study, but his was an extensive and careful experiment using 91 subjects and 13 passages of 300 units (letters, space, and punctuation marks) each. Deletions were made, with no indication where they occurred, in both random and regular fashion, with 10, 20, 25, 33.3, 50, and 66.7 percent removed. His results are interesting, especially since they differ somewhat from Shannon's. At best one would conclude from these data that one-quarter of a passage can be deleted with a fair degree of recovery, but with a 50 percent rate of deletion the percentage of items restored is only about 20 percent, and of these only about one tenth are correct. Even when only 10 percent of the passage is deleted, only 80 percent of that supplied by the subjects is correct. Both the passages and the subjects showed considerable variability. Some passages were comparatively easy for most subjects to supply the missing letters and marks, others, particularly those judged easiest to read by conventional criteria, were uniformly more difficult to complete. The performance of the subjects was highly correlated with verbal and mental ability as measured by standard tests.

With respect to the disagreement between these results and those Shannon mentioned, Chapanis writes: "Dr. Shannon and I now agree that *S*'s can probably reconstruct from about 80% to 100% of deleted text under the following special conditions: (a) The amount deleted is 50%; (b) The deletions are made by taking out every other space, letter, or punctuation mark (other kinds of 50%, regular deletion patterns are more difficult to reconstruct); (c) The *S* is told, or can easily discover, the deletion pattern; (d) The *S* has a high amount of verbal intelligence.

"The results of the present study are valid under these conditions: (a) The *S* is provided with no supplementary information about the amount or kind of deletion; (b) The total context of the situation is such that *S* cannot or does not discover the deletion patterns." ([1954], p. 508).

(*) Mostly he and his wife according to Chapanis [1954, p. 496].

The obvious variant in which the deletions, either random or regular, are indicated by dashes appears not to have been run.

In any case, these results are certainly no more than a lower bound on the redundancy of a language, and probably not a very good one at that. For although the redundancy may be 50% or higher according to other estimates, the removal of 50% of the letters gives the subject a good deal of freedom to reconstruct the message. Once he is on the wrong track, say in the first two or three missing letters, then everything else is almost bound to be in error up to the point where a continuous sequence of three or four letters is not deleted.

Shannon's Upper and Lower Bounds. In his 1951 paper, Shannon carries his estimation procedures further by developing both upper and lower bounds for the entropy, and these data indicate that the redundancy may be nearer 75 per cent than 50 per cent. He selected 100 samples of English text, each consisting of 15 letters. A subject was required to guess at the first letter of a passage until he obtained it correctly. Knowing it, he guessed at the second until it was obtained. In general, knowing $N-1$ letters he guessed at the N^{th} until he was correct. The data may be presented as a table having 15 columns and 27 rows (26 letters and a blank). The entry in column N and row S is the number of times subjects guessed the correct letter on the S^{th} guess given that they knew the $N-1$ preceding letters. A small portion of the table is reproduced:

		N					
		1	2	5	10	15	100
	1	18.2	29.2	51	67	60	80
S	2	10.7	14.8	13	10	18	7
	3	8.6	10.0	8	4	5	–

The column marked 100 was obtained by presenting the subject with 99 letters from a 100 letter passage. The data for columns 1 and 2 were prepared from published word and digram frequencies which are based on far larger samples.

To use these data, Shannon introduced the notion of an ideal predictor who, knowing $p(b_i,j)$, i. e., the probability of all N-grams, would select letters j in order of decreasing probability for the given b_i. Thus each letter of a message can be replaced by a number between 1 and 27 which tells how many guesses will be needed before the correct letter is obtained. For an ideal predictor this sequence of numbers will contain the same informa-

tion as the message, since one can be constructed from the other, but it has the added feature that there will be limited statistical dependencies among the numbers, since the difficulty of one will not generally determine that of the next. Hence, computing the entropy of the number sequence is not difficult, and it can be used to estimate the entropy of the language.

The frequency of the number k in the reduced text will, of course, be given by

$$q_k^N = \sum p(b_i, j)$$

where the sum is taken over all $(N-1)$-grams b_i and over those j's such that it results in the k^{th} largest probability for the given b_i.

Shannon then shows that the N^{th} order entropy, F_N, is bounded by

$$\sum_{k=1}^{27} k(q_k^N - q_{k+1}^N) \log_2 k \leq F_N \leq \sum_{k=1}^{27} q_k^N \log_2 q_k^N .$$

Using the data described above, and smoothing them, Shannon calculated upper and lower bounds for $N = 1,2,\ldots,15,100$.
Some of the values are:

Upper and Lower Bounds on F_N

N	1	2	5	10	15	100
upper bound	4.03	3.42	2.7	2.1	2.1	1.3
lower bound	3.19	2.50	1.7	1.0	1.2	0.6

When both sets of points are plotted for $N = 1,2,\ldots 15$, there still remains some sampling error, but smooth curves can be faired through the points reasonably well.

It should be noted that there is a considerable drop in both bounds between $N = 15$ (at which point the curves are nearly flat) and $N = 100$. Whether or not this is meaningful is difficult to say, but, as we shall see, none of the other estimates suggests that the entropy is as low as 1.3 bits/letter; however, it must be kept in mind that all of these will be upper bounds, and how much too large they are is not known.

The Coefficient of Constraint. Newman and Gerstman [1952] approached the problem in another way which does not depend upon "built-in" knowledge of English statistics, but which does employ an as yet unproved assumption. They define

$$H(1) = -\sum_i p(i) \log_2 p(i)$$

and

$$H(1,\mathcal{N}) = -\sum_i \sum_j p(i,j) \log_2 p(i,j),$$

where i and j are letters in a passage which are separated by $\mathcal{N}-1$ others. That is, $H(1,\mathcal{N})$ measures the average statistical dependence of a choice j upon the choice i which was made $\mathcal{N}$ letters earlier. As $\mathcal{N}$ becomes large it is clear that this dependence decreases. A measure of its magnitude is

$$H_1(\mathcal{N}) = H(1,\mathcal{N}) - H(1).$$

They then define the quantity

$$D(\mathcal{N}) = 1 - \frac{H_1(\mathcal{N})}{H(1)},$$

which is called the *coefficient of constraint.* It is a quantity that is 1 when the $\mathcal{N}^{\text{th}}$ selection is uniquely determined by the first, and 0 when the $\mathcal{N}^{\text{th}}$ is independent of the first. Since only pairs of letters are involved in these quantities, it is comparatively easy to determine them for a given sample of language.

Using a 10,000 word sample from the Bible, they obtained the following data:

	2	3	4	5	6	10
$D(\mathcal{N})$	.223	.103	.064	.039	.027	.012

and a letter frequency entropy of 4.08, which is only slightly different from the 4.14 obtained by Shannon. A plot of these data on log-log paper is approximately linear with a slope of -2.0, or, in other words, $D(\mathcal{N}) = 1/\mathcal{N}^2$, approximately.

The problem now is whether $F_{\mathcal{N}}$ can be estimated from data for $D(\mathcal{N})$. The answer is "yes," provided it is true that

$$F_{\mathcal{N}} \leq [1 - D(\mathcal{N})] F_{\mathcal{N}-1}.$$

This relation is certainly true when $\mathcal{N}=2$. Indeed, the equality holds then. It is also true for any $\mathcal{N}$ such that the symbols are independent, for then $D(\mathcal{N})=0$ and $F_{\mathcal{N}}=F_{\mathcal{N}-1}$. They point out, however, that no proof of the assumption has been found, and they add without further elaboration the

cryptic comment "...and there are limiting cases in which it is proved not to apply." ([1952], p. 120). In any case, if it is assumed, one has

$$F_N \leq F_1 \prod_{i=2}^{N} [1-D(i)]$$

$$= H(1) \prod_{i=2}^{N} (i-1) \prod_{i=2}^{N} (i+1) \prod_{i=2}^{N} \frac{1}{i^2}$$

$$= H(1) \frac{(N+1)}{2N},$$

where the empirically grounded assumption that $D(i) = 1/i^2$ has been used. In the limit $H = \lim_{N \to \infty} F_N = H(1)/2$, which gives an upper bound, if the two assumptions are correct, of 2.04 bits/letter. In addition, for $N = 1,2,\ldots,15$ they computed $H(1)(N+1)/2N$ and compared these points with those obtained by Shannon as an upper bound. This curve seems to fit the points as well as Shannon's faired curve.

Distribution of Words to Estimate Letter Entropy. The third, and last, major approach to setting bounds on the letter entropy rests on a computation of word entropies which is based on known frequencies of word use in the language. This entropy, when divided by the average word length, affords an estimate of the letter entropy which is only an upper bound, since the technique, based as it is only on word frequencies, completely ignores the redundancy due to inter-word influences.

Long before information theory, people had determined the frequency of usage of various words, and it was Zipf [1949] who emphasized(*) that if we rank words $1,2,\ldots r,\ldots$ in order of decreasing frequency, then the frequency of use of a word is simply proportional to the inverse of its rank. That is, the probability p_r that a randomly selected word is of rank r is given, approximately, by

$$p_r = k/r,$$

where k is a proportionality factor *independent* of r. There is a certain ambiguity as to just how many ranks there are and certainly if we consider all

(*) This empirical observation was most systematically explored by Zipf, but it had been noted in certain cases by earlier authors [see Zipf, 1949].

possible English words the approximate law fails for very high ranks. The value of k is chosen so that the empirical relationship, known as Zipf's "law," holds for the lower ranks, and the size N of the vocabulary is given by the condition

$$\sum_{r=1}^{N} p_r = k \sum_{r=1}^{N} \frac{1}{r} = 1 .$$

Newman and Gerstman [1952], Miller [1951 b], and Shannon [1951] have all carried out this computation, but as Newman and Gerstman have pointed out, there are certain discrepancies in the results. Shannon obtained $N =$ 8,727, while Miller, presumably using a definite integral to approximate the series, got 22,000, and Newman and Gerstman obtained 12,370 by taking into account the discontinuity of the first 100 ranks and approximating the rest of the series by an integral.

Using this distribution, it is then possible to calculate the entropy of the independent word selections according to the distribution, i. e.,

$$H = -\sum_{r=1}^{N} \frac{k}{r} \log_2 \frac{k}{r} .$$

Shannon obtained 11.82, Miller 10.6, and Newman and Gerstman 9.7 bits/word. These give estimates of 2.62, 2.36, and 2.16 bits/letter if we take 4.5 letters to be the average word length. There appears to be a further disagreement, as was pointed out by Newman and Gerstman ([1952], p. 124). Considering the different values of N obtained, the Shannon and the Newman and Gerstman results should both be either larger or smaller than the Miller result; they are not.

Another approach to the problem from the point of view of words is due to Bell [1953]. He supposes that the space between words is sent infallibly and then he observes that the length of a word carries some information. "As the simplest example, consider the fact that there are only two words of one letter in normal use: the personal pronoun 'I' and the indefinite article 'a'. Hence only two out of the 26 single-letter 'words' which are mathematically available from the alphabet are admitted to the English language, and it follows that when a word of one letter is received in English the choice is only 1 out of 2 instead of 1 out of 26. An alternative expression of this is that the 'internal information' implicit in the fact that the 1-letter word is in the English language equivalent to a selection of 1 out

of 13 alternatives; and the communication of a selection of 1 out of 13 would be regarded as a communication of 3.7 'bits' of information ($\log_2 13 = 3.7$), so that the average internal information of 1-letter words in the English language may be stated as 3.7 bits per letter." ([1953], p. 384). For longer words such a detailed analysis is impossible, so Bell formed statistical samples from the dictionary. From this he calculated the internal information in bits/letter and obtained:

Number of Letters	1	2	3	4	5	6	7	8
Internal Information.....	3.7	2.2	1.53	1.93	2.36	2.66	2.98	3.21

This curve was smoothly extrapolated for words longer than 8 letters. Using Dewey's word list [1923] to obtain relative frequencies of words of various lengths he calculated the weighted average of the internal information and obtained 2.1 bits/letter.

The Role of Redundancy. Whatever the correct value of the letter entropy is, it is clear that it is not much over 2 bits/letter and not much less than 1. So the redundancy is somewhere between 50 and 75 per cent. In other words, under ideal conditions we could transmit the same information either by using a considerably smaller alphabet and keeping the length of books and articles the same, or by keeping the same number of symbols in the alphabet and reducing sentences and books to from one quarter to one half their present length. That our language is not fully efficient in this statistical sense presumably results from our need to communicate rapidly and accurately under adverse conditions, i. e., where there is noise: in the presence of other voices, in the wind, at sea, etc. It is clear from the little example given in Chapter 4 that even a small amount of noise can result in a serious drop in the information transmitted — in that case a one per cent chance of error resulted in a ten per cent drop in the entropy transmitted. It thus appears reasonable that if a language is designed to cope with even a slight amount of noise, then it must be quite redundant. Of course, when the noise level is so high that the natural redundancy of the language cannot combat it, other methods are used: Words and even whole sentences are repeated. And in such places as noisy factories the vocabulary between two people may be reduced to a few words — e. g., to "stop" and "go."

An example of a purposeful increase in redundancy is found in the very formal language used for air traffic control at an airport. Frick and Sumby [1952] have presented a summary of their findings for this language, but

without much of the data. They used the technique, introduced by Shannon [1951], of having subjects predict the next letter of a message. Using trained personnel as subjects they found that the uncertainty of control tower language is about 28 per cent that of random sequences of letters and spaces. And this, they point out, is a serious overestimation, since in practice the operator almost always knows the pilot's situation and therefore certain messages are excluded. To estimate these situational constraints, they described hypothetical situations to 100 Air Force pilots and asked them to predict the control tower message. Forming equivalence classes of "meaning units" and taking into account the imposed grammar of the language, they found that the uncertainty was no more than 20 per cent of what it would have been had the units been equally likely and randomly selected. The overall effect, they estimate, is a redundancy of about 96 per cent. This is not an implausible result when one considers the high noise level in both the tower and the plane, and the especially low margin of allowable error.

A similar study of tower-pilot communications at the Langley Air Force base has been presented by Felton, Fritz, and Grier [1951]. (Also see Fritz and Grier [1955].) As in the Frick and Sumby work, they divided messages into information elements, "...a word or a group of words representing a type of information, such as runway assignment, elapsed time, etc." ([1951], p. 5). They divided the analysis of redundancy into three levels. First, they simply took into account the frequencies of the various information elements. Second, they determined the predictability within a message. Third, they determined the predictability between messages from the observed conditional probabilities between messages. At the second level, they determined the probability of each message and determined the entropy of whole messages. This divided by the average number of elements per message was taken to be the entropy of each element. A justification of this procedure was given. The data were separated into messages originated in the air and at the tower, and the estimated redundancy using each of the three levels was:

	Redundancy		
Level	1	2	3
Air.....	.35	.72	.81
Tower	.26	.75	.78

The authors estimate that if contextual constraints are taken into account, as they were in the Frick and Sumby paper, then the redundancy would be about 93 per cent, which compares closely with the 96 per cent mentioned above.

8. DISTRIBUTION OF WORDS IN A LANGUAGE

IN THE LAST chapter, I referred to the empirically grounded observation (Zipf's "law") that if the words of a natural language are ranked from the most to the least common, then the frequency of the r^{th} word is approximately inversely proportional to r. Zipf found that more linguistic data could be fit by the more general equation

$$p_r = P\, r^{-B} ,$$

where p_r is the frequency of the r^{th} word and P and B are constants, B being in the neighborhood of 1 for all languages and larger than 1 for most. "Although this relation appears with regularity in linguistic data, no one has claimed more than a vague appreciation of its cause or significance. No one, that is, until Mandelbrot." (Miller [1954], p. 413) Mandelbrot's theory and its applications are presented in a series of five publications [1953 a, 1953 b, 1954 a, 1954 b, 1954 c], and Miller [1954 a] has given a very helpful summary of some of it.

Mandelbrot introduces two basic assumptions which distinguish his theory sharply from Shannon's. First, he supposes that a language — like all known ones — is built of discrete units called words, i. e., the communication is broken into units which are separated by a space. Second, he assumes that the transmitter in the communication system encodes and the receiver decodes word by word. Even though one could accept these as plausible assumptions that are valid for known languages, he defends each by a logical argument based upon the assumption that language must be designed to combat noise. In one paper [1954 a] he shows that the discrete character is needed, and in another [1954 c] he shows how the space and the word by word encoding limit the effects of noise to the word within which the error arises. The redundancy within the word is used to combat this noise.

"Although it may seem trivial, the introduction of the space between words is the crux of Mandelbrot's contribution and the main feature that leads him to results different from Shannon's. In Shannon's problem, the entire message is remembered and then coded in the most efficient form for transmission. In Mandelbrot's problem, the message is remembered only one word at a time, so that every time the space occurs the transmitter makes the most efficient coding he can of that word and then begins anew on the next word. Obviously, a transmitter of the kind Shannon studied will be more efficient, but one of the kind that Mandelbrot is studying will be more practical." (Miller [1954 a], p. 414).

Let us assume that the words are ordered by decreasing frequency of occurrence; denote them by $W_1, W_2, \ldots, W_R$. Let the corresponding frequencies of occurrence be $p_1, p_2, \ldots, p_R$. Suppose that to each word there is a cost C_r for using it — we do not specify what we mean by cost except that it can be summarized by a real number. It might be the number of bits required to transmit it, or the delay, etc. The first problem Mandelbrot attacked — he calls it the "direct problem" — is to find what the costs C_r should be so as to result in the least costly transmission of messages assuming word-by-word coding and the known frequencies p_r. This condition yields, as a first approximation,

$$C_r = [\log_M r],$$

where $[x]$ denotes the smallest integer greater than or equal to x. A better approximation is

$$C_r = [\log_M (r+m) + \log_M d],$$

where M, m, and d are constants independent of r. Observe that the cost depends upon the ranking, but not upon the details of the probability distribution.

Next, we turn to what Mandelbrot has called the "inverse problem." For this problem he assumed the words are given and their costs are fixed, and the task is to determine the frequency distribution p_r such that some economy criterion is met. He has given several criteria which all lead to essentially the same result.

1. Let us suppose that the average cost per word,

$$C = \sum_{r=1}^{R} p_r C_r$$

is fixed in advance, and we look for the best frequency distribution to transport information in Shannon's sense. That is, we maximize $H = -\Sigma p_r \log p_r$ subject to the above constraint. (This problem is formally identical to Boltzman's problem in statistical mechanics: to find the maximum entropy for a given average energy.) The following conditions are necessary and sufficient to solve the problem:

$$p_r = P' M^{-BC_r}$$

$$B > O$$

$$\sum p_r = 1$$

$$\sum p_r C_r = C.$$

The third condition determines P' and the fourth B, provided that $C < \log R$. Note that the cost C_0 of the space does not enter here.

2. A second condition, which is a trivial modification of the first, is to hold H fixed and choose the distribution so as to minimize the average cost C. The only resulting difference is that B is determined by the value of H, provided $H < \log R$. Again the value of C_0 is irrelevant.

3. A more interesting variant occurs when you allow R and C to be free and minimize the average cost per unit of information: i. e., minimize

$$\frac{\sum p_r C_r + C_0}{-\sum p_r \log p_r}$$

subject to the constraint $\Sigma p_r = 1$. As before, Mandelbrot has shown that

$$p_r = P' M^{-BC_r} ,$$

but now B is determined by the value of C_0, and so both the value of C and of H are fixed by the choice of C_0.

Finally, we turn to what Mandelbrot has called the "secrecy problem." He supposes that words are composed of letters $L_1, L_2, \ldots, L_G$, where G is much smaller than R. Let the letters be labeled in order of decreasing frequency, denote the frequency distribution by q_i, and write the cost of the i^{th} letter as c_i. The cost of a word is assumed to be given by the sum of the costs of its component letters.

"The best possible of all weighed vocabularies from the point of view of the secrecy encoder is the one in which the most economical code is also unbreakable. The code must then be a random sequence of elements, space included, and the enemy must either go to word relationships, that is go beyond our approximation, or try all keys, the number of which is astronomical." ([1954 b], p. 131.) His requirement is that an unbreakable random sequence of letters transport information for the smallest possible cost per unit of information. This is similar to condition 3 of the inverse problem, differing however in that there is no element corresponding to the word space. Formally, the condition is that

$$\frac{\sum q_i c_i}{-\sum q_i \log q_i}$$

should be a minimum subject to the condition that $\Sigma q_i = 1$. From this requirement it can be shown that the word distribution must be

$$p_r = P' M^{-BC_r}$$

as before, but with the added conditions that $B > 1$ and $R = \infty$. The latter condition follows from the requirement of a random sequence of letters to sustain secrecy. I shall discuss the condition $B > 1$ a little later.

Let me summarize: To attain the least costly transmission when words are ranked in order of decreasing frequency, then

$$C_r = [\log_M (r + m) + \log_M d].$$

To attain 1) the maximum information transport with the average cost per word fixed, or 2) the minimum average cost per word with the information transported held fixed, or 3) the minimum average cost per unit of information, then the distribution of the words should be

$$p_r = P' M^{-BC_r} .$$

If we combine these two conditions, taking into account the fact that statistical fluctuations in data will smooth over the steps of the former equation, we obtain

$$p_r = P(r + m)^{-B},$$

which Mandelbrot has called the "canonical curve." Observe that if $m = 0$, this is the generalized Zipf law.

As Mandelbrot points out, the fit of Zipf's law with $B = 1$ to most language data is good only in the central range and is in error for the most frequent and the least frequent words. By choosing values of B and m different from 1 and 0 he has been able to achieve far better fits.

The condition $B > 1$ which results from the secrecy criterion has been found to be met by most natural languages. Zipf called those with $B > 1$ "open vocabularies" and those with $B < 1$ "closed vocabularies." Most languages with closed vocabularies are in some way peculiar or special.

Clearly, Mandelbrot's theory, like Shannon's, is normative, but it is much more closely related to a specific empirical field than is Shannon's. Thus the question must be raised as to exactly what Mandelbrot has shown

and what it means for linguistics. "He says that if one wants to communicate efficiently word-by-word, then one must obey Zipf's law. There is a strong temptation to reverse the implication and to argue that because we obey Zipf's law we must therefore be communicating word-by-word with maximal efficiency." (Miller [1954 a], p. 415). Of course, Miller goes on to point out that much other evidence exists — such as the redundancy data discussed in the last section — to suggest that this reversed implication is false. It remains to be shown whether marked deviations in certain directions from perfect efficiency result in only slight deviations from the canonical curve.

It should be pointed out in connection with Mandelbrot's work that a totally different statistical explanation for Zipf's law has been offered by Simon [1955]. His model is not at all concerned with the transmission of information, but is rather of a more traditional statistical type. It has the advantage of suggesting the statistical process whereby the many phenomena other than word distributions are caused to satisfy Zipf's law. For example, Zipf noted that the distribution of cities by population and of incomes by size also satisfy roughly the same relationship. But many doubt that Simon's process accounts for word distributions. Nonetheless, using words in a book as the prototype, let $f(r,k)$ denote the number of different words each of which has occurred exactly r times in the first k words of the book. Simon then makes the following two assumptions concerning the selection of the $(k+1)^{st}$ word.

1. The probability that the $(k+1)^{st}$ word is one which has already appeared exactly r times is proportional to the total number of occurrences of all the words which have appeared exactly r times, i. e., it is proportional to $rf(r,k)$.

2. The probability that the $(k+1)^{st}$ word is a new word, i. e., one which has not already occurred in the first k words, is a constant α.

"If this process correctly describes the selection of words, then the words in a book cannot be regarded as a random sample drawn from a population with a prior distribution." (Simon [1955], p. 427).

From these assumptions, he shows that for a large sample, the probability p_r of words of rank r is given by $AB(r,\rho+1)$, where A and ρ are constants, and $B(r,\rho+1)$ is the Beta function of r and $\rho+1$, i. e.,

$$B(r,\rho+1)=\int_0^1 \lambda^{r-1}(1-\lambda)^{\rho}\,d\lambda=\frac{\Gamma(r)\Gamma(\rho+1)}{\Gamma(r+\rho+1)},$$

where Γ is the Gamma function. This, he shows, is very similar to Zipf's empirical law and it gives good fits for much of the data.

He explores modifications of this model which lead to essentially the same results and which appear to be more reasonable assumptions for the generation of word frequencies, but to examine this in detail would take us too far afield.

In closing this section, let me quote from Simon (p. 435) concerning the two alternative explanations of word frequencies:

"A very different and very ingenious explanation of the observed word-frequency data has been advanced recently by Dr. Benoit Mandelbrot [1953]. His derivation rests on the assumption that the frequencies are determined so as to maximize the number of bits of information, in the sense of Shannon, transmitted per symbol. There are several reasons why I prefer an explanation that employs averaging rather than maximizing assumptions. First, an assumption that word usage satisfies some criterion of efficiency appears to be much stronger than the probability assumptions required here. Secondly, numerous doubts, which I share, have been expressed as to the relevance of Shannon's information measure for the measurement of semantic information."

9. THE CAPACITY OF THE HUMAN BEING AND RATES OF INFORMATION TRANSFER

IN RECENT YEARS it has proved necessary to construct a variety of complex information systems in order to deal with certain military and industrial problems. These systems typically receive a tremendous amount of raw information from diverse sources that must be filtered, recoded, and correlated into what may be called a model of some situation of interest. The model must be sufficiently simple so that a person can grasp it completely, and sufficiently accurate so that it can lead him to useful decisions. For example, an air defense system receives raw information from radars, spotters, airline schedules, weather reports, fighter readiness reports, etc. All of this must be reduced to a simplified model of the enemy attack, the defense facilities, and the defensive response, so that a commanding officer can continuously know the situation with only a few seconds' delay. The officer must make and modify his defensive decisions on the basis of such a model. It is clear that much of this processing — especially where

speed and accuracy are needed — can and should be reduced to machine operations, but, with our present technology, there are certain steps which are far more simply and effectively carried out by a person than by a machine. For example, one of the first steps in an air defense system, and one which is not easily duplicated by a machine, is the isolation and transfer of pertinent information from a radar scope face. From all the random noise and background reflections on the scope an operator must single out those "blips" which represent aircraft. This he must introduce into the rest of the system, say, as a coded telephone message. The question arises as to how much information he can process per second over a sustained period.

It is clear that for any specific problem of this type, an answer can be obtained by direct experiments on the trained personnel using the equipment. On the other hand, one wonders whether it is necessary to study each new situation separately, or whether the pertinent variable is the amount of information in bits/sec which will be presented to the operator as compared with the maximum amount he can handle.

That is, can we treat a human being as a channel and so determine a channel capacity for him? If this were possible, it would certainly simplify the design problem, for it is generally not too difficult to determine the rate of the information flow in the machine components of a system. The question of whether it is useful to treat men as channels in certain situations remains at best an open problem, and there are some, equipped with strong arguments, who believe that it is an illusory hope. The most direct printed attack has been offered by Hake [1955 b]. The gist of his argument is that the information measure is impartial to many aspects of the stimulus set, e. g., to metric relations among the elements, to whether culturally assigned names exist for the individual stimuli, etc., and yet all sorts of experimental evidence suggest that subjects respond to these characteristics of stimuli. "It appears evident to me that a measure of information transmitted is meaningless unless accompanied with an operational definition of the experimental context. The possibility exists that we may discover invariant limits to information measures of performance within a single type of operation which I have described and across several stimulus-response systems. It appears unreasonable to describe such limits as the 'channel capacity,' however, when with a little thought and analysis the limit can be ascribed to some reasonable and known physiological limitation." ([1955 b], p. 253).

This debate, however, is not really my question here; I shall only

recount some of the studies which have been executed to determine the human channel capacity under the dubious assumption that a person can in fact be usefully considered as a channel.

Considering the theory presented in Section I, two procedures to estimate the capacity seem possible. First, estimate the channel capacity from whatever physical, physiological, and psychological facts that are known to be relevant to the type of transmission being employed. Second, by varying certain variables and employing diverse coding schemes, find the maximum amount of information that a person can be caused to handle. This, by the fundamental theorem of information theory, affords a lower bound on the capacity. Roughly speaking, the first procedure has resulted in upper bounds of the order of 10,000 bits/sec, while the second yields a lower bound somewhere in the range of 10 to 100 bits/sec. The consensus is that the lower bound more nearly represents the human capacity, but no really strong argument exists to support this view except that no one has yet devised a way to achieve a higher rate. Presumably, however, they are the more nearly correct and the upper bounds are so large because they ignore so many limitations of the "channel." We shall now examine these estimates in a little more detail.

Upper Bounds. Possibly part of the difficulty in obtaining a satisfactory estimate by the first procedure is the present lack of an adequate model for what happens functionally within a person when he is processing information. Thus, independent measurements on most of the "channel" — which is surely not homogeneous in its properties — cannot be had. As a result, the estimates which have been made are in a sense only concerned with the peripheral aspects of the channel. I will shortly cite another reason which has been offered to explain the difference between the upper and lower bounds.

Licklider and Miller [1951] have pointed out that an estimate of the capacity with respect to auditory signals can be obtained from a result of the theory of information for continuous systems (see the appendix). It is known that if the bandwidth of the channel is W cycles/sec, and if the noise and the signal are simply additive with a power ratio of P/N, then the capacity in bits/sec is given by

$$C = W \log_2 \left(1 + \frac{P}{N}\right).$$

For auditory signals, a bandwidth of 5,000 cycles/sec is conservative and a signal-to-noise ratio of 30 db, or a power ratio of about 1,000, is not unusual, in which case the capacity must be about 50,000 bits/sec. In actual attempts to transmit selective information by auditory means, a rate as high as 50 bits/sec is unusual. In other words, the efficiency of the auditory system must be considered to be about 0.1 per cent. Licklider and Miller, and Peterson [1952], offer the explanation that most of the information transmitted by an auditory signal is personal (and highly redundant) information about the originator — who he is, his way of speaking, his mood, and some of his linguistic history. While this may well be the case, it is interesting that no one has yet devised a way to use this apparently available capacity for the transmission of *preassigned* selective information.

A far more detailed, but rather questionable, estimate of auditory capacity has been made by Jacobson [1950, 1951 a] using various data about hearing, such as the total number of monaurally distinguishable tones. He concludes from his analysis, which ignores all sorts of possible interactions, that one ear should be able to handle about 8,000 bits/sec, and admitting very loud sounds, 10,000 bits/sec. It is known that there are approximately 29,000 ganglion cells from the ear, hence the average rate of information transfer over a nerve fiber is about 0.3 bits/sec. However, he points out that "It is very unlikely that there is any binary or similar coding in the cochlear nerves. It is consequently not particularly meaningful to state that the average informational capacity of a single cochlear fiber is about 0.3 bits/sec." ([1951 a], pp. 470–471).

Jacobson [1951 b] has also carried out a similar calculation for the eye, taking into account facts known about discriminability, etc., but ignoring the effects of color and of the interactions among the several dimensions he has considered. He obtained an estimate of 4.3×10^6 bits/sec for each eye. From this one can conclude the maximum average rate over each neural fiber must be 5 bits/sec. The inclusion of color would, of course, raise this estimate.

So far as I have determined, these are the only estimates of channel capacity which are based on measurements independent of the actual rate of information flow. We turn now to estimates of how rapidly information of a particular type can be, or rather, has been, caused to pass through a person.

Lower Bounds: Maximum Observed Rates of Information Transfer. Let us first consider the transmission of language encoded information.

Miller [1951 b] points out that if we consider the average measured length of vowels and consonants — about 12.5 sounds/sec — and if we were to suppose that they are equi-probable and independently selected, then speech would convey information at a rate of 67 bits/sec. If, however, we take into account their relative frequencies (Dewey, [1923]), then the rate is reduced to about 60 bits/sec. Further, if we take into account the fact that vowels and consonants tend to alternate in English (for more exact information on this for English and other languages, see Newmann, [1951]), the estimate is only 46 bits/sec. Finally, on the basis of Zipf's law, Miller estimated that there are 10.6 bits/word (Chapter 7). Since a speaker can sustain a maximum of about 3 words/sec, the transmission rate using speech can be no more than 32 bits/sec. "The maximum efficiency within the restriction imposed by the phonetic structure of English words, therefore, is about 50 per cent." (Miller [1951 b], p. 798). In practice, however, an ordinary speaking vocabulary is not nearly as large as that assumed when Zipf's law is employed, nor can a person usefully employ a speaking rate of 3 words/sec. An assumption of an equi-probable distribution over a vocabulary of 5,000 words which are spoken at a rate of 1.5 words/sec yields an information rate of 18 bits/sec.

In addition, as Quastler and Wulff [1955] point out, the various rate estimates using Zipf's law ignore constraints among words. They cite evidence which suggests that guessing a missing word within context may be correct as much as 30 per cent of the time. This reduces the information transmission rate to about 7 or 8 bits/word, and if we assume that 15 per cent of the words are incorrectly received, the estimate must be reduced to 6 or 7 bits/word. Using Miller's speaking rate of 1.5 words/sec, it appears that from 10 to 20 bits/sec is a good average rate of transmission, and that with rapid speech the rate may get as high as 25 bits/sec.

Quastler and Wulff report data on several other methods of information transfer, and in summary they find that 25 bits/sec seems to be the maximum rate. In all cases, a motor response was required of the subject, but they verified that mechanical limitations were not determining an apparent rate by showing that higher rates could be achieved if memorized materials were used. One experiment they discussed was based on typing, but it was known *a priori* that this would not lead to the fastest possible rates, since text can be read aloud faster than a typist can take it down. For this experiment, random sequences of letters were drawn from alphabets of 4, 8, 16, and 32 symbols. Seven experienced typists were paced by a metronome at 2, 3, 4, and 6 beats/sec. In general, the errors that oc-

curred were the transposition of letters, and so it is a question as to whether these should be treated as one or two errors. Depending upon our decision, the following upper and lower bounds on information transmitted (Chapter 5) are obtained.

Information Transmitted in bits/sec

Alphabet size	4	8	16	32
Upper Bound	6.7	10.5	13.2	16.7
Lower Bound	3.8	7.4	11.8	13.4

It was found, as would be expected, that with the higher metronome speeds and with the larger alphabets, the greater percentage of errors occurred. For 8 and 16 symbol alphabets a speed of 3.2 ± 0.2 keys/sec represented the highest effective speed, and beyond that their precision so decreased as to keep the transmission rate about constant, and beyond 4.5 keys/sec the quality of their output decreased very rapidly. With 4 symbols the effective speed was 3.6 keys/sec, and with 32 it was 2.9 keys/sec. When the subjects were not driven by a metronome, but were instructed to type as rapidly as possible, it was found that the rate of transmission was down about 9 per cent.

A second experiment drew upon the sight-reading ability of three young pianists. They were presented with random music (notes selected using random numbers) and they were paced by a metronome which was gradually increased in tempo over trials. Tape recordings were made and each of the subjects scored each of the tapes for errors. The agreement was fair, but both a low count (errors detected by each subject) and a high count (those detected by at least one) were determined. The information transmitted was computed from the error count and from assumptions about the error pattern. Again, several different "alphabets" were employed: 3, 4, 5, 9, 15, 25, 37 and 65 keys.

The data show that the highest speed for which the error rate remained low decreases from 7 keys/sec for an alphabet of 3 or 4 keys to 4.4 keys/sec for the 37 key alphabet. This decreased speed, coupled with an increase in error rate, keeps the information transmission rate at about 22 bits/sec over a fairly wide range of speed and alphabet size. However, for very small alphabets and for very large ones, the rate of transmission is less than for alphabets of 15, 25, and 37 keys. The interpretation given is that channel capacity is the controlling factor for the middle sized alphabets, that the sheer range limits performance in the largest ones, and motor limitations determine the performance when the alphabets are very small.

Individual differences became apparent when the subjects attempted to exceed their limits. One subject kept the error rate low by failing to keep up with the metronome, another kept the pace but allowed the error rate to become large, and the third held the pace for periods and then he would lose the beat. But in all cases, the information transmitted was held roughly constant.

Quastler and Wulff have studied a third set of materials for determining capacity: mental arithmetic problems. They point out that if certain plausible assumptions are made about the information involved in calculations, and if the published time data on so-called "lightning calculators" (people who are noted for rapid mental calculations) are used, one obtains an estimate of 22 to 24 bits/sec for the transmission rate. The feat of such people appears, therefore, not to be a high rate of information transmission, but rather a tremendous storage of information for short periods of time. In addition, Quastler and Wulff conducted some simple experiments on mental addition of columns of figures. On the average they found — again by making some plausible, but debatable, assumptions — a rate of 6 to 12 bits/sec, but one exceptional subject sustained a rate of 23 bits/sec.

From these data, and others not published, it appears that it is difficult to cause a subject, employing familiar operations, to exceed — let me be generous — 50 bits/sec, even though present estimates of ear and eye capacity exceed this several hundred times. It seems an open problem to bring these two estimates closer together, either by devising a method to employ much more of the apparent capacity to transmit selective information, or by a more detailed analysis of the human being as a channel to show that 50 or 100 bits/sec is truly his limit. Jacobson's comments on this disparity are of interest. "Thus it is evident that the brain can digest generally less than 1 per cent of the information our ears will pass. It must be appreciated that the ear is a channel vastly wider than its apprehensible output. It is the ability of the brain to *scan* for those portions of the auditory signal which are of interest which makes the wide capacity of the ear maximally useful." ([1951 a], p. 471).

It will be recalled that in the Quastler and Wulff piano experiment, the subjects appeared to be limited by motor factors rather than by "mental channel capacity" when the range of keys was small. Even so, one can raise this question: is there some sort of exchange between speed and error even in this range which keeps the information transmission nearly constant? If this were so, it would allow us to summarize a good deal of tra-

ditional data on motor performance in a comparatively simple way, as was pointed out by Fitts [1954 a]. He ran three experiments on motor performance which gave similar results; we shall describe one of them (a summary can also be found in Fitts [1954 b]).

The subject sits before a panel on which there are two plates (cross hatched in Fig. 5) and for short periods he is alternately to tap these with a stylus. He was instructed to try for accuracy, but within that limitation he was to perform as rapidly as possible. The stylus closed an electric

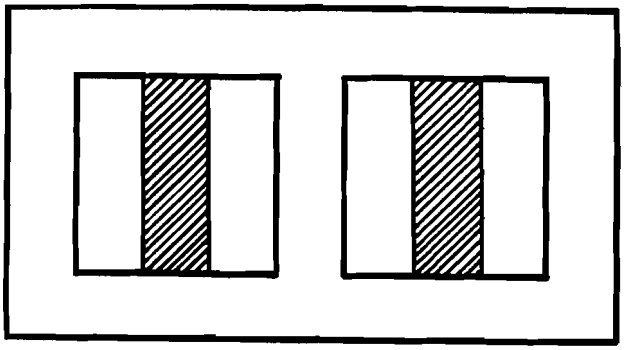

FIG. 5.

circuit when the cross hatched plate was touched, and another one which recorded errors when the outside region was touched. The variables controlled by the experimenter were the distance between the plates (i. e., the amplitude of the movement) and the size of the plate (i. e., the accuracy tolerance). The subject controlled the speed and accuracy of performance. Fitts' hypothesis was this: "If the amplitude and tolerance limits of a task are controlled by E, and S is instructed to work at his maximum rate, then the average time per response will be directly proportional to the minimum average amount of information per response demanded by the particular conditions of amplitude and tolerance." ([1954 a], p. 383).

On the basis of the continuous theory of information (see the appendix) he defined an index of difficulty

$$-\log_2 \frac{W}{2A} \text{ bits/response},$$

where W is the tolerance and A the amplitude measured in the same units. His hypothesis was that this quantity is linearly related to the average time per response, i. e.,

$$\frac{1}{t} \log_2 \frac{W}{2A} \text{ bits/sec,}$$

where t the average time in seconds per movement, is a constant. The data roughly confirm this hypothesis, e. g., the range is 10.3 to 11.5 bits/sec. for one weight stylus. However, at the extremes of amplitude and tolerance he studied, there seemed to be some indication that the information transmitted was falling off. Two other experiments of a similar nature gave similar results.

Other Observed Rates of Information Transfer. Not all the experiments, or the observation taken, on rates of information transfer have resulted in rates as high as those described.

Evidently the mode of presentation of the information vitally affects the rate at which it can be handled; if this conclusion is true, then the naive program outlined at the beginning of this section for determining the channel capacity of a human being must be modified to some degree.

In this connection the results of an experiment performed by Klemmer and Muller [1953] are of interest. The stimuli consisted of five lights arranged in an arc; a corresponding set of telegraph keys was arranged under the subject's fingers. The subject was to press the keys corresponding to those lights which were on. By using various numbers of bulbs — the subjects were told which ones would be employed — 1, 2, 3, 4, and 5 bits could be achieved in the presentation. In addition, the stimulus cycle, which consisted of lights on 50 per cent of the cycle and off the last 50 per cent, was presented at a rate of 2, 3, 4, and 5 cycles per second. The subjects were all trained on the apparatus for several weeks, and the practice curves indicated that they had completely stabilized by the time the experiment was performed.

For a fixed number of bits in the stimulus, it was found that by varying the rate of information presented there was a nearly linear increase in the transmitted information until a peak was reached, after which the transmission rate fell markedly. The location of the peak, and hence its value, was found to be an increasing function of the number of bits in the stimulus.

The approximate values of the peaks were:

	Information presented in bits/stimulus				
	1	2	3	4	5
Peak Transmitted Info. in bits/sec . .	2.7	4.0	5.8	8.4	10.5

The decay of performance following the peak is remarkable. In the case of a stimulus with 5 bits, the peak of 10.5 bits/sec occurred when the input rate was approximately 13 bits/sec. When the rate was increased to 15 bits/sec, the transmitted information dropped to 6 bits/sec. This drop is, of course, due to a radical increase in the error rate.

It should be mentioned that I am reporting average results, and the authors present data to show that there is considerable individual variation.

Now, it is clear that the maximum rates found in this experiment are less than those described above. In many respects this experiment and its conclusions are more closely related to those to be described in the next Chapter on reaction times than it is to either the reading, typing, or music experiments. One important difference is that in the latter experiments the stimuli are before the subject at all times and hence the receptor mechanism can operate with a considerable lead over the response mechanism, whereas such a large lead was not possible in Klemmer and Muller's study. It therefore appears to be more nearly a "continuously" executed reaction-time experiment. This can be supported from data they present. Typical reaction-time experiments were run on the same subjects, and a comparison of the inverse of the reaction time to the stimulus rate (in stimuli/sec) at peak transmission is revealing:

	Bits in Stimulus				
	1	2	3	4	5
$1/RT$	3.8	2.6	2.6	2.4	2.4
Stimulus rate at peak transmission	3.7	2.4	2.4	2.4	2.4

The Felton, Fritz, and Grier [1951] study of communications at Langley, discussed in Chapter 7, yields some data on operational rates of information handling. Using "information elements" on which to base their calculations, they found that during a single landing the following amounts and rates of information were employed by pilots and tower:

	Transmitted in bits	Rate in bits/sec
Air	114	8.4
Tower	133	10.3

However, it will be recalled that they determined that there was a very high redundancy in the transmission, and if only "new" information is considered, the table becomes:

	New Information Transmitted in bits	Rate of new information transmitted in bits/sec
Air.....	22	1.6
Tower.....	29	2.2

Either set of rates is below that which we have seen is possible for speech.

Hick writes, "As a personal speculation from such data as are available, it seems likely that transmission rates fall into three fairly distinct classes: —

1. High rates of 10-15 bits per second.
2. Moderate — 5-6 bits per second.
3. Slow — 3-4 bits per second." ([1952b], p. 68.)

He feels that these rates are closely correlated to the mode of presentation of the information. High rates are obtained only through simple "imitation" codes of the type we learn in childhood. Moderate rates are typical of "arbitrary" specially learned codes in which each signal has a high information content. The low rates result from arbitrary codes having a low information content per signal and a high rate of presentation. As a partial and speculative explanation for rates less than full capacity Hick comments: "But for various reasons I am inclined to suspect — I would certainly not be more definite than that — that there is a tendency, overcome, if at all, only with long practice, to sidetrack one or two bits per discrete movement as a kind of monitoring feedback. It would be originally necessary in the course of developing the skill (the code being, as stated above, relatively arbitrary or 'unnatural'), and may be retained, perhaps as a habit, or perhaps to keep the skill up to full efficiency, for a long time after that." ([1952b,] pp. 70-71).

Experimental results of Fitts and his colleagues (Fitts and Seeger [1953], Fitts [1954], Fitts and Deininger [1954], and Deininger and Fitts [1955]) tend to support part of Hick's position. For example, if the stimulus set is a circle of eight lights and the subject is required to respond according to the corresponding clock positions, the rate of information transfer is significantly higher than if some arbitrary numbering of the lights is used. The one clearly has a well engrained cultural basis, and the other does not. Such results, while hardly surprising, serve as a check on those who have too easily lapsed into speaking of the information in the stimulus set as the determiner of information transfer. "[These results] indicate that it is not permissible to conclude that any particular set of stimuli, or set of responses, will provide a high rate of information transfer; it is the ensemble of S—R combinations which must be considered." (Fitts and Seeger [1953], p. 209).

10. REACTION TIME AND INFORMATION TRANSFER(*)

OUR PRESENT TOPIC may, in a sense, be considered a continuation of the last Chapter on capacity; here we shall deal with what might be called "momentary" capacity. Previously we considered long samples of sequential stimuli to which the subject responded more or less continuously; now we shall consider his reaction time to a single isolated display. The question is what characteristics of the display need to be considered in order to account (simply) for the observed reaction times. The hypothesis, very generally, is that the information content of the display is the relevant variable and that the reaction time will turn out to be a very simple function of it — namely, linear.

There are, according to information theory, a number of ways in which the information transmitted can be varied: a) by varying the number of equi-probable alternatives, b) by altering the probabilities of the various choices, c) by introducing sequential dependencies between choices, and d) by allowing errors (noise) to occur. In the theory these are equivalent; whether they produce equivalent human responses is an empirical problem.

In the first experiment of the series of three I shall discuss, Hick [1952a] considered cases a and d. He presented subjects with a stimulus in which one of n equally likely alternatives would arise, and the subject had to respond as to which occurred. His hypothesis was that the reaction time (RT) would be proportional to the information in the stimulus, or, in other words, the rate of information transfer would be constant. There is, of course, a difficulty in assuming $RT = k \log n$, since when $n = 1$ this would require a zero reaction time. Hick suggests that there are really $n + 1$ alternatives, since we have ignored the case of no stimulus. Furthermore, he assumes that all $n + 1$ are equi-probable and that $RT = k \log (n + 1)$. This assumption is controversial and will be discussed below. Accepting it, he finds that data taken by Merkel [1885] are well fit by choosing $k = 0.626$ and that his own are fit with $k = 0.518$. Since a fixed delay, independent of n, seems plausible, the function $c + k \log n$ might seem intuitively more suited to fitting the data, but it does not fit either set of data as well. These fits were obtained with n in the range 1 to 10, i.e., up to a little more than 3 bits.

Turning to method d of varying the information, Hick points out, "...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." ([1952a], p. 15). An

(*) See Bricker [1955] for a survey of much the same material as discussed here.

experiment was performed in which the subjects were pressed, and the errors were taken into account by computing an equivalent error-free n, n_e. The reaction time data when plotted against n_e were found to be fit pretty well by the curve obtained for the errorless case.

In Hick's experiment the rate of information transfer was about 5.6 bits/sec, a value which is low compared with the largest obtained using a "continuous" stimuli presentation.

We must consider Hick's assumption that there are $n+1$, or in the more general case n_e+1, equiprobable alternatives. In a later paper [1954] he defends this choice as follows: "The discrimination between 'nothing' and 'something,' so to say, was practically perfect.

"But the discrimination between the n_e mathematical fictions was also perfect, by definition; and they are defined as equally probable. It is as if the subject were able to state with certainty — in an average sense, of course — which of the n_e+1 phases the environment was in. In other words, an impartial observer, having no reason to think one phase more probable than another, could receive $\log(n_e+1)$ units of information from him, per response. That is a fact, whatever the reaction time might happen to be, and it implies that the one extra possibility — that of no signal — can be regarded as having the same probability as any particular signal. Whether it really has is neither here nor there — the subject's channel capacity is such that it *can* have." ([1954], p. 400).

Mandelbrot(*) has suggested an alternative model based upon his theory discussed in Chapter 8. The fact that a response occurs, or doesn't, plays a special role, analogous to a space in ordinary language, which is quite different from the particular responses, which are treated as ordinary letters. Using his cost model, where reaction time is now the cost involved, the second approximation leads to

$$RT = \log(n+m) + T,$$

where m and T are unknown constants and n is the number of alternatives. This formula is related to, but different from, Hick's and it is not known whether it fits the data as well. The rationale for it, however, seems to me more substantial and better capable of being extended than does Hick's argument.

"The original evidence that the information measure was the appropriate one to use for interpreting choice-reaction times was simply that the logarithmic function occurs in both. This in itself is not strong, since logarithmic

(*) Personal communication.

relations occur rather often in biological measurement. The case became much stronger with Hick's finding that the reduction in response-time where errors are permitted obeyed the same law." (Crossman [1953], p. 41), Cronbach [1955] also stresses this important point, and he notes that the argument is made even stronger by Hyman's data, to which we turn now.

Hyman [1953] examined methods a, b, and c of varying the information when the performance was kept errorless. He states his hypotheses as,

"1) Reaction time is a monotonically increasing function of the amount of information in the stimulus series.

"2) The regression of reaction time upon amount of information is the same whether the amount of information per stimulus is varied by altering the number of equally probable alternatives, altering the relative frequency of occurrence of particular alternatives, or altering the sequential dependencies among occurrences of successive stimuli." ([1953], p. 189).

The stimuli were a matrix of lights with a range of 0 to 3 bits. The subjects responded by a vocal key, which seems to yield more precise measurements than the hand-operated key of Hick's experiment. The subjects were given complete statistical information about the stimuli and before each test run they were given sample sequences formed according to the appropriate statistics. Four subjects were used. The correlations reported below are the average of the four correlations computed for each subject separately.

In the first phase, the number of equi-probable alternatives were varied and a correlation of 0.983 was found between reaction times and information in the stimuli. This confirms Hick's results. In the second phase, when the relative frequencies were changed, an average correlation of 0.975 was found. In the third phase, introducing sequential dependencies resulted in a correlation of 0.938. The last correlation is significantly lower than the other two.

Hyman concludes from his data that his second hypothesis, while not acceptable at the 1 per cent level, is acceptable at the 5 per cent level. However, he points out two features in the data which suggest to him that the subjects did not react to the fine information structure of the experiment. In the third experiment there were cases where if a stimulus occurred, then in the next presentation it could not possibly occur, and the subjects knew this. Yet instead of reducing the reaction time, this increased it. This result seems disturbing.

He also raises this point: the reaction time to an event with probability p does not depend upon p alone, but also upon the probabilities of the other events in the display. Thus, although the average reaction time is deter-

mined by the average information in the display, the individual reaction times are not determined by the surprise — $\log_2 p$ — of the individual events. It is well to know this, but it does not seem to me to be either surprising or unfortunate, for the main thesis is that the subject responds to the overall statistical features of the display as described by the average information.

The relation between average reaction time and information has been further examined by Crossman [1953]. "When a subject responds to a sequence of signals all of which belong to a known set but some of which occur more frequently than others, his average response-time will be proportional to the average information per signal. This follows from the hypothesis that the subject deals with information at a constant rate." ([1953], p. 41). To test this he used a sorting task on ordinary playing cards. By varying the dimensions on which they were to be sorted he was able to examine the reaction times over a range of 0 to 2 bits/card. The correlation between reaction time and information per card was 0.86, and when the data are plotted it appears that no simple curve will fit them better than a straight line.

Crossman adduced evidence to show that the deviations from linearity were due to differential difficulties in discriminating the cards in different classes. On the basis of this he made the important observation that there is "...a major difficulty in the use of information theory in psychology, for information theory in the discrete case stated by Shannon says nothing about actual signals and the process of distinguishing them one from another; it deals only with abstract symbols already identified and distinct." ([1953], p. 49). This, of course, suggests carrying out a similar experiment using only one dimension of discrimination and causing the entropy to vary along it. This was done and the fit was improved.

On the basis of his data, Crossman concluded "... our hypothesis that rate is constant under variation of relative probabilities is upheld by these observations, with the proviso that 'discriminability' of signals should be equal in a sense yet to be precisely defined." ([1953], p. 50).

From these data it seems reasonable to conclude tentatively that the rate of information transfer in a reaction time experiment is constant when the information in the stimulus is in the range 0 to 3 bits. Since this conclusion is not in conformity with the observations made with a "continuous" stimuli presentation, it would certainly be interesting to see whether the rate remains constant when there are more than 3 bits in the stimulus, and also to see whether an experiment can be found with the rate constant, but much larger than 5 bits/sec, for the range 0 to 3 bits.

11. VISUAL THRESHOLD AND WORD FREQUENCIES

IN THE EARLY 1950's there were a series of experiments performed on the relation between the visual threshold of word recognition (as given by tachistoscopic measurements) and the frequency of their occurrence. Originally, the program stemmed from work on the Bruner-Postman hypothesis that sentences which relate to things liked are recognized with less difficulty than those relating to things disliked. Evidence has accumulated that the major relation is actually between recognition speed and the frequency of occurrence of the word in the language. Howes [1950] cites data involving sentences, and Howes and Solomon [1951] cite similar data involving only words. In the latter case, word frequency counts were obtained from Thorndike and Lorge [1944] and there was found to be a correlation of about —0.7 between recognition time and the logarithm of word frequency. Howes [1950] and Miller [1951a] describe data taken by Solomon in which seven-letter Turkish words were used. These were written on cards which the subjects studied. Some words appeared on many cards, others on only a few, so there was differential exposure to these new words. A correlation of —0.96 was found between recognition time and log frequency. King-Ellison and Jenkins repeated Solomon's experiments with some slight variations, including the use of artificial five-letter words, and they obtained a correlation of —0.99. They point out a relationship to information theory is suggested, namely, that recognition time is a linear function of the information transmitted by a word. The earlier comment I quoted from Crossman is relevant here, namely, that logarithmic relations are so common in biology and psychology that more must be established before an information theoretic model is assumed.

On the other hand, one can argue that this result is *predicted* by Mandelbrot's model of language, provided that one is willing to make one assumption (see Mandelbrot [1954a]). It will be recalled that a central notion of his model is the cost C_r of a word of rank r, and this was left undefined in the general model. It is plausible that recognition time is this cost. If so, then by Zipf's relation, we know that a word of rank r has a probability

$$p_r = Pr^{-B},$$

or taking logarithms,

$$\log r = -\frac{1}{B}\log\frac{p_r}{P}.$$

But Mandelbrot showed, to a first approximation, that $C_r = \log r$, thus we conclude that recognition time should be negatively correlated to the logarithm of the probability of occurrence. The second approximation to the cost expression would lead to a slightly different prediction, and it would be of interest to see whether a careful experiment could discriminate between these two predictions in favor of the second and more exact one.

Work of Krulee, Podell, and Ronco [1954] is related to and consistent with the above data. They established the distance from the eye at which a symbol is first recognized and found a slight decrease in the mean distance as the number of alternatives was increased.

12. THE INFORMATION TRANSMITTED IN ABSOLUTE JUDGMENTS(*)

WHEN A SUBJECT is required to place stimuli varying along one dimension, such as size or loudness, into N simply ordered categories, such as the first N integers, then he is said to be making absolute judgments of the dimension of the stimuli. For example, the stimuli might be pure tones at 100, 150, 200, ..., 1,000 cycles/sec. Each time a tone is presented he must place it in a category as accurately as he can. It is clear that in general errors will occur of the form: a tone with a lower frequency than another will be put in a higher number category. It is also clear that the error rate can probably be diminished by reducing the number of categories. For example, if he must place the above stimuli in 21 categories, we may expect more errors than if he need only report whether a signal is below or above 500 cycles/sec, for then there will be little ambiguity in his mind except for those stimuli near 500 cycles. Such experiments have a long history, but there has always been some difficulty in summarizing the data — just how should the error picture be presented?

Garner and Hake [1951] pointed out that the matrix relating input stimuli to response categories, with the entries the frequencies of pairings between a stimulus and a category, can be treated (with the obvious normalization) as a noise matrix for a communication system, where the communication is of selective information from the stimuli to the experimenter via the subject as a channel. We may, therefore, compute the information of the stimulus set (which, of course, depends on the relative frequencies of

(*) An excellent summary discussion, and interpretation, of much of the data in this and the following two chapters, plus some not so immediately related to information theory, has been presented recently by Miller [1956].

presentation of the different stimuli) and the equivocation of the transmission, and the difference is the information transmitted. If for a certain type of absolute judgement it is found that 21 categories allow the transmission of 3 bits, then in principle as much can be transmitted using only 8 unambiguous categories. Choosing the categories so that there is no ambiguity, i.e., no errors, may be difficult, but Garner and Hake point out that if the errors have a Gaussian distribution the condition is equivalent to a criterion of equal discriminability.

In another paper (Hake and Garner [1951]) they cite the difference between the usual error analysis for experiments of absolute judgments and the proposed information theory analysis. An error analysis ignores the fact that if the error distributions do not overlap, there will be no ambiguity. The information analysis takes this into account, but, unlike the error analysis, it completely ignores the magnitudes of the errors. There are some applications where it is preferable to have a multitude of small errors, provided that there is never a single major one.

A number of applications of this proposal have been made to different classes of absolute judgments. Pollack [1952a] studied tones spaced equidistantly on a logarithmic frequency scale from 100 to 8,000 cycles/sec. The subjects had to assign a number to each tone presented. When there were 2 and 4 tones in the stimulus set, the transmission was perfect, 1 and 2 bits respectively. But with 8 and 16 tones, the curve became flat, and the average maximum transmission was 2.3 bits, or the equivalent of perfect identification among 5 tones. The best subjects reached the equivalent of only 7 tones. On the grounds that there are known to be 40 to 60 identifiable sounds associated with speech and music, Pollack felt that there must have been a serious underestimation of the information transmitted, and so he performed a series of auxiliary experiments to attempt to raise the value. Six different partitions of the frequency space were examined, and the frequency range was varied with the bottom held at 100 cycles/sec and the top moved from 500, 2,000, 4,000 to 8,000 cycles/sec. These variations, and similar ones in a later paper (Pollack [1953]), resulted in only a few percentage points change in the information transmitted. He suggested that the result is so low because of the acute sensitivity of the information measure to error, which we have mentioned earlier (Chapter 4). However, later results which I shall present below show how more information can be transmitted and so suggest indirectly that the low value found may be realistic.

Halsey and Chapanis [1951] have presented similar data on the number of absolutely identifiable spectral hues, and though they did not apply an

informational analysis, their findings are of some interest. The colors were identified sequentially from violet to red by numbers, and the subjects were familiarized with the number-color code until learning was completed. In a test using 10 hues and 20 judgments per hue, they found that two observers were correct in 97.5 per cent of the judgments. These hues were selected on the basis of several earlier experimental runs in which more hues were employed, but a lower accuracy was obtained. They note that absolute identifiability of 10 hues is considerably better than had been previously reported, but they attribute this mainly to different experimental conditions.

If we turn to the sense of taste, similar results hold except that the maximum amount of information transmitted is definitely less than for either pitch or hue. Beebe-Center, Rogers, and O'Connel [1955] report data for both sucrose and saline solutions with the number of stimuli varying from 3 to 17. The concentrations were chosen to be roughly equally spaced in jnd units. The information transmitted reached a peak of about 1.7 bits per judgement for sucrose and a range (for three subjects) from 1.6 to 1.8 bits per judgment with a saline solution. As the number of stimuli in the set were increased, there was a decrease in the information per judgment down to about 1 bit for 17 stimuli. The most notable aspect is that these data are equivalent to perfect discrimination among only three distinct stimuli, as compared with five to seven tones and possibly as many as 10 hues.

Hake and Garner [1951] applied an information theory analysis "... to determine the minimum number of different pointer positions which can be presented in a standard interpolation interval to transmit the maximum amount of information, not about which positions of the pointer are occurring, but about the event continuum being represented." (p. 358). Two variations were run: in the limited response case the subjects were told the values the pointer could assume and they were required to respond only with those numbers; in the unlimited response case no such restriction was made. 5, 10, 20, and 50 possible pointer positions were used, and the data are summarized below:

	Information Transmitted in Bits			
Number of Positions	5	10	20	50
Limited Response	2.31	3.14	3.16	3.19
Unlimited Response	2.29	3.03	3.11	3.41

We observe that beyond 10 pointer positions the amount of information transmitted is roughly constant — equivalent to about 10 errorless positions.

There seems to be little or no difference between limited and unlimited responses as far as this analysis is concerned, but Hake and Garner point out that an error analysis shows that the errors increase when the subjects are allowed unlimited response.

In a later paper, Garner [1953] comments: "A measure of information transmission provides a means of specifying perceptual and judgmental accuracy in situations where absolute judgments about various categories on a stimulus continuum are required. This measurement allows the determination of the maximum number of stimulus categories which could be used with perfect accuracy without the necessity of sampling all the possible numbers of categories. However, this use of information transmission requires the assumption that the inherent judgmental accuracy is independent of the number of stimulus categories used experimentally. Two experiments (Garner and Hake, and Hake and Garner) have shown that this assumption is quite valid for situations involving judgments of position in visual space, and Pollack's experiment demonstrates its validity for judgments of pitch" (p. 373). Garner then proceeded to examine its validity in judgments of loudness of tones using 4, 5, 6, 7, 10, and 20 stimulus categories and a corresponding number of response categories. He found that judgment accuracy was nearly perfect for 4 and 5 categories (perfect being 2 and 2.32 bits respectively), but that it had dropped to 1.62 bits for 20 categories, which is equivalent to perfect accuracy for only three categories. Thus, the assumption is apparently not valid for loudness.

He went on to show, however, that the information transmitted could be improved if both the observers, i.e., the subjects, and the stimuli were taken as inputs to the system and the responses as outputs. (See Chapter 5, Multivariate Theory, for the analysis procedure when there are more than two dimensions.) In other words, there was considerable variability among the subjects when a large number of categories were employed. A further raising of the information transmitted is achieved, so that there is no drop at all, if the stimuli, the observers, and the preceding stimulus are all taken as inputs to the system.

Ericksen and Hake [1955] have obtained data and given a similar analysis for judgments of size. Their interest was not so much with the value of the information transmitted as "...with the extent to which the number of absolutely discriminable stimulus categories can be affected by subjective anchoring effects associated with the range and density of the stimulus dimension and with the number of response categories available to *Ss* for expressing discriminations or judgments." (p. 323). Judgments were made

of squares in two ranges, 2 to 82 mm and 2 to 42 mm, with 5, 11, and 21 stimuli in each range and using 5, 11, or 21 categories. All combinations of number of stimuli and number of categories were examined. It was found, as might be expected, that for a fixed number of stimuli and of response categories, discrimination was better for the larger range than for the smaller one; the difference was significant, but slight (about 0.2 bits).

The interaction between the number of stimuli and number of response categories is shown in the following table:

Information transmitted in bits per judgment

Number of Stimuli		5	11	21
Number of	5	2.08	1.65	1.49
Response	11	1.93	2.07	1.90
Categories	21	2.03	2.14	2.08

We observe that when the number of stimuli match the number of response categories, the information transmitted is constant, and it is nearly so when the number of response categories exceeds the number of stimuli. There is, however, a distinct reduction if there are more stimuli than response categories. This is not unreasonable since as the number of response categories is reduced, the possible response entropy is reduced, so the error entropy would have to diminish an equal amount in order to keep the information transmitted a constant. Detailed study of the data show this did in fact happen for the larger stimuli, but not for the smaller ones. An explanation is given, which we need not enter into here, in terms of the characteristic end (or anchoring) effects of the method of absolute judgments.

Klemmer and Frick [1953] carried out an experiment similar in method and analysis to those above, except that there were two and three stimulus dimensions instead of one. They flashed (0.03 sec) a display consisting of white dots on a black background to subjects who marked on answer sheet grids what they thought the position of the dots to be. The experiment was run both with and without grid lines on the black background, but appreciable differences were not found in the data. With the situation restricted to the presentation of one dot, the information in the stimulus could be varied by changing the order of the matrix of possible positions. From 3.2 bits (order 3) to 5.2 bits (order 6) there was an increase in information transmitted from 3.2 to 4.4 bits. From 5.2 bits to 8.6 bits (order 20) in the display, the information transmitted remained approximately constant.

In addition, the number of dots presented was varied, and it was found

that by using 4 dots and a matrix of order 3 (7.0 bits), 6.6 bits were transmitted. Further, when from 1 to 4 dots were used, then a display having 8.0 bits resulted in almost perfect transmission —7.8 bits. "It is clear that the maximum amount of information that can be assimilated from a brief visual exposure is a function of the type of encoding used. The question immediately arises as to whether or not there is a common metric which may be applied to the different message classes and which will correlate with the maximum information-carrying capacity of that class." (Klemmer and Frick [1953], p. 18). They observe that using only one dimension or coordinate (the location of a point on a line) Hake and Garner found a maximum transmission of 3.1 bits, and using the two coordinates of a matrix plus the one of the number of dots, they found 7.8 bits transmitted. This suggests that the maximum increases with the number of dimensions.

This supposition is confirmed in data taken by other experimenters, particularly those reported by Pollack and Ficks [1954]. In the first of these studies, Pollack [1953] presented auditory stimuli which varied both in pitch and loudness, each dimension being represented by five stimuli roughly spaced at subjectively equal intervals. It was found that the multiple absolute judgments caused a slight reduction in the information transmitted in each dimension, and that the total information transmitted was a little in excess of the sum of the two dimensions analyzed separately(*) and a little less than their sum for the judgments made separately on the two dimensions:

Condition	Information transmitted per judgment in bits
1. Frequency alone, no report on loudness	1.8(†)
2. Frequency alone, loudness report given	1.6
3. Loudness alone, no report on frequency	1.7
4. Loudness alone, frequency report given	1.3
5. Combined frequency and loudness reports	3.1
6. Sum of 1 and 3	3.5
7. Sum of 2 and 4	2.9

(†) Note, this value is not as large as the rate Pollack [1952 a] reported earlier.

Roughly similar results were found by Beebe-Center, Rogers, and O'Connel when they combined the several possible mixtures of sucrose and salt (holding the amount of solvent constant); however, in every case the subjects transmitted slightly more information for the compound stimuli than the sum of the information transmitted for the two dimensions separate-

(*) Recall that McGill's multivariate model (Chapter 5) shows that this is possible because of the interaction term $A(uvy)$.

ly. The excess ranged from 0.04 to 0.20 bits in a total value of roughly 2 bits.

The most vivid demonstration of the increase of information transmitted with increased dimensions is given by Pollack and Ficks [1954]. In one display there were eight dimensions, which were achieved by presenting the subject with a stimulus composed of a tone and noise alternating in time. The eight variables on which he had to report were: frequency range of the noise, loudness of the noise, frequency of the tone, the loudness of the tone, the rate of alternation between the tone and noise, the fraction of time the noise was on, the total duration of presentation of the display, and the direction within the room from which the sound originated. In each case he was asked only to make a binary decision: high or low, loud or soft, fast or slow, etc. In a second variation, only the interupted tone was used and so there were only the last six of the above dimensions; however, in addition to the binary classification on each dimension, subjects were also run having to classify each dimension into 3 and 5 categories.

The subjects were separated into three classes of equal size according to the amount of information transmitted. The results for the poorest and the best classes are given:

Steps per Dimension	Information Transmitted in Bits per Stimulus		
	Maximum possible	Poorest Third	Best Third
two	6	4.8	5.6
two	8	6.4	7.4
five.....	13.9	6.2	7.8

Two aspects of these data are striking. First, the total amount of information transmitted is much greater than was possible using one dimension. Second, increasing the number of classification steps per dimension increases the information only very little as compared with increasing the number of dimensions.

Pollack and Ficks present data on the information transmitted for each of the separate dimensions and there are considerable differences. In the eight dimensional case, direction conveys the most, 0.97 bits, and frequency of the noise least, 0.78 bits. Furthermore, by considering the data for the finer subdivision they find "In general, dimensions associated with a high informational transfer... show a progressive increase in transmission with finer subdivision, whereas dimensions associated with a low information transfer... may show a maximum transmission with a cruder subdivision. Thus, the effectiveness of subdividing dimensions of elementary multidi-

mensional auditory displays is a function of the specific dimensions employed." ([1954], p. 158).

It is reasonably certain that eight dimensions is not the limit to increased information transfer and it would be interesting to know just how far this can be effectively extended. Of course, as Pollack and Ficks point out, such a method of increasing the information transmitted may not be useful in practice. Their study completely ignores the time parameter and presumably as the number of dimensions is increased the rate of information transmitted reaches a peak. Judging by the earlier results presented on rates, this peak can be expected at about eight or ten binary dimensions.

These results appear to tie into some recent work in linguistics. Jakobson, Fant, and Halle [1952] have attempted to show that the various speech sounds of natural languages can be classified according to a number of elementary binary linguistic characteristics: nasal or not, stopped or not, etc. It is thought that discrimination of sounds occurs by recognizing which state obtains for each dimension. Certainly, the above data would indicate that this is the most efficient way to use the auditory characteristics of human beings. (Also see Osgood [1954]).

On the basis of the several experiments we have discussed, one can conclude that for objective ratings there is, up to a point, an increase in the information transmitted with an increase in the number of categories. After that point the information transmitted either remains constant or decreases. Bendig and Hughes [1953] raised this question: Is the same conclusion possible for ratings of subjective feelings? To study this, they had subjects evaluate, according to either 3, 5, 7, 9, or 11 categories, their knowledge of 12 different countries. Anchoring statements of the form "I know (a great deal) (something) (very little) about this country" were employed in three variations: center anchored, both ends anchored, and both ends and the center anchored. Information transmission, they found, was increased by an increase of number of scale categories, except that there was a deceleration in the step from 9 to 11 categories. This is reconfirmed by Bendig [1953b]. This effect is in accord with the diminishing return observed for objective scaling. Bendig [1954] points out that it is also consistent with the hypothesis that the information transmitted is a constant proportion of the maximum possible, and he reports further data which substantiates this assumption.

13. SEQUENTIAL DEPENDENCIES AND IMMEDIATE RECALL, OPERANT CONDITIONING, INTELLIGIBILITY, AND PERCEPTION

ONE OF THE MAIN POINTS of the 1949 Miller and Frick paper was to bring to the attention of psychologists that in information theory they had a tool ideally suited to the characterization of sequential dependencies in the stimulus, in the response data, or in both. There appear to have been four areas of psychological study to which this observation has been applied: to the learning of written material as a function of the statistical dependencies in those materials, to the sequential responses obtained in operant conditioning, to the intelligibility of verbal material as a function of statistical dependencies within the material, and to the ability of subjects to perceive statistical dependencies in materials. I shall discuss them in that order.

Immediate Recall. "Briefly stated, the problem... is, How well can people remember sequences of symbols that have various degrees of contextual constraint in their composition? The experimental literature contains considerable evidence to support the reasonable belief that nonsense is harder to remember than sense. This evidence has suffered, however, from a necessarily subjective interpretation of what was sensible" (Miller and Selfridge [1950]). Using Shannon's method, Miller and Selfridge prepared N^{th} order approximations to English in the following manner: A sequence of N successive words was chosen at random from a connected text, and a subject was asked to imbed the passage in a meaningful sentence. The first word in his sentence following the original group of N words was recorded. The next subject was presented with the last $N-1$ words of the original passage plus the new word, and he placed this N-word passage in a sentence. The first word after the passage was recorded, and so on. In this manner they generated approximations of order 0, 1, 2, 3, 4, 5, and 7 in passages of 10, 20, 30, and 50 words in length. Using these approximations to English, plus meaningful text, a standard recall experiment was executed. With the passage length held constant, they found that the percentage of recall increases with an increase in the order of approximation to English. In particular, for the 30 and 50 word passages the recall of the 5^{th} and 7^{th} order approximations to English is very little different from the recall of text material of the same length — this notwithstanding the fact that the 5^{th} order is quite nonsensical and the 7^{th} order by no means would be considered English. With shorter passages, recall comparable to that of text was achieved for even lower values of N.

"The results indicate that meaningful material is easy to learn, not because it is meaningful *per se*, but because it preserves the short range associations that are familiar to the *Ss*. Nonsense materials that retain these short range associations are also easy to learn. By shifting the problem from 'meaning' to 'degree of contextual constraint' the whole area is reopened to experimental investigations." (Miller and Selfridge [1950], p. 183). For example, one may ask whether their conclusion is valid for the whole memory decay curve, or whether it holds only for short term memory.

Similar results have been found by Aborn and Rubenstein [1952] in a slightly different experimental situation. They devised an "alphabet" of 16 nonsense syllables which fell into four easily distinguished classes of four syllables each; this classification was shown to the subjects. From these syllables six classes of passages of 30-32 syllables were constructed. The members of the first class were formed by random selection of syllables, and the others had increasing amounts of organization. For example, class four passages were marked by commas into groups of four syllables, and the first syllable of each group was chosen from class one, the second from class two, etc. The subjects were allowed 10 minutes to study the formal organization of the passage on which they would be tested and then three minutes to learn the actual passage, after which they were asked to reproduce it as accurately as possible. The authors had two hypotheses: "(a) The amount of learning in terms of syllables recalled is greater as the organization of the passage is greater, i.e., as the average rate of information is smaller. (b) The amount of learning in terms of the information score, computed as the product of the number of syllables recalled and the average rate of information, is constant for all passages." ([1952], p. 261). The data verified the first hypothesis, but not the second. For the first four passages the total amount of information learned was constant, but it dropped in passage 5 and even more so in passage 6. The breaking point was between 1.5 and 2 bits/syllable. This result simply means that the subjects were unable to memorize enough syllables to keep the information score high when the information per syllable was very low. Both these findings are in conformity with those of Miller and Selfridge above.

These same authors have pushed the problem further in a later paper (Rubenstein and Aborn [1954]). They conjectured that the lack of constancy in the information learned as the degree of organization changed was due to both inadequate training and too short study periods. Using the same materials to form passages of from 1 to 4 bits per symbol and of length 80 symbols, they repeated their experiment with a 10-hour training period

and varied the study periods from 1 to 20 minutes. The previous results were not only reconfirmed, but strengthened: the amount of information recalled decreased with every increase of the degree of organization within the message, holding the study time constant. This was not merely a trend, but it was strictly true for each length study period. Consider two passages: it was found that the ratio of information recalled in the one of higher degree of organization to the one of lower degree is less than the ratio of the information per symbol in the two passages. And finally, the information recalled per unit of study time was a decreasing function of the total length of study time.

One appears to be able to conclude that meaningful text is made easy to recall (at least in part) by its redundancy, but that it is not correct to state that holding other things constant the amount of information recalled is constant.

Operant Conditioning. Frick and Miller [1951] have reported an application of their earlier ideas for the measurement of stereotypic behavior (Miller and Frick [1949]) to the operant conditioning of rate in a Skinner box. Two responses were observed: approach to food (A) and bar pressing (B). "Instead of the usual analysis in terms of the *rate* of responding to the bar, the results are analyzed here in terms of the *patterns* of responses" ([1951], p. 21). Three experimental phases were considered separately in the analysis: a) behavior prior to conditioning (operant level), b) conditioning behavior, and c) extinction behavior. During phase b a total of 300 reinforcements were applied.

In all phases the behavior was recorded as sequences of A's and B's, and the uncertainties — in terms of the index of behavioral stereotypy (redundancy) — were computed. It was found that "intersymbol" influences did not extend appreciably beyond two symbols, and the value of the uncertainty in phase a was 0.408 for two symbols. Such a high value when there has been no conditioning is a consequence of the fact that such a sequence as $AAAA$ had a probability of 0.732 of occurring; indeed, the behavior of the rats was more stereotyped before conditioning than after. "The training-period did not introduce order into randomness, but rather caused the animal to abandon one well organized pattern of behavior for another. This needs some qualification. The lower stereotypy after conditioning appears when we consider only the temporal order; when we try to predict which response comes next. If we tried to predict also when the next response would occur and how long it would last, then the conditioned

behavior would look less random than the pre-conditioned behavior." ([1951], p. 25).

Another simple way the data may be described is as points in a two-dimensional plot of $p(B|B)$ vs $p(A|A)$. In phase a of the experiment the rats were approximately at the point (0.9, 0.75). This high perseveration is, in large part, simply a reflection of the topography of the Skinner box, as can be seen from the fact that 96 per cent of the responses separated by less than 10 seconds are of the form AA and BB, while this is reduced to 52 per cent for responses separated by more than 80 seconds.

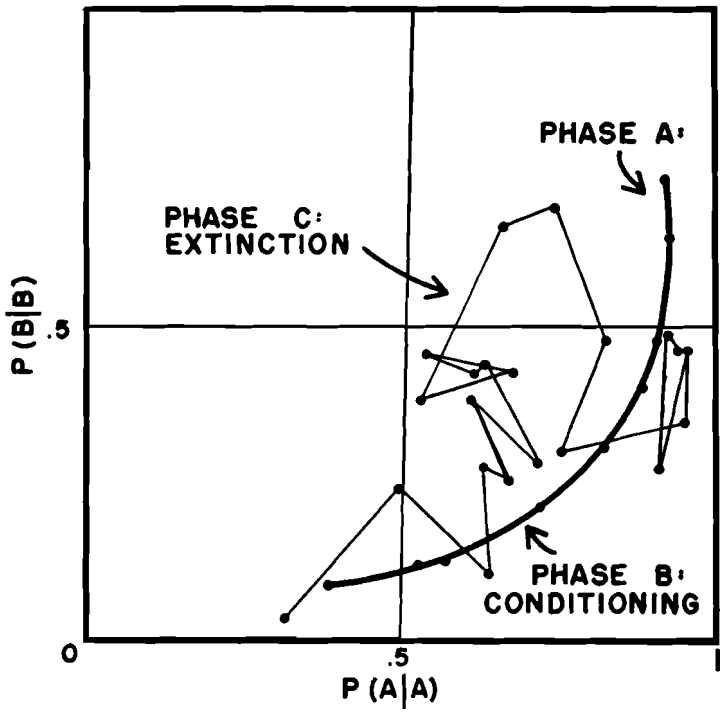

FIG. 6.

During conditioning, phase b, the rats initially move down the plot and then curve slowly over to an equilibrium point of about (0.4, 0.1), as shown in Fig. 6. During the extinction period the movement of a rat in this space is not very clear. There appears to be an initial tendency toward the center (0.5, 0.5) of the plot, or random behavior, but there is considerable random variation over a large portion of the plot. Over a 36-hour period there is a drift toward the initial resting point, but no stability is achieved in that period comparable to that prior to conditioning. It was not determinable

from these data how long it takes for the effects of reinforcement to wear off. As in phase a, there is little difference between the uncertainty determined from two successive responses and from more than two, and after some extinction there is little or no difference in the index based on a single response and that based on successive pairs of responses.

"The data presented and analyzed [in this paper] do not provide any startling new insights into operant conditioning. Most of the conclusions seem perfectly reasonable and obvious to anyone who has worked with rats in a similar situation and observed their general behavior closely. The impressive feature of such an analysis is the extent to which the qualitative aspects of the behavior can be incorporated into a completely quantitative account." ([1951], p. 35).

Intelligibility. The data on the effects of sequential dependencies on intelligibility are less detailed than for learning. There is an experiment by Miller, Heise, and Lichten [1951] in which certain gross effects were examined. They explored the effects of three different contexts on intelligibility, namely: 1) the test item is known to be one of a small vocabulary of possible items, 2) the test item is imbedded in either a word or a sentence, and 3) the test item is known to be a repetition of the preceding item. The materials used were digits, words in sentences, and nonsense syllables, and it was found that intelligibility decreased in that order. Further, the intelligibility of monosyllables, isolated words, and words in sentences was found to increase in each case as the domain of possible items was decreased. Only a very slight increase in intelligibility resulted from the knowledge that the item was a repetition of the preceding one. "The results indicate that far more improvement in communication is possible by standardizing procedures and vocabulary than by merely repeating all messages one or two times." ([1951], p. 335). This conclusion seems to confirm the military practice of using standardized languages when conditions are adverse, as in air traffic control (see Chapter 7).

Perception of Statistical Dependencies. Hake and Hyman [1953] raised the question of just how well and in what way people perceive sequential dependencies that are built into a set of stimuli. They chose to summarize their results in terms of certain conditional uncertainties — entropies — of the subject's responses. Their experiment was divided into four series of runs. Each run consisted of 240 presentations of one or the other of two symbols (H and V), and these presentations were generated according to the following probabilities and conditional probabilities:

	Series			
	1	2	3	4
$p(H)$	.50	.50	.75	.75
$p(H\|H)$	.50	.80	.75	.90
$p(V)$	.50	.50	.25	.25
$p(V\|V)$	.50	.80	.25	.70

Prior to each presentation, a subject was required to predict, or guess, which symbol would occur. The problem of analysis is to determine how accurately we can predict his guess provided we know certain past events such as his previous guesses and the symbols which actually occurred. For the last 120 trials the following conditional entropies were examined: the entropy of the guess y when only the distribution of y is known —$H(y)$, the entropy of y when the distribution of y and the previous guess are known —$H_y(y)$, the entropy of y when the distribution of y, the previous guess, and the previous occurrences are known —$H_{xy}(y)$, the entropy of y when the distribution of y and the previous occurrence are known —$H_x(y)$, and the analogues of each of the last three for the two preceding trials, instead of just one. These data are summarized:

	Series			
	1	2	3	4
$H(y)$	1.00	1.00	.76	.80
$H_y(y)$	1.00	.83	.72	.75
$H_x(y)$	.99	.69	.74	.70
$H_{xy}(y)$	.98	.54	.68	.55
$H_{y,y}(y)$	1.00	.83	.72	.73
$H_{x,x}(y)$	.98	.55	.70	.56
$H_{xy,xy}(y)$	.95	.52	.66	.55

It is clear that the best prediction of the subject's guess, i.e., the lowest entropy, is obtained when both his guesses and the actual occurrences on the two preceding trials are known, but a knowledge of his guess and the actual occurrence on the single preceding trial yields a prediction which is nearly as good, and knowledge of just the occurrence on the two preceding trials is only slightly worse. It thus follows that a subject responded not only to the actual events which occurred but also to his predictions about them. This can be made quite apparent by computing the probability of a guess of H when on the preceding trial a *correct* guess of H was made. For series one this conditional probability is about 0.5, but for the other three series it rises over trials and from trial 100 on it remains approximately constant with a value of 0.9. When the probability of an H guess following two

successive correct H guesses is plotted, the curves rise more rapidly, and even in series one there is a rise from 0.5 to about 0.75.

"We conclude from our evidence that *Ss* do not, in fact, perceive the probability rules by which the series of events was generated. They do perceive, instead, those short sequences of events which precede each prediction, which can be discriminated from other possible sequences, and which are found to provide some information about the future behavior of the symbol series. There are several interesting conclusions which we can make about the way in which *Ss* perceive these previous events.

"1. All combinations of possible previous events were not discriminated with equal ease. Some previous events, especially homogeneous runs of the same symbol, were more easily discriminated and consistently responded to than were others.

"2. The previous events to which our *Ss* responded on each trial included more than just the symbols which had been appearing. They included also the previous predictions of *Ss* and the degree of correspondence between their predictions and the symbols which appeared on previous trials.

"3. There was considerable agreement among our *Ss* as to *when* a particular symbol should be predicted. They tended to respond to some similar or identical previous events in the same way, no matter which series they were predicting..." (Hake and Hyman [1953], p. 72).

Bennett, Fitts, and Noble [1954] report an experiment similar to, but more complicated than Hake and Hyman's. The general structure was the same except that there were five symbols — in this case lights — rather than two to predict. Throughout the experiment the probabilities of the several symbols were held constant, but the digram and trigram frequencies were varied. First, they used one group of subjects to obtain information about sequential guessing habits, which the Hake and Hyman experiment (among others) indicated. The stimulus series consisted of independent selections; however, the subjects' response patterns differed both from independence and from the objective probabilities of the symbols. These data were used to generate conditional probabilities in the stimuli for the succeeding experiments — one sequence, called the concordant one, had sequential dependencies compatible with the observed guessing habits; the other, the discordant one, was designed to be incompatible with those habits. Using different subjects, learning was observed when better than chance behavior was possible if digrams were taken into account but not if only trigrams were considered. It was found that considerable learning did occur in 250 trials with the digrams, and that although there was some initial difference

between the concordant and discordant passages, it was not sustained. Using 500 trials and the trigrams, no learning was apparent. However, behavior for the two types of passages was sharply different: the concordant one elicited a larger number of correct predictions. The last experiment was repeated but extended to 1000 trials and the statistical structure of the passages was explained to the subjects. The differences between the two groups were markedly reduced and there seemed to be some slight indication of learning in the last 250 trials.

The authors point out that one must not conclude from these data that trigrams cannot be learned. With five symbols there are only 25 digrams as compared with 125 trigrams, which is a factor of 5. It is known from other studies that the number of trials required for learning goes up somewhat more rapidly than linearly with the number of stimuli. Furthermore, the four-second interval between trials during which the subject responded tends to prevent natural groupings such as are found in language, so the trigrams cannot easily be dealt with as a whole. "One implication of this line of reasoning, which may have important implications for skill learning, is that when sequences (of stimuli or movements) of a statistical nature have to be learned, it may be very important to give knowledge of results in such a way that *Ss* can observe entire sequences of events." ([1954], p. 311).

14. IMMEDIATE RECALL OF SETS OF INDEPENDENT SELECTIONS

THE SUBJECT OF THIS chapter is closely related to the first part of Chapter 13. The main emphasis of that Chapter was on the effects that inter-symbol dependencies have on immediate recall, whereas here we shall examine the effects of message length and the bits per symbol on immediate recall when there are no dependencies among symbols. Pollack [1952b] prepared messages of from 4 to 24 symbols from sets of 2, 4, 8, 16, and 30 equi-probable Latin consonants and numerals. These were read in a uniform manner to subjects who were told in advance both the set of symbols and the message length. They were required to reproduce them as accurately as possible. When an error was made, the subject was requested to guess as many times as was necessary to obtain the correct response. In one version of the experiment, reading rates were varied, but "Rate of presentation of stimulus materials (over the range considered) appears as a variable with little significance for immediate recall under the conditions considered here" ([1952b] II, p. 13).

The data show that the error entropy per message unit increases both with message length and with an increase in bits per symbol. But, for a message of given length, the percentage of presented information which is lost is approximately independent of the number of bits per symbol. This percentage is, however, an increasing function of the length of the message. The error entropy increased in such a manner that the total information transmitted increased as the message length was increased from 4 to about 10 symbols, it remained roughly constant in the range of 10 to 16 or 18 symbols per message, and it decreased for longer messages. The curves are displaced upward with an increase in bits per symbol, but they are of remarkably similar shape. "The main generalization is that one cannot obtain simultaneously both minimum information loss and maximum information gain by simply varying either the length of a message or the number of possible alternatives per message-unit." "These relations stem from the fact that the percentage of the information presented that is lost or gained is independent of the number of alternatives per unit and is simply a function of the length of the message." ([1952b], I, p. 12).

It is useful to transform these data into plots of error entropy and information transmitted *vs* total informational input. It is then found that for a fixed input, the error entropy is smaller and the information transmitted is larger the greater the number of bits per symbol. Thus, as Pollack points out, if one is interested in the optimal encoding characteristics for messages of fixed length, there are two answers depending upon whether a high error count is tolerable or not. If, however, the question is "What are the optimal encoding characteristics (for immediate recall) for messages of fixed informational content?" then the answer is unequivocal: short messages with a large number of alternatives for each message unit.

In parts III and IV of his report, Pollack systematically studied the error behavior of his subjects. First, his data confirm the familiar finding of this type of experiment that the subjects are most uncertain about the middle portion of the message. For messages of length 7, the relative uncertainty of the middle symbols is slightly higher than the end uncertainty, but it never exceeds .30. However, for messages of length 24, there is a broad plateau in the middle of the message which has a relative uncertainty of about .80. The broadness of this plateau Pollack attributed to the great sensitivity of the information measure to errors. He noted that the uncertainty curve alters its character with increasing message length: for short messages it is positively skewed and for long ones it is negatively skewed.

In the fourth part of the report, he established the conclusion that there

is still information transmitted (as compared with chance responses) by the subjects on the second and third guesses following an incorrect response. "In general, the additional information recovered per message increases as the degree of analysis of the multiple response data becomes more exhaustive. Stated otherwise, we recover more information from the distribution of responses if we utilize the first response following the initial incorrect reproduction, still more if we utilize the first and second responses following the initial incorrect reproduction, and still more if we utilize the first through the third responses following the initial incorrect reproduction. The magnitude of the information increases as the number of alternatives per message-unit increases and is, roughly, independent of message-length (for messages greater than 7 units in length)" ([1952b], IV, p. 8). As would be expected, this effect is a decreasing one, but the decrease is less rapid with larger numbers of alternatives per message-unit.

15. CONCEPT FORMATION

CONSIDER THE EIGHT objects that are characterized by the three "dimensions": triangles or circles, large or small, and black or red. One may attempt to convey to a subject a concept, such as red triangle, by showing him the objects one at a time and stating whether or not they are examples of the desired concept. A positive instance of the concept red triangle is "large red triangle," whereas "small black triangle" or "large red circle" are negative instances. Such experiments in concept learning have long been performed, and the conclusion has been drawn that negative instances are of little value in learning the correct concept. Hovland [1952], however, has raised a question about this conclusion — a question which stems from an information analysis of the situation. "What is not clear... is whether the ineffectiveness of negative instances is primarily attributable to their low value as carriers of information, or whether it is primarily due to the difficulty of assimilating the information which they do convey" ([1952], p. 461).

Certainly it is clear from the above example that positive and negative instances do not transmit the same information, since only two positive ones are required to specify the concept, as compared with six negative. It is, of course, possible to design a situation where the negative instances carry as much or more information as the positive ones. For certain simple general situations, of which the above example is illustrative, Hovland has given formulae for the total number of positive and negative instances re-

quired to specify the concept. In an experimental paper, he and Weiss [1953] examined the effect of positive and negative instances when both the number of instances and the amount of information are held constant, and they conclude that even so the negative instances do not contribute as effectively to learning. "At the same time the data disprove the generalization often cited that negative instances have no value in the learning of concepts. Under appropriate conditions over half of the *Ss* were able to reach the correct solution solely on the basis of negative instances." ([1953], p. 181).

Archer, Bourne, and Brown [1955] establish that additional but irrelevant information of from 1 to 3 bits diminish the rate at which a concept is achieved — the more irrelevant information, the slower the rate.

Bendig [1953a] conducted an experiment which is closely related to concept formation, namely, the identification of a concept after the manner of the game "20 questions." In the experiment, four questions were employed to isolate an animal topic. One experimenter asked the questions in fixed order of another who answered "yes" or "no" according to the topic. Following each question, the subjects were required to guess the concept. The information transmitted by each question was calculated, and theoretically each should have conveyed one bit, but in actuality 0.83, 0.91, 0.21, and 0.78 bits were transmitted. The central conclusion seemed to be that the third question was unfortunately phrased, since answers to it failed to convey much information.

16. PAIRED ASSOCIATES LEARNING

AS A FINAL APPLICATION of information ideas, I shall consider a learning situation where one class of objects — usually words — known as "responses" have been placed by the experimenter in one-to-one correspondence with another class of objects known as "stimuli." Initially, the subject knows nothing of the pairing and he can only guess at the appropriate response to a given stimulus; if he is correct, he is told this, if not, he is told the correct response. After a number of repetitions, R, of the stimulus class, the subject begins to learn the correct pairing, and he obtains a certain number of correct bonds, say C, out of the total of N. The function $C(R)$ is known as his "learning curve" for the paired associates. Several theories, and formulae, for this learning phenomenon have been put forth and are summarized

by Rogers [1952] in a thesis in which he introduces a new learning theory based in part on information theory.

He makes two central assumptions. First, he supposes that the uncertainty which a subject entertains about the stimulus class after R repetitions of the stimulus class is a function of R alone. In particular, he supposes that it is constant — U_{ck} — for the first b repetitions, where b is a "set" parameter which tells when the learning begins, and that from b on it is a linear function of R, i.e.,

$$U_k = U_{ck} - a\,(R - b), \text{ for } R \geq b.$$

Second, let B be the total number of bonds which the subject knows after R repetitions. Rogers shows this is one less than the expected value of the observable C. Let k be a stimulus not among the B that are known, and let i be any response which is not associated with one of the B known stimuli. Then he supposes that the probability that i is the response when k is given is $1/(\mathrm{N}-\mathrm{B})$. In other words, the subject is assumed to distribute his response choices without preference over all the available response elements.

From this second assumption, it is not difficult to obtain an expression for the uncertainty in terms of N and B. Equating this to the assumed expression in terms of R gives an equation between B and R, and so between C and R. This may be solved for C:

$$C = \begin{cases} (N-1)\{1 - exp[-\delta a(R-b)]\} + 1, & \text{for } R \geq b, \\ 1, & \text{for } R < b \end{cases}$$

where $\delta = \log_2 e$. It has long been noted that many learning data are approximately fit by such an exponential learning curve, though in general this has been an empirical observation which was not deduced from other assumptions.

To test the merits of this theory, Rogers drew certain conclusions from it which could be confronted by data. These conclusions were sustained by his data. Three related experiments were performed. 1) Correlated Structure. Stimuli — playing cards having two easily recognized dimensions, suits and denominations — were associated with nonsense syllables of the form consonant-vowel-consonant in a correlated manner. The first letter always corresponded to the denomination and the last to the suit. 2) Unstructured. Pictures of diverse household objects were paired in an arbitrary manner with nonsense syllables. 3) Uncorrelated Structured. The same materials as in 1 were used (so both the stimulus class and the response class were

structured) but there was no systematic relation in the pairing between the stimulus class and the response class. He then examined what two classical theories — Gestalt and the transfer theory of meaning — and what his own information theory of learning predict as to the learning rates in these three cases. Gestalt theory, according to his interpretation, ranks them 1, 3, 2 in order of increasing difficulty, transfer theory gives an ordering of 1, 2, 3, while information theory predicts that 1 and 2 should be equally easy and 3 more difficult. His data are consistent with only the last prediction.

Attempts to fit the learning curve to the data were for the most part successful, although one can note a consistent '*S*' character to the data, which, of course, the exponential does not possess. He points out that if the linear assumption were replaced by an appropriate non-linear one, one could easily produce a learning curve with an '*S*' shape — or, I might add, practically any other shape, for that matter.

APPENDIX
THE CONTINUOUS THEORY

MUCH COMMUNICATION can best be thought of as the transmission of a continuous signal and not as a sequence of temporally ordered selections from a finite set of possible elements. For the most part, as we have seen, the continuous theory has been of little importance in behavioral applications, though it is of considerable importance in electrical ones. I shall, therefore, briefly sketch the theory. This presentation follows Shannon's [1948] closely.

The Continuous Source. A source is said to be continuous if, in effect, it makes but one selection from a continuum of elements; specifically, if it chooses one number from the set of all real numbers. I shall suppose that this selection is characterized by a probability density $p(x)$ over the real numbers x. Since p is a density, $\int_{-\infty}^{\infty} p(x)dx = 1$ and furthermore for any $\epsilon > 0$, no matter how small, one can find finite a and b such that $1 - \epsilon < \int_{a}^{b} p(x)\,dx \leq 1$. Now, for such a and b we may divide the interval from a to b into n equal intervals, and we can treat each of the intervals as an element from a finite set, with probability $\int_{x_i}^{x_{i+1}} p(x)\,dx$ of being selected. All the continuum not

in a to b is an $(n+1)^{st}$ element with probability $1-\int_a^b p(x)\,dx$. Thus we have approximated the continuous source by a discrete one and for each n we can compute a corresponding entropy H_n. As we let n approach infinity, the approximation is better and better, but unfortunately H_n also approaches infinity. This, of course, is reasonable considering the basis of the discrete entropy concept, but that does not make the approach any more satisfactory as a way to compare continuous sources.

In such situations very often the difference between the quantity desired and another quantity, which tends to infinity with increasing n, will itself tend to a finite limit. If this second quantity can be chosen to be the same for all sources, then the resulting differences afford a perfectly acceptable comparison for continuous sources. As before, we choose a and b and we divide the interval from a to b into n equal intervals. Each of these intervals is of length $\Delta x = (b-a)/n$. Whereas before we tried to generalize

$$-\sum_{i=1}^{n} p(x_i)\,\Delta x \log_2[p(x_i)\,\Delta x]$$

and got into trouble, we now examine the entropy of the finite approximation minus the most that approximation might have been, i.e.

$$\log_2 \Delta x - \sum_{i=1}^{n} p(x_i)\,\Delta x \log_2[p(x_i)\,\Delta x].$$

It is not difficult to show that

$$\lim_{\substack{b\to\infty \\ a\to-\infty}} \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \left\{ \log_2 \Delta x - \sum_{i=1}^{n} p(x_i)\,\Delta x \log_2[p(x_i)\,\Delta x] \right\}$$
$$= -\int_{-\infty}^{\infty} p(x) \log_2 p(x)\,dx.$$

This quantity, which is denoted by $H(x)$, is called the *entropy* of a continuous source. It is well to keep in mind that the continuous entropy is not an exact analogue of the discrete entropy, and so certain differences in properties may be anticipated. The surprising thing is how many of the results are independent of the base-line from which the discrete entropy is measured.

If there are two arguments x and y to the distribution (as in the case of noise), the joint and conditional entropies are defined by

$$H(x,y) = -\int\int p(x,y)\log_2 p(x,y)\,dx\,dy$$

$$H_x(y) = -\int\int p(x,y)\log_2 \frac{p(x,y)}{p(x)}\,dx\,dy$$

$$H_y(x) = -\int\int p(x,y)\log_2 \frac{p(x,y)}{p(y)}\,dx\,dy,$$

where

$$p(x) = \int p(x,y)\,dy$$

$$p(y) = \int p(x,y)\,dx.$$

Many of the theorems of the discrete case carry over — usually quite directly — to the continuous case, but in addition there are certain new theorems which rest heavily on the existence of a coordinate system. I shall list some of the more important ones, of which the first is familiar and the other four are new.

1. $H(x,y) \leq H(x)+H(y)$,
   $H(x,y) = H(x)+H_x(y) = H(y)+H_y(x)$,
   $H_x(y) \leq H(y)$.

2. If $p(x) = 0$ except on an interval of length v, then $H(x)$ is a maximum $(=\log_2 v)$ when $p(x) = 1/v$ for x in the interval.

3. Of the class of all continuous one-dimensional distributions with variance σ^2, the normal, or Gaussian, is the one having maximum entropy. The value of the maximum is $\log_2 (2\pi e)^{1/2}\sigma$.

4. Of the class of all continuous one-dimensional distributions with mean $a>0$ and with $p(x) = 0$ for $x\leq 0$, the exponential is the one having maximum entropy. The value of the maximum is $\log_2 ea$.

5. Unlike the discrete case, in which entropy measures the randomness in an absolute way, the continuous entropy is a measure which is relative to a coordinate system. If the coordinate system is changed, the entropy is changed. This is not serious, however, since both the channel capacity and the rate of information transfer depend upon the difference between two entropies, and so they are invariant under coordinate transformation. Reich [1951a] states that he has shown the definition of information rate used by

Shannon is the only one of a broad class of possible definitions which is invariant under coordinate transformation.

The Channel Capacity. As in the discrete noisy case, the channel capacity C is defined to be the maximum rate of transmission $R = H(x) - H_y(x)$ obtained by considering all possible distributions. This is easily shown to be

$$C = \lim_{T \to \infty} \max_{p(x)} \frac{1}{T} \int\int p(x, y) \log_2 \frac{p(x, y)}{p(x)\, p(y)}\, dx\, dy.$$

One particularly important case in applications is that in which the noise is simply added to the signal and is independent of it. In that case the entropy of the noise can be computed. If we denote it by $H(n)$, then

$$C = \max_{p(x)} H(y) - H(n).$$

Of course, if there are restraints on the class of admissible signals, the maximization is taken subject to these restraints.

A simple, but very important, electrical application of the above theorem is to the case of a channel which has a bandwidth of W cycles per second (e.g., a telephone which will pass from 500 to 3,500 cycles per second has a bandwidth of 3,000 cycles per second), in which the transmitter has an average power output of P and the noise is white thermal noise (i.e., all frequencies are equally represented) of average power N. In this case the channel capacity in bits per second is

$$C = W \log_2 \left(1 + \frac{P}{N}\right)$$

Rate of Transmission. "In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with *exact recovery* at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

"This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in trans-

mission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. Of course, as the fidelity requirements are increased the rate will increase. It will be shown that we can, in very general cases, define a rate, having the property that it is possible, by properly encoding the information, to transmit it over a channel whose capacity is equal to the rate in question, and satisfy the fidelity requirements. A channel of smaller capacity is insufficient." ([Shannon and Weaver], 1949, p. 74).

The noise character of the whole system is, as before, given by a distribution $p(x,y)$ which states the probability density that the signal y is received when x is sent. The fidelity of the system is, roughly, an evalution of how different y is on the average from x. It is assumed to be a function of the noise, that is, if it is measured by a real number it can be written in the form $v[p(x,y)]$. Under quite broad conditions, which I shall not attempt to state here (see [Shannon], 1948), it can be shown that v can be represented as

$$v[p(x,y)] = \int\int p(x,y)\,\rho(x,y)\,dx\,dy.$$

The real-valued function $\rho(x,y)$ is essentially a measure of the difference between x and y and in computing the fidelity it is weighted according to the probability density of the joint occurrence of x and y. It may be illuminating to consider two very common electrical criteria of fidelity. The first is the root-mean-square criterion, namely,

$$\rho(x,y) = \frac{1}{T}\int_0^T [x(t)-y(t)]^2\,dt,$$

and the second is the absolute error criterion, namely,

$$\rho(x,y) = \frac{1}{T}\int_0^T |x(t)-y(t)|\,dt.$$

Now, the rate R of generating information corresponding to a given quality of reproduction (fidelity) v is defined to be the minimum R which is obtained by varying $p(y|x)$ with v held constant, i.e.,

$$R = \underset{p(y|x)}{Min}\int\int p(x,y)\log_2\frac{p(x,y)}{p(x)\,p(y)}\,dx\,dy$$

subject to

$$v = \iint p(x,y)\, \rho(x,y)\, dx\, dy.$$

With this definition, and with that of channel capacity given earlier, it can be shown that if a source has a rate R for a valuation of fidelity v, then it is possible to encode the output of the source and to transmit it over a channel with capacity C in such a manner that the fidelity is arbitrarily near v if and only if $R \leq C$. This is the fundamental theorem for the transmission of information in the continuous case.

BIBLIOGRAPHY

The following papers and books were examined in preparing this essay, and many — though not all — have been mentioned in the text. They include the central works on information theory and all of the published reports I have been able to find (as of early 1956) concerned with its application in psychology. The bibliography prepared by Stumpers (1953) is more general than this one in that it covers the whole area of Cybernetics and the applications of information theory in engineering and in the several behavioral sciences (as of early 1953), but it is not so complete for psychological applications.

Aborn, M., and Rubenstein, H., "Information Theory and Immediate Recall," *J. exp. Psychol.*, 44, 1952, 260—266.

Adelson, M., Muckler, F. A., and Williams, A. C., Jr., "Verbal learning and message variables related to amount of information," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 291—299.

Alexander, L.T., and Garner, W. R., "Information transmission in a tracking task," *Amer. Psychologist*, 7, 1952, 276 (abstract).

Archer, E. J., Bourne, L. E., Jr., and Brown, F. G., "Concept identification as a function of irrelevant information and instructions," *J. exp. Psychol.*, 49, 1955, 153—164.

Attneave, F., "The estimation of transmitted information when conditional probabilities are interdependent," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 118—122.

Augenstine, L., "The use of Illiac in determining distributions for information functionals," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 109—115.

Bar-Hillel, Y., "An examination of information theory," *Phil. Sci.*, 22, 1955, 86—105.

Bar-Hillel, Y., and Carnap, R., "Semantic information," *Communication Theory* (Willis Jackson, ed.), Academic Press, New York (1953), 503—511.

Beebe-Center, J. G., Rogers, M. S., and O'Connel, D. N., "Transmission of information about sucrose and saline solutions through the sense of taste," *J. Psychol.*, 39, 1955, 157—160.

Bell, D. A., "The 'internal information' of English words," *Communication Theory* (Willis Jackson, ed.), Academic Press, Inc., New York (1953), 383—391.

Bendig, A. W., "Twenty questions: an information analysis," *J. exp. Psychol.*, 46, 1953 a, 345—348.

—, "The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale," *J. Applied Psychol.*, 1953 b, 37, 38—41.

—, "Transmitted information and the length of rating scales," *J. exp. Psychol.*, 47, 1954, 303—308.

—, and Hughes, J. B., "Effect of amount of verbal anchoring and number of rating-scale categories upon transmitted information," *J. exp. Psychol.*, 46, 1953, 87—90.

Bennett,W. F., Fitts, P. M., and Noble, M., "The learning of sequential dependencies," *J. exp. Psychol.*, 48, 1954, 303—312.

Birdsall, T. G., "The theory of signal detectability," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 391—402.

Blachman, N. M., "Minimum-cost encoding of information," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 3, 1954, 139—149.

Black, J. W., "The information of sounds and phonetic diagrams of one- and two-syllable words," *J. Speech Hearing Disorders*, 19, 1954, 397—411.

Blackwell, D., and Girshick, M. A., *Theory of Games and Statistical Decisions*, John Wiley & Sons, New York (1954).

Blank, A. A., and Quastler, H., *Notes on the estimation of information measures*, Control Systems Laboratory Report R—56, University of Illinois, 1954.

Bricker, P. D., "Information measurement and reaction time: a review," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 350—359.

Carnap, R., and Bar-Hillel,Y., *An outline of a theory of semantic information*, Research Laboratory of Electronics, Technical Report 247, M. I. T., 1952.

Chapanis, A., "The reconstruction of abbreviated printed messages," *J. exp. Psychol.*, 48, 1954, 496—510.

Cherry, E. C., "A history of the theory of information," *Proceedings of the Institution of Electrical Engineers*, III, 98, 1951, 383—393; and *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953, 22—43.

—, "Organisms and mechanisms — an introductory survey," *The Advancement of Science*, 40, 1954, 393—397.

—, Halle, M., and Jakobson, R., "Toward the logical description of languages in their phonemic aspect," *Language*, 29, 1953, 34—46.

Cronbach, L. J., *A generalized psychometric theory based on information measure*, Bureau of Research and Service, College of Education, University of Illinois, 1952 (mimeographed).

Cronbach, L. J., *A consideration of information theory and utility theory as tools for psychometric problems*, Bureau of Educational Research, University of Illinois, 1953 (mimeographed).

—, "On the non-rational application of information measures in psychology," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 14—26.

Crossman, E. R. F. W., "Entropy and Choice time: the effect of frequency unbalance on choice response," *Quart. J. exp. Psychol.*, 5, 1953, 41—52.

Davis, H., "Applications of information theory to research in hearing," *J. Speech Hearing Disorders*, 17, 1952, 189—197.

Deininger, R. L., and Fitts, P. M., "Stimulus-response compatibility, information theory, and perceptual-motor performance," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 316—341.

Dewey, G., *Relative Frequency of English Speech Sounds*, Harvard University Press, Cambridge (1923).

Dolanský, Ladislav, and Dolanský, M. P., *Table of* $\log_2 \frac{1}{p}$, *and* $p\log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p}$, Technical Report 227, Research Laboratory of Electronics, M. I. T., 1952.

Elias, Peter, "A note on autocorrelation and entropy," *Proceedings of the Institute of Radio Engineers*, 39, 1951, 839.

Eriksen, C. W., and Hake, H. W., "Absolute judgments as a function of stimulus range and number of stimulus and response categories," *J. exp. Psychol.*, 49, 1955, 323—332.

Fano, R. M., *The Transmission of Information*, Research Laboratory of Electronics, Technical Report 65, M. I. T., 1949.

—, *The Transmission of Information — II*, Research Laboratory of Electronics, Technical Report 149, M. I. T., 1950 a.

—, "The information theory point of view in speech communication," *J. acoust. soc. Amer.*, 22, 1950 b, 691—696.

—, *Information theory, past, present and future*, M. I. T., 1954 (dittoed).

Faverge, J. M., and Patin, J., "Recherche sur la notation des épreuves composées de questions en vue d'améliorer la validité," *Travail hum.*, 17, 1954, 86—91.

Feinstein, A., "A new basic theorem of information theory," *Transactions of the IRE, Professional Group on Information Theory*, 4, 1954, 2—22.

Felton, W. W., Fritz, E., Grier, G. W., Jr., *Communications measurements at the Langley Air Force Base*, Human Resources Research Laboratory Report No. 31, 1951.

Fitts, P. M., "The information capacity of the human motor system in controlling the amplitude of movement," *J. exp. Psychol.*, 47, 1954 a, 381—391.

—, "The influence of response coding on performance in motor tasks," *Current Trends in Information Theory* (R. A. Patton, ed.), U. of Pittsburgh Press, Pittsburgh (1954 b), 47—75.

—, and Seeger, C. M., "S-R compatability: spatial characteristics of stimulus and response codes," *J. exp. Psychol.*, 46, 1953, 199—210.

Fitts, P. M., and Deininger, R. L., "S-R compatability: correspondence among paired elements within stimulus and response codes," *J. exp. Psychol.*, 48, 1954, 483—491.

Frick, F. C., "Some perceptual problems from the point of view of information theory," *Current Trends in Information Theory* (R. A. Patton, ed.), U. of Pittsburgh Press, Pittsburgh (1954), 76—91.

—, and Miller, G. A., "A statistical description of operant conditioning, *Amer. J. Psychol.*, 64, 1951, 20—36.

—, and Sumby, W. H., "Control tower language," *J. acoust. Soc. Amer.*, 24, 1952, 595—597.

Fritz, E. L., and Grier, G. W., Jr., "Pragmatic communications: a study of information flow in air traffic control," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 232—243.

Gabor, D., "Theory of communications" *Journal of the Institution of Electrical Engineers*, 93, III, 1946, 429—456.

—, "New possibilities in speech transmission," *Journal of the Institution of Electrical Engineers*, 94, III, 1947, 369—390.

—, *Lectures on communication theory*, Research Laboratory of Electronics, Technical Report 238, M. I.T., 1952.

—, "Communication theory, past, present, and prospective," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953 a, 2—4.

—, "A summary of communication theory," *Communication Theory* (Willis Jackson, ed.), Academic Press, Inc., New York (1953 b), 1—23.

—, "Communication theory and physics," *Transaction of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953 c, 48—59.

Garner, W. R., "An informational analysis of absolute judgments of loudness," *J. exp. Psychol.*, 46, 1953, 373—380.

—, and Hake, H. W., "The amount of information in absolute judgments," *Psychol. Rev.*, 58, 1951, 446—459.

—, and McGill, W. J., "The Relation between information and variance analysis," *Psychometrika*, 21, 1956, 219—228.

Glaser, R., and Schwarz, P. A., "Scoring problem-solving test items by measuring information," *Educ. psychol. Measmt.*, 14, 1954, 665—670.

Goldman, S., *Information Theory*, Prentice-Hall, New York (1953).

Grant, D. A., "Information theory and the discrimination of sequences in stimulus events," *Current Trends in Information Theory* (R. A. Patton, ed.), U. of Pittsburgh Press, Pittsburgh (1954), 18—46.

Hake, H. W., "The perception of frequency of occurrence and the development of 'expectancy' in human experimental subjects," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe, (1955 a), 257—277.

—, "A note on the concept of 'channel capacity' in psychology," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955 b), 248—253.

—, and Garner, W. R., "The effect of presenting various numbers of discrete steps on scale reading accuracy," *J. exp. Psychol.*, 42, 1951, 358—366.

—, and Hyman, R., "Perception of the statistical structure of a random series of binary symbols," *J. exp. Psychol.*, 45, 1953, 64—74.

Halsey, R. M., and Chapanis, A., "On the number of absolutely identifiable spectral hues," *J. opt. Soc. Amer.*, 41, 1951, 1057—1058.

Hartley, R. V. L., "Transmission of information," *Bell System Tech, J.*, 7, 1928, 535—563.

Hick, W. E., "Information theory and intelligence tests," *British J. Psychol.*, 4, 1951, 157—164.

—, "On the rate of gain of information," *Quart. J. exp. Psychol.*, 4, 1952 a, 11—26.

—, "Why the human operator?" *Transactions of the Society of Instrument Technology*, 4, 1952 b, 67—77.

—, "Information theory in psychology," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953, 130—133.

—, "The impact of information theory on psychology," *The Advancement of Science*, 40, 1954, 397—402.

Hockett, C. F., "An approach to the quantification of semantic noise," *Phil. Science*, 19, 1952, 257—261.

—, "A review of Shannon and Weaver: the mathematical theory of communication," *Language*, 29, 1953, 69—93.

Hovland, C. I., "A 'communication analysis' of concept learning," *Psychol. Rev.*, 59, 1952, 461—472.

—, and Weiss,W., "Transmission of information concerning concepts through positive and negative instances," *J. exp. Psychol.*, 45, 1953, 175—182.

Howes, D. H., *The definition and measurement of word probability*, Ph D. Thesis, Harvard University, 1950.

—, and Solomon, R. L., "Visual duration threshold as a function of word probability," *J. exp. Psychol.*, 41, 1951, 401—410.

Hyman, R., "Stimulus information as a determinant of reaction times," *J. exp. Psychol.*, 45, 1953, 188—196.

Jackson, Willis (Editor), "Report of proceedings, Symposium on Information Theory, London, 1950," *Transactions of the Institute of Radio Engineers Professional Group on Information Theory*, 1, 1953 a.

—, *Communication Theory*, Academic Press, Inc., New York, (1953 b).

Jacobson, H., "The informational capacity of the human ear," *Science*, 112, 1950, 143—144.

—, "Information and the human ear," *J. acoust. Soc. Amer.*, 23, 1951 a, 463—471.

—, "The informational capacity of the human eye," *Science*, 113, 1951 b, 292—293.

Jakobson, H., Fant, C. G. M., and Halle, M., *Preliminaries to speech analysis*, Technical Report 13, Acoustics Laboratory, M. I. T., Cambridge, 1952.

King-Ellison, Patricia, and Jenkins, J. J., *Visual duration threshold as a function of word frequency: a replication*, The Role of Language in Behavior, Technical Report Number 6, University of Minnesota, Contract No. N8 onr—66216.

Klemmer, E. T., "Tables for computing informational measures," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 71—77.

—, and Frick, F. C., "Assimilation of information from dot and matrix patterns," *J. exp. Psychol.*, 45, 1953, 15—19.

—, and Muller, P. F., Jr., *The rates of handling information: key pressing responses to light patterns*, HFORL memo Report No. 34, 1953.

Krulee, G. K., "Information theory and man-machine systems," *J. Operat. Res. Soc. Amer.*, 2, 1954, 320—328.

—, and Sinclair, E. J., *Some behavioral implications of information theory*, Report 4119, Naval Research Laboratory, Washington, D. C., 1953, 11 pp.

—, Podell, J. E., and Ronco, P. C., "Effect of numbers of alternatives and set on the visual discrimination of numerals," *J. exp. Psychol.*, 48, 1954, 75—80.

Kullback, S., "An application of information theory to multivariate analysis," *Ann. math. Statist.*, 23, 1952, 88—102.

—, and Leibler, R. A., "On information and sufficiency," *Ann. math. Statist.*, 22, 1951, 79—86.

Leonard, J. A., *The effect of partial advance information*, British Medical Research Council, A. P. U. report 217/54, 1954.

—, "Factors which influence channel capacity," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 306—314.

Licklider, J. C. R. (Editor), *Problems in human communications and control*, Paraphrased transcription of a conference sponsored by the National Science Foundation, 1954, M. I.T., Cambridge, dittoed, 203 pp.

—, "Quasi-linear operator models in the study of manual tracking," this volume, 1960.

—, and Miller, G. A., "The perception of speech," *Handbook of Experimental Psychology* (S. S. Stevens, editor), John Wiley & Sons, (1951), 1040—1074.

Lord, F. M., "Scaling," *Rev. educ. Res.*, 24, 1954, 375—392.

Luce, R. D., *Individual Choice Behavior*, John Wiley & Sons, (1959).

MacKay, D., "Quantal aspects of scientific information," *Philosophical Magazine* (series 7), 41, 1950, 289—311; and *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953, 60—80.

—, "The nomenclature of information theory," *Cybernetics* (Heinz von Foerster, ed.), Josiah Macy, Jr. Foundation, New York, (1951 a), 222—233; and *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953, 9—21.

—, "In search of basic symbols," *Cybernetics* (Heinz von Foerster, ed.), Josiah Macy, Jr. Foundation, New York (1951 b), 181—221.

Mandelbrot, Benoît, "Contribution à la théorie mathématique des jeux de communication, "*Publications de l'Institut de Statistique de l'Université de Paris*, 2, 1953 a, 1—124.

—, "An informational theory of the statistical structure of language," *Communication theory* (Willis Jackson, ed.), Academic Press, New York (1953 b), 486—502.

—, "Structure formelle des textes et communication: deux études," *Word*, 10, 1954 a, 1—27.

—, "Simple games of strategy occurring in communication through natural languages," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 3, 1954 b, 125—137.

—, "On recurrent noise limiting coding," *Proceedings of the Symposium on Information Networks*, Microwave Research Institute, Polytechnic Institute of Brooklyn, New York (1954 c), 205—221.

McGill, W. J., *Multivariate transmission of information and its relation to analysis of variance*, Report No. 32, Human Factors Operations Research Laboratory, M. I. T., 1953.

—, "Multivariate information transmission," *Psychometrika*, 19, 1954, 97—116; and *Transactions of the IRE, Professional Group on Information Theory*, 4, 1954, 93—111.

—, "Isomorphism in statistical analysis," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955 a), 56—62.

—, "The relation between uncertainty and variance," *Proc. 1954 Conf. Test Probl. Educ. Test. Serv.*, 1955 b, 37—42.

—, and Quastler, H., "Standardized nomenclature: an attempt," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 83—92.

McMillan, B., "The basic theorems of information theory," *Ann. math. Statist.*, 24, 1953, 196—219.

—, "Mathematical aspects of information theory," *Current Trends in Information Theory* (R. A. Patton, ed.), U. of Pittsburgh Press, Pittsburgh (1954), 1—17.

Merkel, J., "Die zeitlichen Verhältnisse der Willensthätigkeit," *Philos, St.*, 2, 1885, 73—127.

Miller, G. A., "Language engineering," *J. acoust. Soc. Amer.*, 22, 1950, 720—725.

—, *Language and Communication*, McGraw-Hill, New York (1951 a).

—, "Speech and language," *Handbook of Experimental Psychology* (S. S. Stevens, ed.), John Wiley & Sons (1951 b), 789—810.

—, "What is information measurement?" *Amer. Psychologist*, 8, 1953, 3—11.

—, "Communication," *Annual Review of Psychology*, 5 (Stone, C, P., and McNemar, Q., eds.), Annual Reviews, Inc., Stanford (1954 a), 401—420.

—, "Information theory and the study of speech," *Current Trends in Information Theory* (R. A. Patton, ed.), U. of Pittsburgh Press, Pittsburgh (1954 b), 119—139.

—, "Note on the bias of information estimates," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 95—100.

—, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychol. Rev.*, 63, 1956, 81—97.

—, and Frick, F. C., "Statistical behavioristics and sequences of responses," *Psychol. Rev.*, 56, 1949, 311—324.

—, Heise, G. A., and Lichten, W., "The intelligibility of speech as a function of the context of the test materials," *J. exp. Psychol.*, 41, 1951, 329—335.

—, and Madow, W. G., *On the maximum likelihood estimate of the Shannon-Wiener measure of information*, Air Force Cambridge Research Center, Technical Report, 54—75, 1954.

—, and Selfridge, J. A., "Verbal context and the recall of meaningful material," 63, *Amer. J. Psychol.* 1950, 176—185.

Munsow, W. A., and Karlin, J. E., "Measurement of human channel transmission characteristics," *J. Acoust. Soc. Amer.*, 26, 1954, 542—553.

Newman, E. B., "Computational methods useful in analysing series of binary data," *Amer. J. Psychol.*, 64, 1951 a, 252—262.

Newman, E. B., "The pattern of vowels and consonants in various languages," *Amer. J. Psychol.*, 64, 1951 b, 369—379.

—, and Gerstman, C. J., "A new method for analysing printed English," *J. exp. Psychol.*, 44, 1952, 114—125.

Osgood, C. E. (Editor), "Psycholinguistics: a survey of theory and research problems," *J. abnorm. soc. Psychol.*, 49 (4, pt. 2 — suppl.), 1954, 203 pp.

—, "Fidelity and reliability," *Information Theory in Psychology* (H.Quastler, ed.), The Free Press, Glencoe (1955), 374—384.

Patton, R. A., (Editor), *Current Trends in Information Theory*, Univ. of Pittsburgh Press, Pittsburgh (1954).

Peterson, G. E., "Applications of information theory to research in experimental phonetics," *J. Speech Hearing Disorders*, 17, 1952, 175—188.

Pollack, Irwin, "Information of elementary auditory displays," *J. Acoust. Soc. Amer.*, 24, 1952 a, 745—750.

—, *The Assimilation of sequentially-encoded information. 1. Methodology and an illustrative experiment. 2. Effect of rate of information presentation. 3. Serial position analysis. 4. The informational contribution of "wrong" responses.* Human Resources Research Laboratories, Memo Report No. 25, Washington, 1952 b.

—, "The information of elementary auditory displays. II," *J. Acoust. Soc. Amer.*, 25, 1953, 765—769.

—, and Ficks, L., "Information of elementary multidimensional auditory displays," *J. Acoust. Soc. Amer.*, 26, 1954, 155—158.

Pratt, Fletcher, *Secret and Urgent*, Blue Ribbon Books, Garden City (1942).

Proceedings of the London Symposium on Information Theory, 1950, see Jackson [1953 a].

Proceedings of the London Symposium on Information Theory, 1952, see Jackson [1953 b].

Proceedings of the Symposium on Information Networks, Microwave Research Institute. Polytechnic Institute of Brooklyn, New York (1955).

Quastler, Henry (Editor), *Essays on the Use of Information Theory in Biology*, University of Illinois Press, Urbana (1953).

—, *Information Theory in Psychology*, The Free Press, Glencoe (1955 a).

—, "Approximate estimation of information measures," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955 b), 124—139.

—, "Information theory terms and their psychological correlates," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955 c), 143—171.

—, and Wulff, V. J., *Human performance in information transmission*, Control Systems Laboratory Report No. 62, University of Illinois, 1955.

Rappaport, M., *The role of redundancy in discrimination of visual forms*, Ph. D. dissertation, The Ohio State University, 1954.

Reich, E., "Definition of information," *Proceedings of the Institute of Radio Engineers*, 39, 1951 a, 290.

—, "The game of 'gossip' analysed by the theory of information," *Bulletin of Mathematical Biophysics*, 13, 1951 b, 313—318.

Rogers, M. S., *An application of information theory to the problem of the relationship between meaningfulness of material and performance in a learning situation*, Ph. D. thesis, Princeton University, 1952, mimeographed.

Rogers, M. S., and Green, B. F., "The moments of sample information when the alternatives are equally likely," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 101—108.

Rubenstein, H., and Aborn, M., "Immediate recall as a function of degree of organization and length of study period," *J. exp. Psychol.*, 48, 1954, 146—152.

Savage, L. J., *The Foundations of Statistics*, John Wiley & Sons, New York (1954).

Schafer, T. H., "A basic experiment in detection," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 415—418.

Schützenberger, M. P., "Sur les rapports entre la quantité d'information au sense de Fisher et au sens de Wiener," *Comptes Rendus de l'Académie des Sciences*, 233, 1951, 925—927.

Senders, J. W., "Man's capacity to use information from complex displays," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 360—363.

—, "The effect of number of subjects on the estimate of transmitted information," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 368—372.

—, and Cohen, J., "The effects of sequential dependencies on instrument reading performance," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 282—289.

Shannon, C. E., "A mathematical theory of communication," *Bell System Tech. J.*, 27, 1948, 379—423 and 623—656.

—, "Communication theory of secrecy systems," *Bell System Tech. J.*, 28, 1949, 656—715.

—, "The redundancy of English," *Cybernetics* (H. von Foerster, ed.), Josiah Macy, Jr. Foundation, New York (1950), 123—158.

—, "Prediction and entropy of printed English," *Bell System Tech. J.*, 30, 1951, 50—64.

—, "Communication theory, exposition of fundamentals," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953 a, 44—47.

—, "General treatment of the problem of coding," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953 b, 102—104.

—, "The lattice theory of information," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 1, 1953 c, 105—107.

—, and Weaver, Warren, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1949).

Simon, H. A., "On a class of skew distribution functions," *Biometrika*, 42, 1955, 425—440.

Slepian, D., "Information theory," *Operations Research for Management* (J. F. McCloskey and F. N. Trefethen, eds.), The Johns Hopkins Press, Baltimore (1954), 149—167.

Stumpers, F. L., "A bibliography of information theory (communication theory — cybernetics)," *Transactions of the Institute of Radio Engineers, Professional Group on Information Theory*, 2, 1953; and Research Laboratory of Electronics, Technical Report, M. I. T., 1953.

Tanner, W. P., Jr., "On the design of psychophysical experiments," *Information Theory in Psychology* (H. Quastler, ed.), The Free Press, Glencoe (1955), 403—414.

Thorndike, E. L., and Lorge, I., *The teacher's word book of 31,000 words*, Bureau of Publications, Teachers College, Columbia University, New York (1944).

Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1 (see Jackson [1953 a]); 2 (see Stumpers [1953]); 3 (1954); 4 (1955).

Van Meter, D., and Middleton, D., "Modern statistical approaches to reception in communication theory," *Transactions of the IRE, Professional Group on Information Theory*, 4, 1954, 119—145.

Wald, A., *Sequential Analysis*, John Wiley & Sons, New York (1947).

Watanabe, S., "A study of ergodicity and redundancy based on intersymbol correlations of finite range," *Transactions of the IRE, Professional Group on Information Theory*, 4, 1954, 85—92.

Weaver, W., Peterson, G. E., and Davis, H., "Information theory: 1) Information theory to 1951 — a non-technical review, 2) Applications of information theory to research in experimental phonetics, 3) Applications of information theory to research in hearing," *J. Speech Hearing Disorders*, 17, 1952, 166—197.

Weinstein, M., *Stimulus complexity and the recognition of visual patterns*, Ph. D. dissertation, The Ohio State University, 1955.

Wiener, Norbert, *Cybernetics*, John Wiley & Sons, New York (1948).

—, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley & Sons, New York (1949).

Willis, R., "Estimating the scalability of a series of items — an application of information theory," *Psychol. Bull.*, 1954, 51, 511—516.

Wilks, S. S., "The likelihood test of independence in contingency tables," *Ann. math. Statist.*, 6, 1935, 190—196.

Woodward, P. M., *Probability and Information Theory, with Applications to Radar*, McGraw-Hill, New York (1953).

Zipf, G. K., *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press, Cambridge (1949).