Psychological Review
1996, Vol. 103, No. 3, 566-581

0033-295X/96/$3.00

Reconciling Simplicity and Likelihood Principles in Perceptual Organization

Nick Chater
University of Oxford

Two principles of perceptual organization have been proposed. The likelihood principle, following H. L. F. von Helmholtz (1910/1962), proposes that perceptual organization is chosen to correspond to the most likely distal layout. The simplicity principle, following Gestalt psychology, suggests that perceptual organization is chosen to be as simple as possible. The debate between these two views has been a central topic in the study of perceptual organization. Drawing on mathematical results in A. N. Kolmogorov's (1965) complexity theory, the author argues that simplicity and likelihood are not in competition, but are identical. Various implications for the theory of perceptual organization and psychology more generally are outlined.

How does the perceptual system derive a complex and structured description of the perceptual world from patterns of activity at the sensory receptors? Two apparently competing theories of perceptual organization have been influential. The first, initiated by Helmholtz (1910/1962), advocates the *likelihood principle:* Sensory input will be organized into the most probable distal object or event consistent with that input. The second, initiated by Wertheimer and developed by other Gestalt psychologists, advocates what Pomerantz and Kubovy (1986) called the *simplicity principle:* The perceptual system is viewed as finding the simplest, rather than the most likely, perceptual organization consistent with the sensory input[1].

There has been considerable theoretical and empirical controversy concerning whether likelihood or simplicity is the governing principle of perceptual organization (e.g., Hatfield, & Epstein, 1985; Leeuwenberg & Boselie, 1988; Pomerantz and Kubovy, 1986; Rock, 1983). The controversy has been difficult to settle because neither of the key principles, likelihood and simplicity, is clearly defined. Moreover, there have been suspicions that the two principles are not in fact separate, but are two sides of the same coin. Pomerantz and Kubovy (1986) cited Mach (1906/1959)—"The visual sense acts therefore in conformity with the principle of economy [i.e., simplicity], and at the same time, in conformity with the principle of probability [i.e., likelihood]" (p. 215)—and themselves have suggested that some resolution between the two approaches might be possible. Moreover, the close mathematical relationship between simplicity and likelihood has been widely acknowledged in a range of technical literatures, in computational modeling of perception (e.g., Mumford, 1992), artificial intelligence (e.g., Cheeseman, 1995), and statistics (e.g., Wallace & Freeman, 1987). But this relationship has not been used to demonstrate the equivalence between simplicity and likelihood principles in perceptual organization.

This article shows that the likelihood and simplicity principles of perceptual organization can indeed be rigorously unified by using results linking simplicity and probability theory developed within the mathematical theory of Kolmogorov complexity (Chaitin, 1966; Kolmogorov, 1965; Li & Vitanyi, 1993; Solomonoff, 1964).

Likelihood Versus Simplicity: The Debate

Both the likelihood and simplicity principles explain, at least at an intuitive level, a wide range of phenomena of perceptual organization. Consider, for example, the Gestalt law of good continuation, that perceptual interpretations that involve continuous lines or contours are favored. The likelihood explanation is based on the observation that continuous lines and contours are very frequent in the environment (e.g., Brunswick, 1956). Although it is possible that the input was generated by discontinuous lines or contours that happen, by coincidence, to be arranged so that they are in alignment from the perspective of the viewer, this possibility is rejected because it is less likely. The simplicity explanation, by contrast, suggests that continuous lines or contours are imposed on the stimulus when they allow that stimulus to be described more simply.

Another example is the tendency to perceptually interpret ambiguous two-dimensional projections as generated by three-dimensional shapes containing only right angles (Attneave, 1972; Perkins, 1972, 1982; Shepard, 1981). The likelihood explanation is that right-angled structures are more frequent in the environment (at least in the "carpentered" environment of the typical experiments subject; Segall, Campbell & Herskovits, 1966). The simplicity explanation is that right-angled struc-

I would like to thank Mike Oaksford and Julian Smith for their valuable comments on the manuscript.

Correspondence concerning this article should be addressed to Nick Chater, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England. Electronic mail may be sent via Internet to chater@psy.ox.ac.uk.

[1] This principle is also known as *prägnanz* or the *minimum principle.* The term minimum principle comes from the minimization of complexity; the term *prägnanz* was used with somewhat broader scope by the Gestalt school, to include regularity, symmetry, and other properties (Koffka, 1935/1963).

tures are simpler—for example, they have fewer degrees of freedom—than trapezoidal structures.

There is a vast range of phenomena that appear consistent with both likelihood and simplicity interpretations. From the standpoint of this article—that the likelihood and simplicity principles are equivalent—the common coverage of the two approaches is not surprising. More interesting, however, are phenomena that have been taken as evidence for one view or the other. On the present interpretation, such evidence cannot, of course, be taken at face value, as we shall see below. For now, let us consider a typical example of evidence adduced on each side of the debate.

Putative Evidence for Likelihood

Likelihood is widely assumed to be favored by evidence that shows that preferred perceptual organization is influenced by factors concerning the structure of the everyday environment. For example, consider two-dimensional projections of a shaded pattern, which can be seen either as a bump or an indentation (see, e.g., Rock, 1975). The preferred interpretation is consistent with a light source from above, as in natural lighting conditions. Thus, the perceptual system appears to choose the interpretation that is most likely, but there is no intuitive difference between the simplicity of the two interpretations.

Putative Evidence for Simplicity

Cases of perceptual organizations that violate, rather than conform to, environmental constraints are widely assumed to favor the simplicity account. Leeuwenberg and Boselie (1988) show a schematic drawing of a symmetrical, two-headed horse. The more likely interpretation, also consistent with the drawing, is that there are two horses, one occluding the other. But the perceptual system appears to reject likelihood; it favors the interpretation that there is a single, rather bizarre, animal.

These considerations suggest that likelihood and simplicity cannot be the same principles; the two appear to give rise to different predictions. I now discuss likelihood and simplicity in turn and argue that, despite appearances, they are identical.

Likelihood

The likelihood principle proposes that the perceptual system chooses the organization that corresponds to the most likely distal layout consistent with the sensory input. But what does it mean to say that a hypothesized distal layout, H_i, is or is not likely, given sensory data, D? The obvious interpretation is that this likelihood corresponds to the conditional probability of the distal layout, given the sensory input: $P(H_i \mid D)$. But this step does not take us very far, because there are a variety of ways in which probabilities can be interpreted, and it is not clear which interpretation is appropriate in the present case.

A first suggestion is that the probability can be interpreted in terms of frequencies. This frequentist interpretation of probability (von Mises, 1939/1981) is that the probability of an outcome is the limit of the frequency of the outcome divided by the total number of "trials" in a repeated experiment. For example, suppose that the repeated experiment is tossing a fair coin. The limiting frequency of the outcome, "heads," divided by the total number of trials will tend toward 0.5 as the number of trials increases. According to the frequentist interpretation of probability, this is what it means to say that the probability of the coin falling heads, on any given trial, is 0.5. The frequentist interpretation of the conditional probability $P(A \mid B)$ is simply the frequency of trials on which B and A occur, divided by the total frequency of trials on which B occurs. Although the frequentist interpretation of probability is used in many philosophical and mathematical contexts, it does not give a meaningful interpretation of probability in the present context. According to the frequentist account, the conditional probability of the distal layout, given the sensory input, $P(H_i \mid D)$, is defined as the limit of the frequency of trials on which both the sensory input, D, and the distal layout, H_i, occur, divided by the total frequency of trials on which D occurs. But this limit is not well defined, because the same sensory input will never occur again—and hence a limiting frequency can never be obtained. Of course, similar inputs will occur, but this does not help—because classifying inputs as similar requires that some organization is imposed upon them, and it is organization that we want to explain. In short, then, the frequentist interpretation of probability is defined in terms of limiting frequencies in a repeated experiment; and this is inapplicable to probabilities involving sensory inputs, because sensory inputs are never repeated.

Therefore a different notion of probability is required. The natural alternative is a subjective conception of probability.[2] According to a subjectivist conception (de Finetti, 1972; Keynes, 1921), probability is a measure of degree of belief in some event or state. Conditional probability corresponds to degree of belief in an event or state, given some other belief in an event or state. How can these ideas be translated into the context of perceptual organization? The likelihood, $P(H_i \mid D)$, is interpreted as the degree of belief in the hypothesis (H_i) concerning the distal layout, given data (D) concerning the sensory state.[3]

In order to calculate likelihood, Bayes's theorem must be applied:

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)}, \qquad (1)$$

where

$$P(D) = \sum_j P(D \mid H_j)P(H_j),$$

Bayes's theorem allows the likelihood to be calculated from two kinds of quantity: (a) terms of the form: $P(H_j)$—"prior" probabilities for each distal layout; and (b) terms of the form

[2] There is a further possibility, an objectivist or propensity approach to probability (e.g., Mellor, 1971), but it is not clear how this approach might be applicable in this case.

[3] I use the term *likelihood* in the sense used by perceptual theorists. In statistics, it is frequently used to denote $P(D \mid H_i)$, the probability of data given a hypothesis. Classical sampling theory approaches to statistical inference often involve maximizing likelihood (Fisher, 1922), in the statisticians sense, rather than posterior probability, which is the appropriate sense of maximum likelihood for perceptual theory.

$P(D|H_j)$—conditional probabilities for each sensory input, given the distal layout.

The likelihood approach assumes that the perceptual system chooses the most likely H_j, (i.e., the H_j that maximizes Equation 1). To state this formally, and for use in the calculations below, I introduce some notation:

$$\arg\max_x [f(x)] = i \underset{definition}{\Leftrightarrow} f(i) \text{ maximizes } f(x) \quad (2)$$

and

$$\arg\min_x [f(x)] = i \underset{definition}{\Leftrightarrow} f(i) \text{ minimizes } f(x). \quad (3)$$

Using this notation, the likelihood principle states that the chosen hypothesis is the H_k such that

$$k = \arg\max_i [P(H_i | D)]. \quad (4)$$

Simplicity

Applying a simplicity criterion to perceptual organization requires clarification on two points: what is assessed for simplicity? and how is simplicity measured? I address these in turn.

What Is Assessed for Simplicity?

The first question invites an easy answer: that the perceptual organization is made as simple as possible. But, taken at face value, this means that a very simple organization (perhaps perceiving the distal scene to be a uniform, unstructured field) would always be a preferred organization. This possibility is, of course, ruled out by the constraint that the organization is consistent with the sensory input—and most sensory inputs will not be consistent with this organization, because they are highly nonuniform. But this point itself raises difficult questions: What does it mean for an organization to be consistent, or compatible, with a perceptual input? Can consistency with the input be traded against simplicity of interpretation?[4] If so, how are simplicity and consistency with the input to be jointly optimized? The theoretical account of simplicity presented below suggests how these questions may be answered.

There is, however, a further, and more subtle difficulty: What rules out the simplest possible, "null," perceptual organization? This organization is completely consistent with the sensory input, since it adds nothing to it. Mere consistency or compatibility with the sensory input is plainly not enough; the perceptual organization must also, in some sense, capture regularities in the sensory input. This naive use of simplicity of perceptual organization as a guiding principle is analogous to a naive use of simplicity as a guiding principle in science: Preferring the simplest theory compatible with the data would lead to null theories of great simplicity, such as "anything whatsoever can happen." Such null theories, although simple, are unsatisfactory because they do not explain any of the regularities in the natural world. In science, simplicity of theory must be traded against explanatory power (Harman, 1965); the same point applies for perceptual organization. But this appears to imply that perceptual organization involves the joint optimization of two factors, and the relative influence of these two factors is unspecified. Moreover, this conclusion is unattractive because two notions, simplicity and explanatory power, must be explicated rather than just one.

Fortunately, there is an alternative way to proceed. This is to view perceptual organization as a means of encoding the sensory stimulus; and to propose the perceptual organization chosen is that which allows the simplest encoding of the stimulus. This view disallows simple perceptual organizations that bear little or no relation to the stimulus, because these organizations do not help encode the stimulus simply. It also provides an operational definition of the *explanatory power* of a perceptual organization—as the degree to which that organization helps provide a simple encoding of the stimulus. If a perceptual organization captures the regularities in the stimulus (i.e., if it "explains" those regularities), then it will provide the basis for a brief description of the stimulus; if an organization fails to capture regularities in the stimulus, then it will be of no value in providing a brief description of the stimulus. Explanatory power is therefore not an additional constraint that must be traded off against simplicity; maximizing explanatory power is the same as maximizing the simplicity of the encoding of the stimulus.

How Can Simplicity Be Measured?

I have established what simplicity should apply to: namely, the encoding of the perceptual stimulus. But how is simplicity to be measured? The measurement of simplicity has been considered extensively in philosophy (e.g., Sober, 1975) where no quantitative measure has gained acceptance. In psychology, theorists concerned with perceptual organization have taken the pragmatic step of identifying the simplicity of an encoding with its length. Attneave (1954, 1981), for example, explicitly suggested that "what the [perceptual] system *likes* is short descriptions" (Attneave, 1981, p. 417). According to this point of view, the preferred perceptual organization is that which allows the briefest possible perceptual encoding. It is interesting that the suggestion that the perceptual system has economical encoding as an important goal has also been suggested in a variety of other contexts (e.g., Atick & Redlich, 1990; Barlow, Kaushal, & Mitchison, 1989; Blakemore, 1990). Moreover, as I discuss below, this is an appropriate choice—an independent tradition within mathematics and computer science, Kolmogorov complexity theory, shows that the identification of simplicity with brevity provides a deep and important theory of simplicity (Chaitin, 1966; Kolmogorov, 1965; Li & Vitanyi, 1993; Solomonoff, 1964).

Psychologists have used two approaches to operationalize the notion of brevity of encoding: Shannon's (1948) information theory (Attneave, 1959; Garner, 1962) and the tradition known as coding theory (Hochberg & McAlister, 1953; Restle, 1979; Simon, 1972), one elaboration of which is structural information theory (Buffart, Leeuwenberg & Restle, 1981). I consider these in turn.

---

[4] Koffka (1935/1963) allowed the possibility of organizations that are not consistent with the perceptual stimulus, allowing distortion to be traded with simplicity. The empirical evidence indicates that such trade-offs, if they occur, may be quite small (Attneave, 1982).

Information theory and brevity. The information-theoretic approach quantifies brevity in terms of the number of bits required to distinguish the stimulus (or some part of the stimulus) from a range of mutually exclusive and exhaustive alternatives, known as an *information source*. Each alternative, A_i, in an information source, A, is associated with some probability of occurrence, $P(A_i)$. The amount of information, $I(A_i)$, associated with the choice of a particular alternative, A_i, is called the *surprisal* of A_i and is defined

$$I(A_i) = \log_2\left(\frac{1}{P(A_i)}\right). \tag{5}$$

Later, I shall consider the surprisal of A_i, conditional on some other event, B_j. I denote this by $I(A_i \mid B_j)$, and the definition parallels Equation 5:

$$I(A_i \mid B_j) = \log_2\left(\frac{1}{P(A_i \mid B_j)}\right). \tag{6}$$

The average surprisal of a source, A, is known as its entropy, $H(A)$, and is simply the surprisal of each alternative, weighted by its probability of occurrence:

$$H(A) = \sum_j P(A_j)I(A_j). \tag{7}$$

Surprisal can be viewed as a measure of brevity of encoding because of basic ideas from information theory, which I now discuss. Suppose that a sequence of alternatives is independently chosen according to the probabilities of the information source, and that this sequence of alternatives must be encoded in a binary sequence. Let us stipulate that the encoding must be noiseless (i.e., the sequence of alternatives can be reconstructed with perfect accuracy). Suppose, moreover, that the encoding proceeds by associating each alternative, A_i, with a "code word"; that is, a sequence of binary digits (so, e.g., a particular alternative, A_{13}, might be associated with the code word 001101). A particular sequence of alternatives is then encoded by concatenating the corresponding code words into a single binary sequence.

How should the code words be assigned to alternatives in order to minimize the average length of the binary string required to transmit the sequence of alternatives? The length of the string encoding the sequence is the product of the length of the sequence and the average length of code words for elements in the sequence; hence we must assign code words in order to minimize the average code word length. Suppose that the code word for alternative A_i is a binary string of length, l_i. Then the average code word length for the source A is specified:

$$\sum_j P(A_j)l_j. \tag{8}$$

Let us call the minimum value of this average $L(A)$. Shannon's (1948) *noiseless coding theorem* is that this minimum is very close to the entropy of the source:

$$H(A) \leq L(A) \leq H(A) + 1. \tag{9}$$

Crucially, this minimum is obtained when the code length for an alternative is its surprisal (rounded up to the nearest integer, because binary sequences can only have integer lengths). In symbols:

$$l_j = \lceil I(A_j) \rceil = \left\lceil \log_2\left[\frac{1}{P(A_j)}\right] \right\rceil \tag{10}$$

where the notation $\lceil x \rceil$ denotes x rounded up to the nearest integer. This means that the surprisal of an alternative can be viewed as a measure of its code length in an optimal binary code. Thus, surprisal can be viewed as a measure of brevity of encoding.

Despite the theoretical elegance of information theory in many contexts, it proves to have a number of difficulties when applied to individual perceptual stimuli, as we now see. Suppose, to borrow an example from Leeuwenberg and Boselie (1988), that the stimulus consists of a sequence of letters: *aaabbbbbgg*. The amount of information in this stimulus depends on the information source being considered. Suppose that it is assumed that each letter is chosen independently and that there is an equal ($\frac{1}{3}$) chance that the letter chosen will be *a*, *b*, or *g*. Then the information associated with the specification of, say, the initial *a* is $\log_2(1/\frac{1}{3}) = \log_2(3)$ bits of information. The information associated with the specification of each of the 10 letters in this sequence is, in this case, the same. Hence the information required to specify the entire sequence is 10 $\log_2(3)$ bits. But a different result is obtained if we suppose that the letters have different probabilities of being chosen. Suppose that we assume that *b* is chosen with probability $\frac{1}{2}$, and *a* and *g* with probabilities $\frac{1}{4}$. Now, five of the letters (the *b*s) can each be specified with $\log_2(1/\frac{1}{2}) = \log_2(2) = 1$ bit; and five (the *a*s and *g*s) can each be specified with $\log_2(1/\frac{1}{4}) = \log_2(4) = 2$ bits, making a total of 15 bits of information. Furthermore, a different result again is obtained if it is assumed that the letters are chosen from the entire alphabet, or the entire computer keyboard, or all possible shapes of a certain size, and so on. The larger the set from which the stimulus is presumed to have been chosen, the more information is required to specify it. Moreover, we might assume that the individual elements of the stimulus are not chosen independently (as the sequential runs in our sample sequence would tend to suggest). Perhaps the stimulus was generated by a more complex process, such as a Markov source, or a stochastic context-free grammar. Each distinct specification of the source from which the stimulus was generated gives rise to a different answer concerning the amount of information in the stimulus.

In short, the information-theoretic approach does not measure the information in a particular stimulus per se, but rather measures the amount of information in that stimulus relative to the probabilities of all the other stimuli that might have been generated. In experimental settings, where a small number of stimuli are presented many times, this range of possibilities has a relatively natural definition, in terms of the probabilities of each stimulus presented in the experiment (e.g., Garner, 1962). But in many experimental contexts, and in natural perception, the range of possibilities from which the current stimulus is drawn can be defined in innumerably many different ways, as is clear even with the simple letter sequence described above. Hence, in most contexts, information theory does not provide a useful measure of the brevity with which a particular stimulus

can be encoded because it is not defined relative to the stimulus alone (see Garner, 1962, for related discussion).[5]

Even putting this difficulty aside, information theory merely specifies the length of the code (the number of bits) required to encode the stimulus, but it does not pick out any particular code as the best code for the stimulus. Yet it is the nature, not just the length, of the code that is crucial from the point of view of understanding perceptual organization (Garner, 1974). Suppose, for example, that the sequence above was drawn from the eight equally likely alternatives shown in Table 1. These sequences can be viewed as consisting of three "segments" of repeated letters—the first segment being three letters long, the second segment five letters long, and the final segment two letters long. The eight alternatives can be viewed as generated by three binary decisions, concerning whether the first segment consists of *a*s or *x*s, whether the second segment consists of *b*s or *y*s, and whether the third segment consists of *g*s or *z*s. By information theory, the number of bits required to specify a choice between eight equally probable alternatives is three bits. But optimal three-bit codes need not relate to the organization present in the perceptual stimuli. Table 1 illustrates two optimal codes for the stimuli. The first is "meaningful" in the sense that each bit carries information about the identity of the letters in a particular segment; it therefore reflects this (minimal) organization of the stimulus into three segments. By contrast, the second code is "meaningless"; codes are arbitrarily assigned to strings, and hence the organization of the stimulus is ignored. From an information-theoretic point of view there is no distinction between these codes; information theory does not favor the code that reflects the underlying organization of the stimulus over that which does not. But from the point of view of perceptual organization, the difference between codes that express the structure of the stimuli and those that do not would appear to be crucial.

Table 1
"Meaningful" and "Meaningless" Codes for a Simple Sequence

Stimulus	"Meaningful" code	"Meaningless" code
aaabbbbbgg	111	101
aaabbbbbzz	110	011
aaayyyyygg	101	001
aaayyyyyzz	100	111
xxxbbbbbgg	011	000
xxxbbbbbzz	010	100
xxxyyyyygg	001	110
xxxyyyyyzz	000	010

Note. Table 1 shows a set of eight stimuli that are assumed to be equally probable. By information theory, any given stimulus can be encoded in just three bits. Two possible optimal codes are shown. The first is "meaningful," in the sense that each element of the code can be interpreted as specifying part of the structure of the stimulus. Specifically, the first element of the code specifies whether the first three letters are *a*s or *x*s, the second element specifies whether the next five letters are *b*s or *y*s, and the third element specifies whether the final two letters are *g*s or *z*s. The second code is "meaningless," in that the relation between stimuli and codes is chosen at random. Information theory does not distinguish between the code that "organizes" the stimuli and the code that does not.

Table 2
A "Meaningless" Code for Heterogeneous Stimuli

Stimulus	Code
aaabbbbbgg	111
24£1fhq+3*	110
(Hh8cQ1a1	101
wwwww	100
%	011
T	010
PERCEPTION	001
42^{42}	000

Note. Table 2 reinforces the point made by Table 1, again showing a set of eight stimuli, assumed to be equally probable. Here the stimuli are completely heterogeneous. The sequence *aaabbbbbgg* has the same code as in Table 1, but it is not "meaningful." Information theory does not recognize this distinction: In both Tables 1 and 2, the code used for the sequence is optimal.

This point is reinforced by the example shown in Table 2, of eight further stimuli assumed to be equally likely. Unlike the stimuli in Table 1, these stimuli cannot be organized in any coherent way, but are completely heterogeneous. Nonetheless, the same code as before (111) is used to specify *aaabbbbbgg* as in the meaningful code in Table 1. But while the stimulus and the code are the same as before, now there is no meaningful interpretation of the individual parts of the code as specifying the structure of the sequence. From the point of view of perceptual organization this is a crucial difference, the difference between stimuli that can be organized and those that cannot; but it is ignored by information theory.

Coding theory and brevity. The difficulties with the application of information theory have led psychologists to develop an alternative approach to measuring the brevity with which a perceptual organization allows a stimulus to be encoded. This method is to define what Simon (1972) calls *pattern languages* in which different organizations of the stimulus can be expressed. The preferred organization is that which allows the shortest description of the stimulus, when measured in terms of the length of the expression in the pattern language. This constraint is stated in terms of number of symbols in the description (e.g., Simon, 1972) and sometimes the number of parameters (e.g., Leeuwenberg, 1969). These formulations are equivalent according to the natural stipulation that the values of each parameter are coded by a distinct symbol.

The coding theory approach may be illustrated with the *aaabbbbbgg* sequence described above. The coding corresponding to the null organization requires 10 parameters ($I = 10$), one for each symbol of the sequence. A spurious organization, dividing the sequence $(aa)(ab)(bb)(bb)(gg)$ can support the code $2(a)ab2(b)2(b)2(g)$; where $2(a)$ is a code in the pattern

[5] Various ingenious notions such as *inferred subsets* (Garner, 1974) and *perceived subsets* (Pomerantz, 1981; Royer & Garner, 1966) have been used to specify a relevant set of alternatives from which the stimulus is chosen, although these apply only in very specific contexts.

language meaning a run of two *a*s. But this expression contains 10 parameters, either numbers or letters, and hence no economy is achieved ($I = 10$, as before).[6] A good organization, (*aaa*) (*bbbbb*) (*gg*), can support the code $3(a)5(b)2(g)$, which requires just six parameters ($I = 6$). This intuitively sensible organization thus corresponds to the organization that supports a short code.

This general approach has been applied in a variety of contexts, from the organization of simple sequences, such as the example just considered (Leeuwenberg, 1969; Restle, 1970; Simon, 1972; Simon & Kotovsky, 1963; Vitz & Todd, 1969), to judgments of "figural goodness" (Hochberg & McAllister, 1953), the analysis of Johansson's (1950) experiments on the perception of motion configurations (Restle, 1979), and figural completion (Buffart, Leeuwenberg & Restle, 1981). It has also been advanced as a general framework for understanding perceptual organization (e.g., Attneave & Frost, 1969; Leeuwenberg, 1971; Leeuwenberg & Boselie, 1988).

Approaches based on short description length appear to be dogged by two problems: (a) that a fresh description language must be constructed for each fresh kind of perceptual stimulus and (b) that the predictions of the theory depend on the description language chosen and there is no (direct) empirical means of deciding between putative languages. In practice, as Simon (1972) noted, the second problem is not terribly severe; description lengths in different proposed description languages tend to be highly correlated. The mathematical theory of Kolmogorov complexity provides a useful generalization of coding theory that addresses these issues.

Reconciling Likelihood and Simplicity: I

I now show that the likelihood and simplicity principles can be viewed as different sides of the same coin. I present the discussion in two stages. This section describes how to effect the reconciliation, if simplicity is measured using standard information theory. This connection between simplicity and likelihood has previously been discussed in the context of computational models of visual perception by Mumford (1992; see also Grenander, 1976–1981) and is well-known in the literature on information theory (e.g., Cover & Thomas, 1991) and statistics (e.g., Rissanen, 1989). The next, much longer, section, generalizes this analysis, using Kolmogorov complexity instead of standard information theory, to provide a deeper reconciliation of simplicity and likelihood. I now turn, then, to the analysis in terms of standard information theory.

Let us begin with the likelihood view. I noted above that the likelihood view recommends the most probable hypothesis, H_k, about perceptual organization, given the sensory data, D. In symbols (Equation 4) this is

$$k = \arg\max_i [P(H_i|D)].$$

Applying Bayes's theorem (Equation 1), this implies that

$$k = \arg\max_i \left[\frac{P(D|H_i)P(H_i)}{P(D)}\right].$$

The k that maximizes this quantity will also maximize the logarithm of that quantity (because log is monotonically increasing). This implies that

$$k = \arg\max_i \left\{\log_2\left[\frac{P(D|H_i)P(H_i)}{P(D)}\right]\right\}$$
$$= \arg\max_i \{\log_2[P(D|H_i)] + \log_2[P(H_i)] - \log_2[P(D)]\}.$$

Because $P(D)$ is independent of the choice of i, the final term is irrelevant, such that:

$$k = \arg\max_i \{\log_2[P(D|H_i)] + \log_2[P(H_i)]\}$$
$$= \arg\min_i \{-\log_2[P(D|H_i)] - \log_2[P(H_i)]\}$$
$$= \arg\min_i \left\{\log_2\left[\frac{1}{P(D|H_i)}\right] + \log_2\left[\frac{1}{P(H_i)}\right]\right\}.$$

This gives the result:

$$k = \arg\min_i [I(D|H_i) + I(H_i)], \quad (11)$$

where this last step follows from the definitions of surprisal, I (Equations 5 and 6).

Having considered the likelihood principle, I now consider simplicity, using information theory. The length of code required to encode data D (the stimulus) needs to be determined, using a particular hypothesis, H_i, concerning perceptual organization. This code will have two parts: first, an encoding of the choice of hypothesis, H_i, and second an encoding of the data D, given H_i. The total code length, l_{total_i}, will be the sum of the length of the code for the hypothesis, l_{H_i}, and the length of the code for the data given the hypothesis, $l_{D|H_i}$. In symbols:

$$l_{total_i} = l_{H_i} + l_{D|H_i}. \quad (12)$$

The simplicity strategy is to choose the k' that minimizes the total code length, l_{total_i}, or, in symbols:

$$k' = \arg\min_i (l_{total_i}) = \arg\min_i (l_{H_i} + l_{D|H_i}).$$

Using Equation 10 to convert code lengths into surprisals (assuming that we use optimal codes, at each step), this is

$$k' = \min_i [\lceil I(D|H_i)\rceil + \lceil I(H_i)\rceil]. \quad (13)$$

Aside from the rounding terms (which can in any case be eliminated by a more sophisticated treatment), the goal of choosing a hypothesis to minimize description length (13) is the same as the goal of choosing a hypothesis to maximize likelihood (11). That is, the most likely hypothesis, H_i, about perceptual organization is the H_i that supports the shortest description of the data, D, concerning the stimulus. Thus, the likelihood principle

[6] This code also contains 10 symbols (ignoring brackets, which are present only for clarity). In this code, each parameter value corresponds to a single symbol (although this would cease to be true for numerical values larger than 9, which are expressed as compound symbols in base 10).

(choose the organization that is most likely) and the simplicity principle (choose the organization that allows the briefest encoding of the stimulus) are equivalent.

It is important to stress that this result shows not just that likelihood and description length are sometimes identical. Rather, it shows that for any problem of maximizing likelihood there is a corresponding "dual" problem of minimizing description lengths. Specifically, given any specification of subjective probabilities $P(H_k)$ and $P(D|H_k)$, there will be a code that is optimal with respect to those probabilities. The hypothesis that minimizes the length of code required for the data, D, will be the same as the hypothesis that maximizes likelihood. Similarly, any code can be viewed as an optimal code with respect to a set of subjective probabilities. Therefore, choosing the hypothesis that minimizes the code length of the data will be equivalent to maximizing likelihood with respect to those probabilities. This equivalence has not been noted in the psychology of perceptual organization, but in the study of computational models of perception, it has been widely exploited. For example, problems of maximizing likelihood can sometimes be made more tractable when prior knowledge of relevant probabilities is not available, by switching to a formulation of minimizing code length (Mumford, 1992). Conversely, problems of minimizing code length can sometimes be solved by switching to a probabilistic formulation, where relevant prior knowledge is available (e.g., Atick & Redlich, 1992).

It follows, therefore, that any evidence that the perceptual system chooses perceptual organizations that follow from the likelihood principle (where certain assumptions are made about subjective probabilities) can be viewed equally well as evidence for the simplicity principle (where certain and equivalent assumptions are made about the coding language).

Reconciling Simplicity and Likelihood: II

The analysis above is suggestive, but unsatisfactory in two ways. First, regarding simplicity, I have used the information-theoretic measure of simplicity, which I have already noted is not an appropriate measure of the brevity with which an individual stimulus can be encoded. Second, regarding likelihood, I have sidestepped a fundamental difficulty: that there are infinitely many possible distal states (i.e., infinitely many H_i) consistent with any given perceptual stimulus, D, and it is by no means obvious that these can consistently be assigned prior probabilities, $P(H_i)$. I now show how both of these problems have been addressed by the mathematical theory of Kolmogorov complexity. Specifically, I show how a measure of the information in an individual stimulus can be defined, which can be viewed as a generalization and unification of the information theory and coding theory approaches developed in psychology; and how a coherent definition of prior probability for a distal state can be specified. I then show that the equivalence between the likelihood and simplicity criteria holds, using these more satisfactory notions of likelihood and simplicity.

The literature on Kolmogorov complexity is not well-known in psychology, and I therefore sketch relevant aspects of the theory below, referring the reader to Li and Vitanyi's (1993) excellent textbook for more details.

Kolmogorov Complexity as a Measure of Simplicity

Coding theory suggests a means of deciding between rival perceptual organizations of the stimulus. This involves defining a pattern language, expressing the rival perceptual organizations in that pattern language, and favoring the organization that allows the briefest description of the stimulus. By extension, coding theory suggests a way of measuring the complexity of stimuli themselves: the length of the shortest description in the pattern language that encodes the stimulus. Indeed, this notion has been used to account for subjects' judgments of the complexity of stimuli by a number of researchers (Leeuwenberg, 1969; Simon, 1972; Simon & Kotovsky, 1963; Vitz & Todd, 1969). The theory of Kolmogorov complexity (Chaitin, 1966; Kolmogorov, 1965; Solomonoff, 1964), developed independently, can be viewed as a more general version of this approach to measuring complexity (and it can also be viewed as a generalization of standard information theory).

I noted above two apparently unattractive features of coding theory: (a) that different pattern languages must be developed for different kinds of stimuli and (b) that the measure of simplicity depends on the pattern language used. Kolmogorov complexity avoids the first problem by choosing a much more general language for encoding. Specifically, the language chosen is a *universal programming language.* A universal programming language is a general purpose language for programming a computer. The familiar programming languages such as PROLOG, LISP, and PASCAL are all universal programming languages. How can an object, such as a perceptual stimulus, be encoded in a universal programming language such as, for example, LISP? The idea is that a program in LISP encodes an object if the object is generated as the output or final result of running the program.

Whereas coding theory requires the development of special purpose languages for coding particular kinds of perceptual stimulus, Kolmogorov complexity theory can describe all perceptual stimuli using a single universal programming language. This follows because, by the definition of a universal programming language, if an object has a description from which it can be reconstructed in any language, then it will have a description from which it can be reconstructed in the universal programming language. It is this that makes the programming language universal. The existence of universal programming languages is a remarkable and central result of computability theory (Odifreddi, 1989; Rogers, 1967). Perhaps even more remarkable is that there are so many universal programming languages, including all the familiar computer languages.[7]

One of the simplest universal languages, which we will consider below, is that used to encode programs on a standard universal Turing machine (Minsky, 1967). A Turing machine is a simple computational device, with two components. The first component is a linear "tape" consisting of squares that may contain the symbols *0* or *1* or may be left blank. The tape can be extended indefinitely in both directions, and hence there can be infinitely many different patterns of 0s and 1s on the tape. The second component is a "control box," consisting of a finite

[7] Specifically, any language rich enough to express the partial recursive functions will be universal (Rogers, 1967).

number of states, which operates upon the tape. At any time, the control box is located over a particular square of the Turing machine's tape. The control box has a small number of possible actions: It can move left or right along the tape, one square at a time; it can read the symbol on the tape over which it is currently located; it can replace the current symbol with a different symbol; and the current state of the control box may be replaced by one of the finite number of other possible states. Which actions the control box performs is determined by two factors: the current state of the machine and the symbol on the square of the tape over which it is located.

A Turing machine can be viewed as a computer in the following way. The input to the computation is encoded as the string of 1s and 0s that comprise the initial state of the tape. The nature of the control box (that is, which symbols and states lead to which actions and changes of state) determines how this input is modified by the operation of the control box. The control box might, for example, leave the input intact, delete it entirely and replace it with a completely different string of 1s and 0s, or more interestingly perform some useful manipulation of the input. This resulting string encodes the output of the computation.[8] Each control box can therefore be associated with a mapping from inputs to outputs, defining the computation that it performs. According to the Church–Turing thesis (Boolos & Jeffrey, 1980), every mapping that can be computed by any physical device whatever can be computed by some Turing machine.

A universal Turing machine is a Turing machine that is programmable. There are many different possible programming languages corresponding to different universal Turing machines (in the same way that there are many different programming languages for conventional computers). In the following, one universal Turing machine (it does not matter which) has been chosen, which is called U. The input to U can be thought of as divided into two parts (these will typically be separated, for example, by blank spaces between them on the tape). The first part is the *program*, which encodes the series of instructions to be followed. The second part is the *data*, upon which the instructions in the program are to operate. U's control box is designed so that it reads and carries out the instructions in the program as applied to the data provided. In this way, U can perform not just a single computation from input to output, but many different mappings, depending on what is specified by its program. In fact, it is possible to write a program that will make U compute the same mapping as any specified Turing machine, and thus every mapping that can be computed at all (assuming the Church–Turing thesis). So U is universal in that it can perform all possible computations, if it is given the appropriate program. Notice, finally, that the program of the universal Turing machine is just a string of 1s and 0s on the input tape—this is the same form as the data on which the program operates. We shall see that this is useful when considering the notion of universal *a priori* probability below.

For any object,[9] including perceptual stimuli, the definition of the complexity of that object is the length of the shortest code (i.e., the shortest program) that generates that object, in the universal programming language of choice. By using a universal language, the need to invent special purpose languages for each kind of perceptual stimulus is avoided, thus solving the first problem noted with coding theory.

Moreover, in solving the first problem, the second problem, that different patterns languages give different code lengths, is solved automatically. A central result of Kolmogorov complexity theory, the *invariance theorem* (Li & Vitanyi, 1993), states that the shortest description of any object is invariant (up to a constant) between different universal languages. Therefore, it does not matter whether the universal language chosen is PROLOG, LISP or PASCAL, or binary strings on the tape of a universal Turing machine; the length of the shortest description for each object will be approximately the same. Let us introduce the notation $K_{LISP}(x)$ to denote the length of the shortest LISP program that generates object x; and $K_{PASCAL}(x)$ to denote the length of the shortest PASCAL program. The invariance theorem implies that $K_{LISP}(x)$ and $K_{PASCAL}(x)$ will only differ by some constant, c (which may be positive or negative), for all objects, including, of course, all possible perceptual stimuli. In symbols:

$$\exists c \forall x (K_{LISP}(x) = K_{PASCAL}(x) + c). \qquad (14)$$

(where $\exists$ denotes existential quantification and $\forall$ denotes universal quantification). In specifying the complexity of an object, it is therefore possible to abstract away from the particular language under consideration. Thus the complexity of an object, x, can be denoted simply as $K(x)$; this is known as the *Kolmogorov complexity* of that object.

Why is complexity language invariant? To see this intuitively, note that any universal language can be used to encode any other universal programming language. This follows from the preceding discussion because a programming language is just a particular kind of computable mapping, and universal programming language can encode any computable mapping. For example, starting with LISP, a program can be written, known in computer science as a compiler, that translates any program written in PASCAL into LISP. Suppose that this program has length c_l. Now suppose that $K_{PASCAL}(x)$, the length of the shortest program that generates an object x in PASCAL, is known. What is $K_{LISP}(x)$, the shortest program in LISP that encodes x? Notice that one way of encoding x in LISP works as follows: The first part of the program translates from PASCAL into LISP (of length c_l), and the second part of the program, which is an input to the first, is simply the shortest PASCAL program generating the object. The length of this program is the sum of the lengths of its two components: $K_{PASCAL}(x) + c_l$. This is a LISP program

[8] It is also possible that the control box will continue modifying the contents of the tape forever, so that there is no well-defined output. We shall ignore such nonhalting Turing machines later for simplicity. See Footnote 12.

[9] There is an implicit restriction, of course, to abstract, mathematical objects, both in this general context and in the context of perception. The perceptual stimulus is considered in terms of some level of description (e.g., in terms of pixel values, or activity of receptors); it is the abstract description that is encoded in the pattern language. It would be incoherent to speak of the perceptual stimulus itself being encoded. Encoding concerns information, and information is by definition an abstract quantity, dependent on the level of description (see Chater, 1989; Dretske, 1981, for discussion).

that generates x, if by a rather roundabout means. Therefore $K_{LISP}(x)$, the shortest possible LISP program must be no longer than this: $K_{LISP}(x) \leq K_{PASCAL}(x) + c_1$. An exactly similar argument based on translating in the opposite direction establishes that $K_{PASCAL}(x) \leq K_{LISP}(x) + c_2$. Putting these results together, $K_{PASCAL}(x)$ and $K_{LISP}(x)$ are the same up to a constant, for all possible objects x. This is the Invariance Theorem (see Li & Vitanyi, 1993, for a rigorous proof along these lines) and establishes that Kolmogorov complexity is language invariant.

Both limitations of coding theory—the need to develop special purpose languages for particular kinds of pattern and the dependence of code length on the pattern language used—are overcome. Only a single language need be used, a universal programming language, and moreover it does not matter which universal programming language is chosen, because code lengths are invariant across universal languages.

In addition to providing a measure of the complexity of a single object, x, Kolmogorov complexity can be generalized to measure the complexity of transforming one object, y, into another object, x. This quantity is the length of the shortest program that takes y as input and produces x as output and is called *conditional Kolmogorov complexity*, which is written $K(x|y)$. $K(x|y)$ will sometimes be much less than $K(x)$. Suppose, for example, that x and y are both random strings of n binary numbers (where n is very large). Because random strings have no structure, they cannot be compressed by any program, and hence $K(x) = K(y) = n$ (Li & Vitanyi, 1993). But suppose that x and y are closely related (e.g., that x is simply the same string as y but in reverse order). Then the shortest program transforming y into x will be the few instructions needed to reverse a string, of length c_3, where c_3 is very much smaller than n. Thus, in this case, $K(x|y)$ will be much smaller than $K(x)$. On the other hand, $K(x|y)$ can never be substantially larger than $K(x)$. This is because one way of transforming y into x is to first run a (very short) program that completely deletes the input y and then reconstruct x from scratch. The length of the former program is very small (say, c_4), and the length of the latter is $K(x)$. The shortest program transforming y into x cannot be longer than this program, which makes the transformation successfully. Therefore $K(x|y) \leq K(x) + c_4$, which establishes that $K(x|y)$ can never be substantially larger than $K(x)$ (see Li & Vitanyi, 1993, for discussion). Conditional Kolmogorov complexity is important below as a measure of the complexity of perceptual, D, given a hypothesized perceptual organization, H_i.

Kolmogorov complexity also has very close relations to Shannon's notion of information. For an information source, A, it can be shown (Li & Vitanyi, 1993, p. 194) that the entropy $H(A)$ of the source is approximately equal to the expected Kolmogorov complexity of the alternatives A_i, which comprise the source:[10]

$$H(A) = \sum_i P(A_i)I(A_i) \cong \sum_i P(A_i)K(A_i)$$
$$= \text{expected Kolmogorov complexity.} \quad (15)$$

Intuitively, this is plausible, because entropy is the expected value of surprisal $I(A_i)$, and that surprisal (rounded up to the nearest integer) was noted earlier as the minimal code length for the alternative A_i, in an informationally optimal code. Kolmogorov complexity is simply a different measure of code length for an alternative; but on average it has the same value as the original measure. There are many other parallels between the standard notion of information theory and Kolmogorov complexity, which has given rise to algorithmic information theory, a reformulation of information theory based on Kolmogorov complexity (see, e.g., Kolmogorov, 1965; Zvonkin & Levin, 1970). Notice that Kolmogorov complexity overcomes the crucial difficulty with the classical notion of information in the context of the study of perceptual organization, because it applies to a single object in isolation, not in relation to the set of alternatives.

It is therefore possible to view Kolmogorov complexity as a measure of simplicity, which is both a generalization of information theory and a generalization of coding theory. It thus provides an attractive unification of the two principal approaches used by psychologists to quantify simplicity; and it overcomes the standard difficulties with both notions from the point of view of measuring complexity of perceptual stimuli. More important than unifying the two approaches to simplicity, however, is that it allows the reconciliation of the apparently distinct likelihood and simplicity principles of perceptual organization. Before this reconciliation can be demonstrated, it is necessary to provide a more detailed mathematical treatment of the likelihood view, to which I now turn.

Likelihood and Prior Probabilities

I showed above how Bayes's theorem (Equation 1) can be used to calculate the probability of a particular hypothesis concerning the distal layout, given data concerning the perceptual stimulus.

An immediate possible concern regarding the application of Equation 1 is that the number of possible sensory stimuli is very large indeed (indeed it is infinite, aside from the limits of resolution of the sensory systems), and the probability $P(D)$ of any specific stimulus, D, will be very close to 0. This possible division by 0 is not actually problematic, however, because the numerator will always be even smaller than the denominator; otherwise $P(H_i|D)$ would be greater than 1, violating the laws of probability theory. That is, $P(H_i|D)$ is the ratio of two very small quantities, but the ratio is well-defined. Indeed, in typical applications of Bayesian statistics, $P(D)$ is typically very close to 0, but no problems arise (e.g., Lindley, 1971).

A more difficult problem arises, however, with respect to the prior probabilities of hypotheses, $P(H_i)$. Applying Bayes's theorem requires being able to specify the prior probabilities of the possible hypotheses. But, as I noted earlier, there are infinitely many distal layouts (or perceptual organizations) that are consistent with a given stimulus (at least for the degraded stimuli studied in experiments on perceptual organization[11]). For ex-

[10] This result requires the weak condition that the function from the index, i, of a state to its probability $P(A_i)$ is recursive. See Li and Vitanyi (1993, p. 194) for details.

[11] Gibson (1950, 1966, 1979) has, of course, argued that in the rich, ecologically valid conditions of normal perception, this underdetermination of the distal layout by the perceptual input does not apply. Although I would argue against this point of view, I simply note here that

ample, a two-dimensional line drawing may be the projection of an infinite number of different three-dimensional shapes; a pattern of dots may be joined up by infinitely many different curves; and so on.

Priors must be assigned so that each of this infinite number of alternatives is assigned a nonzero prior (so that it is not ruled out a priori) and so that the sum of the probabilities is 1 (otherwise the axioms of probability theory are violated). These constraints rule out the possibility of assigning each hypothesis an equal probability, because the sum of an infinite number of finite quantities, however small, will be infinite and hence not equal to 1 (the problems associated with such "improper" priors have been extensively discussed in the philosophy of science and the foundations of statistics; e.g., Carnap, 1952; Jeffrey, 1983; and Keynes, 1921). Therefore, an uneven distribution of prior probabilities is required.

In a perceptual context, an uneven distribution of priors is quite reasonable. I have already noted that, according to the likelihood interpretation, the perceptual system can be viewed as favoring (i.e., assigned a higher prior probability to) certain hypotheses (such as three-dimensional shapes containing right angles) over other hypotheses (such as highly skewed trapezoids). The likelihood view typically (though not necessarily) takes the empiricist view that certain hypotheses are favored because of past perceptual experience (e.g., that humans live in a "carpentered world"). Hence the question arises: How should priors be set for the newborn, before any perceptual experience? The obvious approach is to suggest that each hypothesis is given equal probability. However, this is not possible, because there are an infinite number of hypotheses.

The problem of assigning priors to an infinite number of hypotheses has been most intensively studied in the context of scientific inference (Horwich, 1982; Howson & Urbach, 1989; Jeffreys & Wrinch, 1921; Keynes, 1921). Indeed, the inability to find a satisfactory solution to this problem proved to be a serious difficulty for Carnap's (1950, 1952) program of attempting to devise an inductive logic (see Earman, 1992; for discussion). In the context of attempting to solve the problems of inductive inference raised by Carnap, Solomonoff (1964) showed how priors could be assigned consistently and neutrally to an infinite range of hypotheses and in doing so provided the first formulation of the principles of Kolmogorov complexity.

Solomonoff suggested that hypotheses could be neutrally assigned probabilities as follows. First, a programming language is selected. For simplicity, the very simple language of the universal Turing machine, U, is chosen. Recall that a program for U consists of arbitrary strings of 0s and 1s on a particular portion of U's tape. I now consider a two-stage process for generating objects, x, whose a priori probability we wish to assign. The first stage involves generating programs, p, for the universal Turing machine at random. This simply involves generating random strings of 0s and 1s, for example, by tossing a coin. The second stage is to run each program, p, until it halts, having generated some object (some programs will not halt, but we can ignore these). Solomonoff defines the universal a priori probability, $Q_U(x)$, of an object x as the probability that the object produced by this process is x.

Intuitively, the universal a priori probability of an object depends on how many programs there are that generate it (i.e., how many descriptions it has). If any of these programs is generated in the first stage, then the x will be produced at the second stage. In addition, it is important how long the programs (descriptions) of the object are: The shorter the program, the more likely it is to be generated at random at the first stage.

Specifically, consider a particular program, p', for U that generates the object x; that is, $U(p') = x$. The length of program (i.e., the number of 0s and 1s it contains) is denoted by $l(p')$. Then the probability of the program p' being generated at random at the first stage of the process above is the probability of $l(p')$ consecutive coin tosses coming up in a particular way: $(1/2)^{l(p')} = 2^{-l(p')}$. If p' is generated at the first stage, then at the second stage U runs p' and generates the object x.

The above calculation gives the probability of x being generated by this particular program. The universal prior probability $Q_U(x)$ is the probability of x being generated by any program. To calculate this, we must sum over all programs, p, which generate x (in symbols, $p: U(p) = x$) Thus, universal prior probability $Q_U(x)$ is

$$Q_U(x) = \sum_{p:U(p)=x} 2^{-l(p)}. \qquad (16)$$

Although the language of a particular universal Turing machine, U, has been considered, universal a priori probability, like most quantities in Kolmogorov complexity theory, is language independent.[12]

The intuition behind Solomonoff's (1964) approach to setting priors concerning sets of alternative objects is that neutrality should consist of evenhandedness between processes (programs) that generate alternative objects, not evenhandedness between objects themselves. Thus, objects that are easy to generate should be those that are expected, a priori. Furthermore, evenhandedness among processes means generating programs at random; this favors simpler processes (i.e., those with shorter programs), since they are more likely to arise by chance. It is here that the first sign of a relationship between a priori probability and simplicity is seen.

Recall that the application of Bayes's theorem requires the specification not just of prior probabilities, $P(H_i)$, but also conditional probabilities $P(D|H_j)$. These can be specified in a directly analogous way to the prior probabilities, defined as follows:

$$Q_U(x|y) = \sum_{p:U(p,y)=x} 2^{-l(p)}. \qquad (17)$$

research on perceptual organization typically concerns ecologically invalid stimuli, such as line drawings.

[12] There are a number of technical details that I ignore for simplicity. For example, it is important that the programs for the universal Turing machine are what is known as prefix codes. That is, no complete program is also the initial portion (prefix) of any other program. I also ignore the question of how, or if, to take account of those Turing machines that do not halt. On the current definition, the prior probabilities sum to less than one because of the nonhalting machines. These and other technical issues have been tackled in different ways by different researchers (e.g., Solomonoff, 1978; Zvonkin & Levin, 1970), but do not affect the present discussion.

$Q_U(x|y)$ is known as the *conditional universal distribution*. It represents the probability that a randomly generated program for U will generate object x (according to the two stages given earlier), given y as an input. Intuitively, if x is probable given y (e.g., y is a hypothesis that correctly describes some aspect of data x), then it should be easy to reconstruct x given y.

Aside from intuitive appeal, universal a priori probability (and its generalization to conditional probabilities) has a large number of attractive mathematical characteristics that have led to it, or close variants, being widely adopted in mathematics (Li & Vitanyi, 1993). Moreover, entirely independent mathematical arguments, drawing on nonstandard measure theory, converge on the same notion of universal a priori probability (Zvonkin & Levin, 1970). For these reasons, universal a priori probability has become a standard approach to assigning probabilities to infinite numbers of alternatives, in the absence of prior experience. It therefore seems reasonable to apply this notion to assigning probabilities to alternative hypotheses concerning the distal layout.

Reconciling Simplicity and Likelihood: II

The first reconciliation of simplicity and likelihood relied on Shannon's (1948) noiseless coding theorem, which relates the probability of an alternative to its code length. It was noted that

$$l_j = \left\lceil \log_2 \left[\frac{1}{P(A_j)} \right] \right\rceil$$

(by Equation 10), where l_j is the length of the code for alternative A_j, given an optimal code. The second, and deeper, reconciliation between simplicity and likelihood also requires relating probability and code length, via what Li and Vitanyi (1993) call the *coding theorem* (due to Levin, 1974), a direct analog of Shannon's result. This states that (up to a constant)

$$K(x) = \log_2 \left[\frac{1}{Q_U(x)} \right]. \tag{18}$$

The length $K(x)$ of the shortest program generating an object, x, is related to its universal prior probability by the coding theorem in the same way as optimal code length l_j is related to the probability of the alternative A_j, which it encodes.

There is another analogous result that applies to conditional probabilities, known as the conditional coding theorem, which states that (up to a constant)

$$K(x|y) = \log_2 \left[\frac{1}{Q_U(x|y)} \right]. \tag{19}$$

The argument for the reconciliation of likelihood and simplicity runs as before. As above, the likelihood principle recommends that we choose k so that

$$k = \arg\max_i (P(H_i|D))$$

Following our previous analysis this implies

$$k = \arg\min_i \left\{ \log_2 \left[\frac{1}{P(D|H_i)} \right] + \log_2 \left[\frac{1}{P(H_i)} \right] \right\}.$$

We can now apply the coding theorem and the conditional coding theorem, to derive

$$k = \arg\min_i [K(D|H_i) + K(H_i)]. \tag{20}$$

Thus, choosing hypothesis H_k in order to maximize likelihood is equivalent to choosing the H_k, which minimizes the description length of the data, D, when that data is encoded using H_k. That is, maximizing likelihood is equivalent to maximizing simplicity. The simplicity and likelihood principles are equivalent.[13]

As I noted in discussing the information-theoretic reconciliation between likelihood and simplicity above, every problem of maximizing likelihood has a dual problem of minimizing code length. This also holds in terms of Kolmogorov complexity.

The possibility of establishing an equivalence between simplicity and likelihood is not merely a matter of mathematical curiosity. It has been part of the motivation for the development of approaches to statistical inference couched in terms of simplicity, rather than probabilities, known variously as minimum message length (Wallace & Boulton, 1968; Wallace & Freeman, 1987) and minimum description length (e.g., Rissanen, 1978, 1987, 1989). These ideas have also been applied in the literatures on machine learning (e.g., Quinlan & Rivest, 1989), neural networks (e.g., Zemel, 1993), and even to problems quite closely connected to perceptual organization: automatic handwritten character recognition (Goa & Li, 1989) and computer vision approaches to surface reconstruction (Pednault, 1989). It is interesting that such a range of important applications have resulted from the reconciliation of the two psychologically motivated principles of simplicity and likelihood. It is possible to speculate that the reconciliation may have potentially important consequences for perceptual theory and for the study of cognition more generally. I consider some possible implications in the discussion.

Discussion

I now consider the wider implications of this analysis. First, I reconsider the empirical evidence that has been viewed as favoring either the likelihood, or the simplicity, principle and show how this evidence does not distinguish between the two views. Second, I outline possible residual debates concerning which principle should be viewed as fundamental. Third, I note that it can be mathematically proved that the cognitive system cannot follow either principle, as traditionally formulated, and suggest a minor modification of the principles, to retain psychological plausibility. Fourth, I consider possible applications of simplicity–likelihood principles in relation to other aspects of cognition. Finally, I discuss what the simplicity–likelihood view of perceptual organization leaves out.

[13] Li and Vitanyi (1995) provide a rigorous and detailed discussion of the mathematical conditions under which this relation holds, in the context of the relationship between minimum description length (Rissanen, 1987) and Bayesian approaches to statistical inference.

Reconsidering the Empirical Evidence

If the simplicity and likelihood principles are identical, then evidence from experiments on perceptual organization cannot favor one rather than the other. Indeed, as noted above, a wide range of phenomena in perceptual organization have been interpreted equally easily in terms of both principles. Nonetheless, various lines of empirical evidence have been viewed as favoring one view or the other. I argued that this evidence does not distinguish between the simplicity and likelihood principles of perceptual organization.

Putative evidence for the likelihood principle comes from preference for organizations that "make sense" given the structure of the natural world, but are not in any intuitively obvious sense simpler than less "natural" organizations, such as the tendency to interpret objects as if they are illuminated from above. The mathematical analysis above suggests that there must, however, be an explanation in terms of simplicity. The simplicity-based explanation can be intuitively understood as follows. Consider the simplest description not of a single stimulus, but of a typical sample of natural scenes. Any regularity that is consistent across those scenes need not be encoded afresh for each scene; rather, it can be treated as a "default." That is, unless there is an specific additional part of the code for a stimulus that indicates that the scene violates the regularity (and in what way), it can be assumed that the regularity applies. Therefore, other things being equal, scenes that respect the regularity can be encoded more briefly than those that do not. Moreover, perceptual organizations of ambiguous scenes that respect the regularity will be encoded more briefly than those that violate it. In particular, then, the perceptual organization of an ambiguous stimulus obeying the natural regularity of illumination from above will be briefer than the alternative organization with illumination from below. In general, preferences for likely interpretations also give rise to preferences for simple interpretations: If the code for perceptual stimuli and organizations is to be optimal when considered over all (or a typical sample of) natural scenes, it will reflect regularities across those scenes.

Putative evidence for simplicity involves cases of perceptual organizations that appear to be very *un*likely. Recall Leeuwenberg and Boselie's (1988) schematic drawing of what is seen as a symmetrical, two-headed horse. People do not perceive what seems to be a more likely interpretation, that one horse is occluding another. This appears to be at variance with the likelihood principle, and Leeuwenberg and Boselie hinted that this is evidence in favor of simplicity. But a likelihood explanation of this phenomenon, where likelihood applies locally rather than globally, can also be provided. That is, the perceptual system may determine the interpretation of particular parts of the stimulus according to likelihood (e.g., the fact that there are no local depth or boundary cues may locally suggest a continuous object). These local processes will not always be guaranteed to arrive at the globally most likely interpretation (see Hochberg, 1982).

Residual Debates Between Likelihood and Simplicity

From an abstract point of view, simplicity and likelihood principles are equivalent. But from the point of view of perceptual theory, one may be more attractive than the other.

Leeuwenberg and Boselie (1988) give an important argument against the likelihood principle and in favor of the simplicity principle: that the likelihood principle presupposes that patterns are interpreted, rather than explaining the interpretation of those patterns. This is because the likelihood principle holds that the structure in the world explains the structure in perceptual organization; but the theorist has no independent way of accessing structure in the world, aside from relying on the results of the principles of perceptual organization. Hence, likelihood cannot be taken as basic in explaining perceptual organization. This point of view has parallels in the literature on inductive inference using Kolmogorov complexity. For example, Rissanen (1989) argues that, although minimizing description length and maximizing likelihood are formally equivalent, the former is a preferable viewpoint as a foundation for inductive inference, because the likelihood approach presupposes that the world has a certain (probabilistic) structure, and the only way to access this structure is by inductive inference. Therefore, likelihood cannot be taken as basic in inductive inference. This line of argument provides a motivation for preferring simplicity over likelihood; but it will not, of course, be persuasive to theorists with a strong realist point of view, according to which the structure of the world is independent of human cognition, and can be objectively studied without embodying presuppositions about the structure of the human cognitive system.

A very different reason to prefer one or other principle may be derived by considering the nature of mental representations and algorithms used by the cognitive system. This issue is usefully framed in terms of Marr's (1982) distinction between "computational level" and "algorithmic level" explanation of an information-processing device. At the computational level, the questions "What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it is carried out?" are asked (Marr, 1982, p. 25). It is clear that the principles of simplicity and likelihood are typically understood at this level of explanation; they define goals for the perceptual system: maximizing simplicity or likelihood. I have shown that these goals are equivalent, and therefore that the simplicity and likelihood are equivalent principles at Marr's computational level. By contrast, at the algorithmic level the questions "How can this computation be implemented? · · · what is the representation for the input and output, and what is the transformation?" are asked (Marr, 1982, p. 25). If the simplicity and likelihood principles are viewed as rival accounts of the algorithmic level, then there may be a genuine debate between them. An algorithmic-level elaboration of the simplicity view based on description length is that algorithmic calculations are defined over description lengths, rather than probabilities. Specifically, perceptual organizations are chosen to minimize description length in some language of internal representation. This notion of description length could then be interpreted not only as an abstract measure of simplicity but also as a concrete measure of the amount of memory storage space used by the cognitive system in storing the stimulus, using the internal language. Evidence for the existence of such an internal language, along with evidence that short codes in that language correspond to preferred organizations, might be taken as evidence in favor of the priority of the simplicity view at the algorithmic level. Accord-

ing to this view, perceptual organization would not involve probabilistic calculations, but the minimization of the amount of storage space required in memory. Similarly, evidence for the internal representation of information concerning probabilities and evidence that this information was used in algorithms for probabilistic calculations could be used in favor of the likelihood view. According to this view, the cognitive system would not be minimizing the amount of storage space required in memory but would be conducting explicit probabilistic reasoning.

How could simplicity and likelihood accounts of the algorithmic level be distinguished empirically? Evidence distinguishing between these competing algorithmic-level accounts is likely to be very difficult to collect, particularly because the most obvious line of attack, experimental study of the structure of perceptual organizations for various kinds of stimulus, is inapplicable because of the equivalence of the two principles at the computational level. Nonetheless, the two principles can (although they need not) be interpreted as making distinct empirical claims concerning the algorithms underlying perceptual organization.

Overall, even the provable equivalence of the simplicity and likelihood principles may not preclude arguments over which is fundamental, based either on philosophical or algorithmic-level concerns.

A Psychologically Plausible Modification of the Simplicity–Likelihood Principle

In this article, I have been concerned with showing that the simplicity and likelihood principles are the same. I now note that it is provably impossible that the cognitive system follows either principle in its strongest form. For concreteness, I discuss the simplicity principle, although the same considerations apply to the likelihood principle.

Is it possible that the perceptual system invariably chooses the simplest (most probable) organization of the stimulus? Assuming that the perceptual system is computational, there are strong reasons to suppose that it cannot. The general problem of finding the shortest description of an object is provably uncomputable (Li & Vitanyi, 1993). It is important to note that this result applies to any kind of computer, whether serial or parallel, and to any style of computation, whether symbolic, connectionist or analog.[14] It is also intuitively obvious that people cannot find arbitrary structure in perceptual scenes. To pick an extreme example, a grid in which pixels encoded the binary expansion of π would, of course, have a very simple description, but this structure would not be identified by the perceptual system; the grid would, instead, appear completely unstructured.

It is clear, then, that the perceptual system cannot, in general, maximize simplicity (or likelihood) over all perceptual organizations, pace traditional formulations of the simplicity and likelihood principles. It is, nonetheless, entirely possible that the perceptual system chooses the simplest (or most probable) organization that it is able to construct. That is, simplicity may be the criterion for deciding between rival organizations. To retain psychological plausibility, the simplicity (likelihood) principles must be modified, to say that the perceptual system chooses the simplest (most likely) hypothesis it can find; this will not, in general, be the simplest possible hypothesis.

Notice, however, that this revised formulation—that the perceptual system chooses the simplest (most probable) organization that it can construct—does not affect the equivalence between the simplicity and likelihood principles. Because there is a one-to-one relationship between probabilities and code lengths, a small (but not quite minimal) code will correspond to a probable (but not quite most probable) organization. So maximizing simplicity and maximizing likelihood are equivalent, even when maximization is approximate rather than exact.

Implications for Other Psychological Processes

Could the relationship between simplicity and likelihood principles be relevant to other areas of psychology? One possible application is to areas of low-level perception in which compression of the sensory signal has been viewed as a central goal (Atick & Redlich, 1990; Barlow, Kaushal, & Mitchison, 1989; Blakemore, 1990).[15] The goal of compression is frequently viewed as stemming from limitations in the information-carrying capacity of the sensory pathways. However, the equivalence of maximizing compression (i.e., minimizing description length) and maximizing likelihood indicates a complementary interpretation. It could be that compressed perceptual representations will tend to involve the extraction of features likely to have generated the sensory input. According to this complementary interpretation, perceptual inference occurs in the very earliest stages of perception (e.g., as implemented in mechanisms such as lateral inhibition in the retina), where neural coding serves to compress the sensory input.

The relationship between the simplicity and likelihood principles may also be relevant to the relationship between inference and memory. Perceptual, linguistic, or other information is not remembered in a "raw" form, but in terms of high-level categories and relations organized into structured representations (e.g., Anderson, 1983; Hinton, 1979; Johnson-Laird & Stevenson, 1970) such as "sketches" (Marr, 1982), schemata (Bobrow & Norman, 1975), scripts (Schank & Abelson, 1977), or frames (Minsky, 1977). Two constraints on such memory organizations suggest themselves: (a) that they allow the structure of the world to be captured as well as possible, and (b) that they allow the most compact encoding of the information to be recalled, so that memory load is minimized. Prima facie, these goals might potentially conflict and require the cognitive system to somehow make appropriate trade-offs between them. But, as we have seen, the goal of capturing the structure of the world and the goal of providing a compressed representation can be seen as equivalent.

---

[14] Strictly speaking, this statement requires a caveat. It applies only to neural network and analog styles of computation where states need not be specified with infinite precision. This restriction is very mild, since it seems extremely unlikely, to say the least, that an infinite precision computational method could be implemented in the brain, particularly in view of the apparently noisy character of neural signals.

[15] These theorists advocate a variety of distinct proposals concerning the objectives of perception. They are all closely related to the goal of minimizing description length, although not necessarily couched in these terms.

More generally, the relationship between simplicity and likelihood principles may be useful in integrating psychological theories that stress probabilistic inference (e.g., Anderson, 1990; Fried & Holyoak, 1984; Oaksford & Chater, 1994) and those that stress the importance of finding compressed representations (Redlich, 1993; Wolff, 1982). Furthermore, the possibility of viewing cognitive processes from two complementary perspectives may throw valuable light on both kinds of accounts.

What the Simplicity and Likelihood Principles Leave Out

This article has been concerned with showing that the simplicity and likelihood principles are identical. I have not considered whether the unified simplicity–likelihood principle really governs perceptual organization. This question is clearly a large topic for future research, and a full discussion is beyond the scope of this article. Nonetheless, two preliminary comments are worth making.

First, on the positive side, the evidence for the unified simplicity–likelihood principle is the sum of the evidence that has been previously adduced in favor of the simplicity and likelihood principles.

Second, on the negative side, the simplicity–likelihood principle ignores a factor that may be of considerable importance: the interests and potential actions of the agent. The application of the simplicity–likelihood principle can be viewed as "disinterested contemplation" of the sensory stimulus: The simplest encoding, or the most likely environmental layout, is sought without any concern for the specific interests of the perceiver. But perceivers are not disinterested; they are concerned with particular goals and actions and hence with aspects of the environment relevant to those goals and actions. The frog's perceptual system is, for example, geared toward the detection of dark, fast, moving concave blobs (among other things), because this information allows the frog to perform actions (snapping in the appropriate direction) that satisfy its interests (eating flies; Lettvin, Maturana, McCullough, & Pitts, 1959). Similarly, the fly is sensitive to correlates of optic expansion because of the importance of this information in the timing of landing (Poggio & Reichardt, 1976; Reichardt & Poggio, 1976). It has been suggested that many aspects of human perception, too, may be explained in terms of people's interests and motor abilities (e.g., Gibson, 1979). Indeed, one of the most important tenets of Gibson's "direct" perception is that agents pick up properties of the world that afford various actions to the agent, such as being lifted, reached, grasped, or climbed. The important point here is that affordances are defined in terms of the actions and goals of the agent: just those factors that the simplicity and likelihood principles ignore. The importance of affordances and like notions has been widely discussed in many areas of perception, but has not been a focus of interest in the literature on perceptual organization. It is possible that the interests and actions of the agent are not relevant in organizational processes in perception. On the other hand, it is possible that here, too, it is necessary to view the perceiver not merely as a disinterested observer of sensory stimuli, but as using sensory input to determine appropriate actions. Devising experimental tests of the importance of interests and actions in the context of perceptual organization is an important empirical challenge. If these factors do influence perceptual organization, then the simplicity–likelihood principle must be elaborated or replaced.

Conclusion

I have shown that the simplicity and likelihood principles in perceptual organization are equivalent, given natural interpretations of simplicity in terms of shortest description length and of likelihood in terms of probability theory. This implies that the empirical and theoretical debate over whether perceptual organization maximizes simplicity or maximizes likelihood is misguided. Instead, the fundamental question is whether, or to what extent, perceptual organization is maximizing simplicity and maximizing likelihood.

References

Anderson, J. A. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. A. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation, 2,* 308–320.

Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation, 4,* 196–210.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61,* 183–193.

Attneave, F. (1959). *Applications of information theory to psychology.* New York: Holt, Rinehart & Winston.

Attneave, F. (1972). Representation of physical space. In A. W. Melton & E. J. Martin (Eds.), *Coding processes in human memory* (pp. 283–306). Washington, DC: Winston.

Attneave, F. (1981). Three approaches to perceptual organization. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 417–422). Hillsdale, NJ: Erlbaum.

Attneave, F. (1982). Prägnanz and soap bubble systems: A theoretical exploration. In J. Beck (Ed.), *Organization and representation in perception* (pp. 11–29). Hillsdale, NJ: Erlbaum.

Attneave, F., & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception & Psychophysics, 6,* 391–396.

Barlow, H. B., Kaushal, T. P., & Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural Computation, 1,* 412–423.

Blakemore, C. (Ed.). (1990). *Vision: Coding and efficiency.* Cambridge, England: Cambridge University Press.

Bobrow, D. G., & Norman, D. A. (1975). Some principles of memory schemata. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding: Essays in cognitive science* (pp. 131–150). New York: Academic Press.

Boolos, G. S., & Jeffrey, R. C. (1980). *Computability and logic.* Cambridge, England: Cambridge University Press.

Brunswick, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.

Buffart, H., Leeuwenberg, E., & Restle, F. (1981). Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 241–274.

Carnap, R. (1950). *Logical foundations of probability.* Chicago: University of Chicago Press.

Carnap, R. (1952). *The continuum of inductive methods.* Chicago: University of Chicago Press.

Chaitin, G. J. (1966). On the length of programs for computing finite

binary sequences. *Journal of the Association for Computing Machinery, 13,* 547–569.

Chater, N. (1989). *Information and information processing.* Unpublished doctoral dissertation, Centre for Cognitive Science, University of Edinburgh, Edinburgh, Scotland.

Cheeseman, P. (1995). On Bayesian model selection. In Wolpert, D. (Ed.), *The mathematics of generalization* (pp. 315–330). Redwood City, CA: Addison-Wesley.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley.

de Finetti, B. (1972). *Probability, induction and statistics.* New York: Wiley.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* Cambridge, MA: MIT Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A222,* 309–368.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* New York: Wiley.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MO: Erlbaum.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Goa, Q., & Li, M. (1989). An application of the minimum description length principle to on-line recognition of handprinted alphanumerals. In *11th International Joint Conference on Artificial Intelligence* (pp. 843–848). San Mateo, CA: Morgan Kaufman.

Grenander, U. (1976–1981). *Lectures in pattern theory I, II and III: Pattern analysis, pattern synthesis and regular structures.* Heidelberg, Germany: Springer-Verlag.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review, 74,* 88–95.

Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin, 97,* 155–186.

Helmholtz, H. L. F. von (1962). *Treatise on physiological optics* (Vol. 3; J. P. Southall, Ed. and Trans.). New York: Dover. (Original work published 1910)

Hinton, G. E. (1979). Some demonstrations of the effects of structural description in mental imagery. *Cognitive Science, 3,* 231–250.

Hochberg, J. (1982). How big is a stumulus? In J. Beck (Ed.), *Organization and representation in perception* (pp. 191–218). Hillsdale, NJ: Erlbaum.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figure "goodness." *Journal of Experimental Psychology, 46,* 361–364.

Horwich, P. (1982). *Probability and evidence.* Cambridge, England: Cambridge University Press.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach.* La Salle, IL: Open Court.

Jeffrey, R. C. (1983). *The logic of decision.* New York: McGraw-Hill.

Jeffreys, H., & Wrinch, D. (1921). On certain fundamental principles of scientific enquiry. *Philosophical Magazine, 42,* 269–298.

Johansson, G. (1950). *Configurations in event perception.* Stockholm: Almqvist & Wiksell.

Johnson-Laird, P. N., & Stevenson, R. (1970). Memory for syntax. *Nature, 227,* 412.

Keynes, J. M. (1921). *A treatise on probability.* London: Macmillan.

Koffka, K. (1963). *Principles of Gestalt psychology.* New York: Harcourt, Brace & World. (Original work published 1935)

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission, 1,* 1–7.

Leeuwenberg, E. (1969). Quantitative specification of information in sequential patterns. *Psychological Review, 76,* 216–220.

Leeuwenberg, E. (1971). A perceptual coding language for perceptual and auditory patterns. *American Journal of Psychology, 84,* 307–349.

Leeuwenberg, E., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review, 95,* 485–491.

Lettvin, J. Y., Maturana, W. S., McCullough, W., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers, 47,* 1940–1951.

Levin, L. A. (1974). Laws of information conservation (non-growth) and aspects of the foundations of probability theory. *Problems in Information Transmission, 10,* 206–210.

Li, M., & Vitanyi, P. (1993). *An introduction to Kolmogorov complexity and its applications.* New York: Springer-Verlag.

Li, M., & Vitanyi, P. (1995). *Computational machine learning in theory and praxis.* (Tech. Rep. No. NC-TR-95-052) London: Royal Holloway College, University of London, Department of Computer Science.

Lindley, D. V. (1971). *Bayesian statistics: A review.* Philadelphia: Society for Industrial and Applied Mathematics.

Mach, E. (1959). *The analysis of sensations and the relation of the physical to the psychical.* New York: Dover Publications. (Original work published 1914)

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Mellor, D. H. (1971). *The matter of chance.* Cambridge, England: Cambridge University Press.

Minsky, M. (1967). *Computation: Finite and infinite machines.* Englewood Cliffs, NJ: Prentice Hall.

Minsky, M. (1977). Frame-system theory. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 355–376). New York: McGraw-Hill.

Mumford, D. (1992). Pattern theory: A unifying perspective. In Joseph, A., Mignot, F., Murat, F., Prum, B., & Rentschler, R. (Eds.), *Proceedings of the First European Congress of Mathematics* (pp. 187–224). Basel, Switzerland: Birkhäuser Verlag.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Odifreddi, P. (1989). *Classical recursion theory.* Amsterdam: Elsevier.

Pednault, E. P. D. (1989). Some experiments in applying inductive inference principles to surface reconstruction. In *11th International Joint Conference on Artificial Intelligence* (pp. 1603–1609). San Mateo, CA: Morgan Kaufman.

Perkins, D. N. (1972). Visual discrimination between rectangular and nonrectangular parallelepipeds. *Perception and Psychophysics, 12,* 396–400.

Perkins, D. N. (1982). The perceiver as organizer and geometer. In J. Beck (Ed.), *Organization and representation in perception* (pp. 73–93). Hillsdale, NJ: Erlbaum.

Poggio, T., & Reichardt, W. (1976). Visual control of orientation behavior in the fly: Part II. Towards the underlying neural interactions. *Quarterly Review of Biophysics, 9,* 377–438.

Pomerantz, J. R. (1981). Perceptual organization in information processing. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 141–180). Hillsdale, NJ: Erlbaum.

Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In K. R.

Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Volume II. Cognitive processes and performance* (pp. 36:1–45). New York: Wiley.

Quinlan, J., & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation, 80,* 227–248.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation, 5,* 289–304.

Reichardt, W., & Poggio, T. (1976). Visual control of orientation behavior in the fly: Part I. A quantitative analysis. *Quarterly Review of Biophysics, 9,* 311–375.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77,* 481–495.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review, 86,* 1–24.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica, 14,* 465–471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B, 49,* 223–239.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry.* Singapore: World Scientific.

Rock, I. (1975). *An introduction to perception.* New York: Macmillan.

Rock, I. (1983). *The logic of perception.* Cambridge, MA: MIT Press.

Rogers, H., Jr. (1967). *Theory of recursive functions and effective computability.* New York: McGraw-Hill.

Royer, F. L., & Garner, W. R. (1966). Response uncertainty and perceptual difficulty of auditory temporal patterns. *Perception and Psychophysics, 1,* 41–47.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Erlbaum.

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). *The influence of culture on visual perception.* Indianapolis, IN: Bobbs-Merrill.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 279–342). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review, 79,* 369–382.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70,* 534–546.

Sober, E. (1975). *Simplicity.* Oxford, England: Clarendon Press.

Solomonoff, R. J. (1964). A formal theory of inductive inference, Parts 1 and 2. *Information and Control, 7,* 1–22, 224–254.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions in Information Theory, 24,* 422–432.

Vitz, P. C., & Todd, T. C. (1969). A coded element of the perceptual processing of sequential stimuli. *Psychological Review, 76,* 433–449.

von Mises, R. (1981). *Probability, statistics, and truth.* New York: Dover. (Original work published 1939)

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computing Journal, 11,* 185–195.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B, 49,* 240–251.

Wolff, J. G. (1982). Language acquisition, data compression, and generalization. *Language and Communication, 2,* 57–89.

Zemel, R. (1993). A minimum description length framework for unsupervised learning. Unpublished doctoral dissertation, Department of Computer Science, University of Toronto, Toronto, Canada.

Zvonkin, A. K., & Levin, L. A. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys, 25,* 83–124.

Received June 7, 1995
Revision received December 1, 1995
Accepted December 8, 1995 ■