Information, run structure and binary pattern complexity[1]

PAUL C. VITZ
NEW YORK UNIVERSITY

Two information processing models of the perception of binary patterns are outlined. The H (k-span) model assumes S evaluates the transitional uncertainty associated with the different k-tuples constituting the pattern. The H (run-span) model assumes S codes the pattern into runs and then evaluates the transitionaluncertainty associated with the pattern's run-tuples. Both models give a measure of the complexity of a pattern. The models are tested by using the complexity measures to predict various response indices of pattern complexity; e.g. judged pattern complexity, mean number of words to describe a pattern, accuracy of recall of a pattern. H (run-span) proved to be highly correlated with all of the pattern complexity indices. H (k-span) received only moderate support.

The purpose of this paper is to present two information measures of the complexity of binary sequences or strings with special emphasis on repeating binary patterns such as: aab, aab...; abaabb, abaabb...; etc. Such measures are of interest because they are based on hypotheses of how Ss code and organize sequential stimuli; binary patterns are of special interest because they are easily generated, and their simplicity presumably exposes the process of perceptual organization more clearly than other patterns.

Existing measures of the complexity of sequential binary patterns are primarily response measures. Glanzer and Clark (1963a, b) propose that the mean number of words used to describe a pattern be used as a measure of pattern complexity. They presented tachistoscopically the 256 binary numbers of length eight. After a brief exposure S reproduced the pattern and an accuracy of recall score was obtained for each number. Another group of Ss was shown each number for 30 sec and then asked to write a short description of it. The mean number of words used to describe the number correlated highly with the accuracy scores. Depending upon the particular stimulus representation of the binary number, the correlation between mean verbalization length (MVL) and accuracy ranged from -.79 to -.83. Glanzer and Clark (1964) also demonstrate that MVL is correlated to the same extent with judged complexity of conventional geometric figures. Thus, MVL is a widely applicable response measure of pattern complexity.

Another response measure of binary pattern complexity has been proposed by Royer and Garner (1966). They also used binary numbers of length eight, but presented the patterns as repeating series of two distinct tones. There are equivalences existing within the 256 numbers due to complementarity and cyclic repetition. When these equivalences are taken into account there remain 20 different repeating 8-place binary patterns. The S listened to the repeating patterns and at whatever time he believed he could follow the pattern he began predicting. Although the Ss had equal opportunity to select any of the logically possible starting points they usually showed marked preferences for particular starting points. As a measure of perceptual complexity Royer and Garner computed the uncertainty associated with the probabilities of starting each pattern at different possible starting points. This response point uncertainty (RPU) measure is proposed as a measure of S's difficulty in organizing a pattern and can be considered a measure of pattern complexity.

Although they are useful, there are fundamental difficulties with response measures of pattern complexity. One important limitation is that complexity cannot be evaluated in advance. Because of the large number of stimulus patterns of experimental interest, having to run Ss to get a measure is a serious restriction. In addition, a correlation between two response measures is often hard to interpret causally; e.g., does the mean number of words to describe a pattern determine perceptual accuracy?

The complexity measures developed here, H(k-span) and H(run-span), are derived from stimulus characteristics of the pattern, and are based on simple models of how binary sequences are perceptually organized.

Model 1: H (k-span)

For the present we will consider only repeating binary patterns. (The model's restriction to binary patterns is only for convenience since it can be applied to repeating patterns consisting of any number of different elements, e.g., trinary patterns such as aabbbcc, aabbbcc, etc.) The major operation of the H(k-span) model is evaluation of the uncertainties in the pattern starting with H_0, the uncertainty associated with the absolute probabilities of the two events, then progressing to first order transitional uncertainty, H_1, then to second order transitional uncertainty, H_2, to $H_3, \ldots H_n$. Evaluation of the

Copyright 1968, Psychonomic Journals, Santa Barbara, Calif.

higher order transition uncertainties stops when H_n equals zero. Of course, to evaluate the increasingly higher transition uncertainties an increasingly longer memory span is required, hence the name H(k-span). The complexity of a pattern is defined as the total of the average uncertainties evaluated during this process, i.e.,

$$H(\text{k-span}) = \sum_{n=0}^{N} H_n$$

The following transition matrices summarize the model for the repeating pattern, aabbab, aabbab,.... Although something like this analysis is described in Attneave (1959), the representation of the transition matrices and uncertainties used here follows the Markov model as presented by Binder and Wolin (1964), in which $H = -\Sigma p_i\, p_{ij} \log_2 p_{ij}$, where p_i and p_{ij} are the absolute probability of being in state i and the transition probability of going to state j after being in state i, respectively.

Repeating Pattern: a a b b a b (H_0: memory = k = 0) In this special case the Markov measure becomes the familiar $H = -\Sigma p_i \log_2 p_i$, i.e., the information associated with the absolute probabilities of a and b. For the binary pattern, aabbab,

$$H(\text{k-span}) = \sum_{n=0}^{3} H_n = 1.00 + 0.92 + 0.66$$

$$= 2.58 \text{ bits.}$$

The measure H(k-span), which summarizes the preceding process, is based on a model of a theoretically perfect perceiver since there are no adjustments or parameters included to account for human deviation from the model. It is assumed that human Ss closely approximate the model.

A more detailed description of the k-span model will not be given since its general adequacy can be tested without further development. The H(k-span) values for all the repeating binary numbers of lengths 1-4, 6, and 8 are presented in Table 1.

For computational purposes it is important to note that the laborious transition matrix analysis previously outlined can be avoided. The H(k-span) value of any pattern is equal simply to the $\log_2$ of the length of the pattern, where the length is the number of elements in the pattern. Thus, all repeating patterns of length 6 have H(k-span) values of 2.58, etc. The basis of this interesting and use-

Transition Matrix Summary of H (k-span) for Repeating Pattern: aabbab

p_i	state i	
1/2	a	$h_0a = 1/2 \log_2 1/2 = 0.50$
1/2	b	$h_0b =$ 0.50
		$H_0 = h_0a + h_0b = 1.00$

H_1: memory = k = 1.

p_i	state i	state j		
	k = 1	a	b	$h_1a = p_a ((p_{aa} \log_2 p_{aa}) + (p_{ab} \log_2 p_{ab}))$
				$h_1a + 1/2 ((1/3 \log_2 1/3) + (2/3 \log_2 2/3))$
1/2	a	1/3	2/3	
1/2	b	2/3	1/3	$h_1a + 1/2 (0.92)$
				$h_1b + 1/2 (0.92)$
				$H_1 = h_1a + h_1b = 0.92$

H_2: memory = k = 2

p_i	state i	state j		
	k = 2	a	b	$H_2 = 1/6 (0) + 1/3 (1.00) + 1/3 (1.00) + 1/6 (0)$
1/6	aa	0	1	
1/3	ab	1/2	1/2	
1/6	bb	1	0	$H_2 = 0.66$
1/3	ba	1/2	1/2	

H_3: memory = k = 3

p_i	state i	state j		
	k = 3	a	b	
1/6	aab	0	1	
1/6	abb	1	0	$H_3 = 0.00$
1/6	bba	0	1	
1/6	bab	1	0	
1/6	aba	1	0	
1/6	baa	0	1	

Table 1. Measures of the Complexity of Repeating Binary Patterns

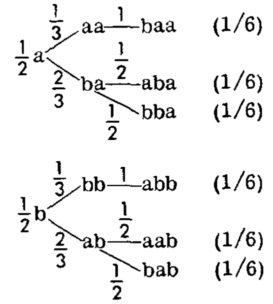
No.	Repeating Pattern	H (k-span) Complexity	H (run-span) Complexity	Mean Judged Complexity Exp. 1* N=22	Mean Judged Complexity Exp. 2** N=15	Mean Verbalization Length*** Exp. 1 N=22
1	a	0.00	0.00	.14	1.00	1.23
2	ab	1.00	1.00	1.91	2.40	4.68
3	aab	1.58	1.00	3.96	--	5.05
4	aabb	2.00	1.00	--	5.60	--
5	aaab	2.00	1.00	--	5.40	--
6	aaabbb	2.58	1.00	2.92	--	5.18
7	aaaabb	2.58	1.00	3.82	--	5.32
8	aaaaab	2.58	1.00	2.77	--	5.14
9	aaabab	2.58	2.00	5.95	--	10.36
10	abaabb	2.58	2.00	6.96	--	10.14
11	aaaabbbb	3.00	1.00	--	5.33	--
12	aaaaabbb	3.00	1.00	--	7.07	--
13	aaaaaabb	3.00	1.00	--	5.93	--
14	aaaaaaab	3.00	1.00	--	3.60	--
15	aaababab	3.00	2.58	--	14.33	--
16	aaabbbab	3.00	2.00	--	14.00	--
17	aaaabbab	3.00	2.00	--	14.13	--
18	aaaaabab	3.00	2.00	--	10.80	--
19	aaaababb	3.00	2.00	--	15.60	--
20	aaaabaab	3.00	2.00	--	11.33	--
21	aaabaabb	3.00	2.00	--	15.40	--
22	aaabbaab	3.00	2.00	--	14.47	--
23	aaabbabb	3.00	2.00	--	14.13	--
24	aabbabab	3.00	2.58	--	16.33	--
25	aababbab	3.00	2.58	--	17.80	--
26	aabaabab	3.00	2.58	--	15.33	--

* *differences between means > .80 significant at .01 level, 2-tail.*

** *for patterns 1-14, differences between means > 1.00 significant at .05 level, 2-tail. for patterns 15-26, differences between means > 1.99 significant at .05 level, 2-tail.*

*** *differences between means > .30 significant at .05 level, 2-tail.*

ful equivalence can be observed by representing the transition probabilities outlined above with a tree diagram. For the pattern aabbab this diagram is:



This diagram, which ends when perfect predictability of each point in the pattern has been reached, demonstrates that any pattern must have n final and equally probable branches for unique prediction to be possible at each point. Thus, in the final analysis all patterns have H(k-span) values equal to $\log_2$ of the pattern length. The transition matrix is shown to give a more descriptive account of pattern processing than the simple measure, $\log_2$ of the length provides.

Model 2: H (run-span)

The probabilities which are used in the H(k-span) measure are associated with the k-tuples of different lengths. Assuming these are the stimulus elements is mathematically reasonable but makes little psychological sense. There is much evidence strongly implying that S codes a series of binary events into coded elements quite different from the k-tuples. Perhaps the simplest assumption about this coding process is that the sequence is coded into runs of like events. Evidence that Ss code binary patterns into runs is presented by Restle (1961, 1966), Keller (1963), Rose and Vitz (1966), and Vitz and Todd (1967).

The H(run-span) model assumes that the sequence is first coded into runs and that runs are treated as the only stimuli in the pattern. A run is defined as any event repeated n times and preceded and followed by any event other than itself. Except for this stimulus coding principle the H(run-span) model is identical with H(k-span). That is, the model represents the perceiver as analyzing the uncertainties associated with the run-tuples just like the H(k-span) model analysis of the k-tuple uncertainty. For example, the repeating pattern aabbab is first broken into runs: aa, bb, a, and b. There are four runs and each occurs with probability 1/4 and hence H_{r0} = 2.00 bits. Next the first order run transition uncertainty, H_{r1}, is evaluated with the runs as the stimuli and responses. In this case each run perfectly predicts the next run and $H_{r1} = 0.00$. The H(run-span) total complexity of a pattern is the sum of the average uncertainties associated with the run-tuple transition matrices, i.e.,

$$H(\text{run-span}) = \sum_{n=0}^{N} H_{rn}$$

For the pattern, aabbab, this value is 2.00 bits. As with H(k-span) the transition matrix evaluation of H(run-span) can be avoided, since H(run-span) also is equal to $\log_2$ of the length of the pattern, where the length is the total number of runs in the pattern, e.g., the pattern aaababab has a length of six runs and H(run-span) = 2.58.[2]

The H(run-span) values for the repeating binary patterns of lengths 1-4, 6, and 8 also are presented in Table 1.

Although H(k-span) and H(run-span) are models of perceptual processing they are related to well known models of learning. H(k-span) is similar to a model first stated by Burke and Estes (1957) and later extensively elaborated by Restle (1961, pp. 109-111). The H(run-span) model is similar to Restle's run-model of learning (1961, 1966) and to his (1967) recent grammatical or rule analysis of binary pattern learning.

The following two experiments are tests of the H(k-span) and H(run-span) measures of pattern complexity. Both experiments require Ss to judge the complexity of repeating binary patterns, and it is assumed that these judgments should correlate highly with the stimulus complexity measures. Failure to do so would cast considerable doubt on any measure's validity. In addition to the complexity judgments the relation of H(k-span) and H(run-span) to other response measures of perceptual complexity are examined.

EXPERIMENT 1

Method

Stimuli. The eight patterns used are those in Table 1, Nos. 1-3 and 6-10. These are all the logically different repeating binary patterns of lengths 1, 2, 3, and 6; i.e., all other possible binary patterns were equivalent either by complementarity or cyclic repetition. Each pattern was typed on a white 3 x 5 card and repeated until the string of characters was 30 in length, e.g., pattern No. 2 was repeated 15 times, etc. The binary characters were a small circle and a small isosceles triangle. Two series were constructed. Series 1 was the same as Series 2 except circles were substituted for triangles and vice versa.

Procedure. The S was given a randomized deck of 28 3 x 5 cards; on each card were two of the patterns, one above the other. The task was to record on a response sheet which of the repeating patterns was simpler: The simpler pattern was described as the pattern S found the "easiest and fastest to comprehend." The cards presented the 28 possible paired comparisons. A pattern's complexity score is the mean number of times it was not chosen as the simpler. After judging, S was given four cards; on each were two of the eight patterns, and S was told to write a short accurate description of the repeating part of the pattern, avoiding all unnecessary words.

The MVL is the mean number of words used to describe the repeating pattern. Words used to describe the binary characters, e.g., the size of the circle or triangle, or words mentioning the number of times the pattern was repeated were not counted. Such words were easily identified and since they did not refer to the repeated pattern their omission seems required. The Ss were college undergraduates fulfilling an introductory psychology course experiment participation requirement and were run in small groups. Eleven Ss judged each series; N = 22.

Results

The complexity values for each of the eight stimulus patterns are presented in Table 1, in the columns titled H(k-span) and H(run-span), Mean Verbalization Length, and Judged Complexity, Experiment 1. Table 2 presents the correlations among the four measures. It is clear from these correlations that the judged complexity values are related to both H(k-span) and H(run-span) with H(run-span) the markedly better predictor. In addition the mean verbalization length (MVL) is equally highly correlated with judged complexity.

The stimuli used in Experiment 1 are a small number of the possible repeating binary patterns and it is possible the above correlations are not representative of the larger population of such patterns. Experiment 2 is designed to evaluate the same relationship with a larger number of patterns.

EXPERIMENT 2

Method

Stimuli. The 20 patterns used were identical with those numbered 1, 2, 4, 5, and 11-26 in Table 1. These are all of the logically different repeating binary numbers of lengths 1, 2, 4, and 8. Each pattern was typed on a white 3 x 5 card and repeated

Table 2. Correlations between the measures of pattern complexity used in Experiment 1. (N = 22)

	H (run-span)	MVL	Judged Complexity
H (k-span)	.73	.71	.73
H (run-span)	--	.99	.95
Mean Verbalization Length (MVL)	--	--	.96

until the string of characters was 32 in length. The binary characters were lower case a and b.

Procedure. The greater number of patterns precluded a paired-comparison procedure, hence the following rank order method was used. The 20 cards were placed haphazardly on a desk in front of S. He was asked to first look at each card and observe the repeating pattern. Next S was told to divide the cards into two groups, the 10 simplest and the 10 most complex. The term "simple" was defined as in Experiment 1. Next, S set aside the 10 complex patterns and rank ordered the 10 simplest from the simplest to the most complex. The S's first seven ranks were recorded and those seven cards were removed. The remaining three cards and the 10 in the complex group were again placed haphazardly in front of S. The S placed the three most complex patterns aside and ranked the remaining 10 from simplest to most complex. When this was done E recorded the first seven ranks out of this 10 and removed the corresponding seven cards. The remaining three cards and the three previously judged as most complex were then placed in front of S, and he made a final ranking of the last six cards from simplest to most complex. The Ss were undergraduate and graduate student volunteers: N = 15.

RESULTS AND DISCUSSION

The mean complexity ranks for the 20 patterns are shown in Table 1 in the column: Judged Complexity, Experiment 2. For these 20 patterns, H(k-span) correlated with judged complexity .67 while H(run-span) correlated .94. Because of the weak empirical support for the H(k-span) measure in both experiments it will be omitted from further consideration.

The published response point uncertainty data of Royer and Garner (1966) previously described allow a test of H(run-span). For the same 20 repeating binary numbers, H(run-span) correlates .92 with RPU. (The RPU values for each pattern are presented by Royer and Garner in their Table 1. The simple pattern, a, a, a,... was included in the preceding correlation.)

One of the limitations of H(run-span) which is part of the previous rationale is the restriction to repeating or periodic patterns. The following analysis argues that with a few plausible assumptions the model can be applied to nonperiodic binary patterns, such as the 8-place binary numbers used by Glanzer and Clark (1963a, b).

H (run-span): Nonperiodic Patterns

Consider the nonperiodic binary pattern aaabbbaa. (If repeated, of course, this pattern would become aaaaabbb.) The previous development of H(run-span) would allow the computation of H_{r0}, as the pattern consists of three runs: aaa, bb, and aa. But H_{r1} cannot be defined since the run aa is not followed by any other run. To avoid this problem we will assume that all finite binary patterns consist not only of the runs as defined but also of two additional runs: (1) a run representing the stimulus preceding the first binary element which will be symbolized BR (Beginning Run). (2) Also, a run following the last element is postulated. This end run will be symbolized ER. Further, it is assumed that the pattern is scanned from left to right and when ER is reached, the process is re-set, that is, it returns to BR.

The pattern aaabbbaa is now represented as consisting of five runs: BR, aaa, bbb, aa, and ER, each occurring with probability 1/5. At the first run transition level, H_{r1}, the pattern is completely predictable since each run is followed by only one of the five possible elements. The new measure, nonperiodic H(run-span) or $_{np}$H(run-span) is, again, the sum of the average uncertainties at each of the run-tuple transition levels. And its computation also is simplified by using the equivalent measure of $\log_2$ of the number of runs, including the BR and ER. For aaabbbaa, $_{np}$H(run-span) = $\log_2 5$ or 2.32 bits.

To test the $_{np}$H(run-span) measure, the data from the previously described experiment by Glanzer and Clark (1963b) were obtained.[3] In their Experiment 2, each of the 256 binary numbers of length 8 received a mean verbalization length score (MVL) and an accuracy of recall score. The accuracy score was the proportion of the 80 Ss who reproduced the pattern without any error. Glanzer and Clark report a correlation of -.81 between the 256 MVL and accuracy scores. However, of the 256 patterns it seems reasonable to combine the scores of logically equivalent patterns. Thus, the scores of each pattern and its complement were averaged and the scores of each pattern and its right-left reversal, if it had one, were averaged. This averaging reduced the 256 patterns to 72 different patterns and presumably results in a more representative score for a pattern since differences between the scores of equivalent patterns are due to random error, to the easier perception of one of the binary events, or to pre-experimental perceptual habits. For example, there is some evidence in Glanzer and Clark's accuracy data that 0 was more easily perceived than 1.

Using these data the correlations between $_{np}$H(run-span) and MVL is .89 and between $_{np}$H(run-span) and accuracy the correlation is -.86.[4] Again, H(run-span) received considerable support. However, with these data there was some evidence pointing to weakness in the H(run-span) model. As is shown in Table 1, there are only four different H(run-span) values for the 26 patterns, and yet many of the patterns with identical H(run-span) values have different judged complexity scores. In many cases these differences are quite reliable. The measure's failure to discriminate among these patterns is probably because coding into runs is too simple a description of the coding process. For example, Ss might also code patterns into alternations (see Royer & Garner, 1966; and Royer, 1967). An examination of the correlation scatterdiagrams of $_{np}H$(run-span) and MVL and accuracy supports this contention. The pattern 10101010 and other patterns close to complete alternation had $_{np}H$(run-span) values that were too high. If alternations, instead of runs, were treated as coded elements then the H values for these patterns would have been lowered and prediction improved. Perhaps, a coding principle or rationale can be developed for treating some patterns as composed of runs and/or of alternations. If so an H measure on these two kinds of coded elements would correlate quite highly with most of the response indices treated here.

References

Attneave, F. *Applications of information theory to psychology.* New York: Holt, 1959.

Binder, A., & Wolin, R. B. Informational models and their uses. *Psychometrika,* 1964, 29, 29-54.

Burke, C. J., & Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika,* 1957, 22, 133-145.

Glanzer, M., & Clark, W. H. Accuracy of perceptual recall: An analysis of organization. *J. verbal Learn. verbal Behav.,* 1963a, 1, 298-299.

Glanzer, M., & Clark, W. H. The verbal loop hypothesis: binary numbers. *J. verbal Learn. verbal Behav.,* 1963b, 2, 301-309.

Glanzer, M., & Clark, W. H. The verbal loop hypothesis: Conventional figures. *Amer. J. Psychol.,* 1964, 77, 621-626.

Keller, L. Run structure and the learning of periodic sequences. Unpublished doctoral dissertation, Indiana University, 1963.

Restle, F. *Psychology of judgment and choice.* New York: Wiley, 1961.

Restle, F. Run structure and probability learning: Disproof of Restle's model. *J. exp. Psychol.,* 1966, 72, 382-389.

Restle, F. Grammatical analysis of the prediction of binary events. *J. verbal Learn. verbal Behav.,* 1967, 6, 17-25.

Rose, R. M., & Vitz, P. C. The role of runs of events in probability learning. *J. exp. Psychol.,* 1966, 72, 751-760.

Royer, F. L. Sequential complexity and motor response rates. *J. exp. Psychol.,* 1967, 74, 199-202.

Royer, F. L., & Garner, W. R. Response uncertainty and perceptual difficulty of auditory temporal patterns. *Percept. & Psychophys.,* 1966, 1, 41-47.

Vitz, P. C., & Todd, T. C. A model of learning for simple repeating binary patterns. *J. exp. Psychol.,* 1967, 75, 108-117.

Notes

[1] This research was supported by grants from the National Institute of Mental Health (MH-11040-01 and 13083-01). I am indebted to Thomas C. Todd for his contributions to this research and to James J. Candalino for his assistance in the collection of the data.

[2] The number of runs in a pattern, N_r, is a still simpler measure and in the experiments described here almost an equivalent measure; N_r correlated .98 with H (run-span) in Experiment 1 and .97 with H (run-span) in Experiment 2. But N_r like $\log_2 N_r$ suggests little about a model of perceptual processing.

[3] I wish to thank Murray Glanzer for graciously providing his data.

[4] Interestingly, the correlation between MVL and accuracy is now −.94. This substantial increase over the −.81 originally reported by Glanzer and Clark is due to averaging the results of equivalent patterns.

(Accepted for publication December 1, 1967.)