Memory performance and subject-defined depth of processing

JOHN G. SEAMON and SUSAN VIROSTEK
Wesleyan University, Middletown, Connecticut 06457

This experiment found that subjects could order a series of classification questions in terms of their relative depth of processing. The subject-defined processing depth was used to predict performance in an incidental learning task which employed the same questions with different subjects. A significant rank correlation was obtained between question depth of processing and stimulus word free recall. These data provide an independent assessment of processing depth and memory performance and thereby satisfy a necessary, but not sufficient condition for the depth-of-processing hypothesis. Little success was achieved, however, in attempting to define the scheme relating processing depth and performance. While memory is related to encoding conditions, the search for the relational scheme continues.

Several years ago, Craik and Lockhart (1972) hypothesized that memory performance varies directly with the level or depth of stimulus encoding. Better performance in studies of incidental learning is typically found when stimuli are classified on the basis of meaning or conceptual relationships than visual or phonological features. The former tasks are felt to reflect deep semantic processing, while the latter are thought to entail surface-structural analysis.

Evidence for the difference between structural and semantic processing tasks appears to be strong. Hyde and Jenkins (1969) found associative clustering to be greater following a semantic than a structural orienting task. Elias and Perfetti (1973) showed false recognitions on acoustically or semantically related distractors to vary directly with encoding conditions. Finally, Seamon and Murray (1976) found that stimulus meaningfulness directly affected the recall and recognition of words analyzed on a semantic basis, but had no effect on items classified on a physical basis.

The above studies clearly show that structural and semantic orienting tasks entail different kinds of processing, but as Nelson (1977) effectively argues, different kinds of processing need not reflect different depths of processing. Differential depth of processing is an assumption based on different types of processing which produce different measures of memory performance.

The purpose of this research is to determine if a large number of classification tasks can be ordered in terms of processing depth to be used subsequently as a basis

for predicting performance on a free recall task. As stated by Nelson (1977, p. 165), "the derivation of the empirical ordering from the theoretical ordering must occur independently for memory performance and for depth of processing so that the examination of data can potentially falsify the principle." A nonsignificant rank correlation between the ordering of classification tasks and subsequent memory performance would constitute evidence contrary to the depth-of-processing hypothesis. Alternatively, a significant rank correlation would satisfy a necessary, but not sufficient condition for the Craik and Lockhart (1972) position.

PART 1

Method

Subjects. The subjects were 17 Wesleyan undergraduates who were enrolled in the senior author's memory seminar during the spring semester of 1977. At the time of the experiment, the subjects had read and discussed sensory memory, attention, and pattern recognition as topics in the course. The students were not yet familiar with the levels-of-processing approach to memory.

Materials. The 13 classification tasks listed by Nelson (1977, Table 3) and shown in Table 1 were used in this study. The 13 questions were printed separately on pieces of paper, shuffled, and placed in an envelope for each subject. The numbers associated with each question in Table 1 will be used for later data presentations in Part 2 and were not available to the subjects. Table 1 merely lists the questions with an associated numerical code.

Procedure. The first part of this experiment was conducted in a classroom with each of the 17 students seated at a separate desk. On a blackboard in front of the room, the word FROG was printed in uppercase letters. Each subject was given an envelope containing the 13 classification questions and asked to order the questions in terms of their relative processing depths from low to high on the surface of the desk. Depth of processing was defined as the amount of processing or degree of difficulty associated with each question. No mention of ease of learning was made to the subjects at any time. The emphasis was placed on the difficulty of different kinds of processing per se. Subjects were told that there was no "correct" sequence; they could

The present study was supported by a Wesleyan University Research Grant awarded to the first author. Requests for reprints should be sent to John G. Seamon, Department of Psychology, Wesleyan University, Middletown, Connecticut 06457. Appreciation is expressed to Thomas O. Nelson and Fergus I. M. Craik for comments on an earlier version. Susan Virostek is now at Columbia University.

Table 1
Nelson's (1977) Exemplary Classification Tasks

Classification Questions	
1	Is it printed in capital letters?
2	Does it contain the letter R?
3	Does it contain an r sound?
4	Does it rhyme with LOG?
5	Does it contain only one syllable?
6	Can you imagine it?
7	Is it a noun?
8	Is it green?
9	Can the letters G, O, R, F be rearranged to form a word?
10	Is it an animal?
11	Is it an associate of JUMP?
12	Can you use it in a sentence "A _____ fell down?"
13	Is it good?

Note—The above questions would be answered affirmatively for the stimulus FROG.

arrange the stimuli in any order and in as many piles (from 1 to 13) as they deemed necessary. No subject took more than 10 min to finish the task. Upon completion, the individual question slips were transferred to a larger sheet of paper to preserve their spatial arrangement for later data analysis.

Results

Of the 17 subjects who took part in this portion of the experiment, 8 ordered the classification questions individually into 13 piles. The other nine subjects viewed one or more of the questions as requiring the same amount of processing depth; these subjects averaged 9.33 piles with a range of 5-12.

The order of classification questions for each subject was ranked from 1 to 13 in terms of processing depth. In those instances where a subject had more than one question in a pile, the tied observations were assigned the average of the ranks they would have received had no ties occurred. The median ranks for each question were obtained over subjects and are shown in Figure 1. The medians are arranged in order of increasing magnitude from left to right. The numbers along the abscissa correspond to the question numbers in Table 1. Figure 1 clearly shows that the subjects thought the classification questions differed in processing depth and that there was strong agreement on the ordering of questions. This observation is supported by the finding of a significant coefficient of concordance. Kendall's W, which measures the degree of agreement among subject rankings, yielded a value of .60 when corrected for ties [$\chi^2(12) = 121.92, p < .001$]. This suggests that the subjects applied similar standards in ranking the 13 classification tasks (Siegel, 1956, p. 237).

In summary, the results of Part 1 indicate that the subjects had a common understanding of processing depth as it relates to the present stimulus (FROG) and classification questions. If relative processing depth is operationally defined by the median rank order of classification questions, the correlation between that order and the probability of free recall may be obtained.

PART 2

Method

Subjects. The subjects were 24 Wesleyan undergraduates who served as paid volunteers. None of the subjects in this portion of the experiment took part in the previous portion or in any other incidental learning experiment.

Stimulus materials. The stimuli consisted of 44 English nouns and four verbs derived from nouns obtained from the Paivio, Yuille, and Madigan (1968) word norms. Half of the words were rated high in imagery, concreteness, and meaningfulness (mean ratings of 6.14, 6.54, and 7.06, respectively) and half were rated significantly lower on these dimensions (mean ratings of 2.60, 2.27, and 4.65, respectively). Words that differed in terms of part of speech (noun-verb) and imagery (high-low) were necessary to have positive and negative instances for all of the classification questions. Initial letters of the words were evenly distributed over the whole range of the alphabet.

Procedure. Subjects made decisions about individual words on the basis of the classification questions used in Part 1. The subjects were told that the experiment measured the speed of decision making and that all questions should be answered as quickly and as accurately as possible to insure accurate response times. A surprise free recall test was given after all of the words were classified. With the exception of Question 9, all of the classification questions used before were repeated. Question 9 was not used because a failure to solve an anagram would result in no word to recall on the memory test.

Each of the 12 classification questions was used four times with a different word each time for a total of 48 trials. The words and classification questions were arranged such that for each question, a positive response was appropriate two times (e.g., Is it an animal?—whale) and a negative response was correct on the other two trials (e.g., Is it an animal?—thought). The 48 trials were divided into four blocks of 12 trials; each block was composed of one trial of each classification question. The associated positive or negative responses for each question were randomly ordered over the four blocks in each case. The subjects saw one of two different lists of 48 question-word pairs. Each subject received a counterbalanced order of blocks and a random order of questions within each block in each list.

Figure 1. The median rank for each classification question used in Part 1. The numbers along the abscissa correspond to the associated questions listed in Table 1. Relative depth of processing increases with an increase in the value along the ordinate.

On a given trial, an individual subject was shown a classification question through a narrow aperture in a viewing screen. The subject's task was to read the question orally into a microphone, which recorded all vocalizations on tape. Immediately after the question was read, the question was removed by raising the card and exposing the word to be classified, which was printed on the bottom of the card. A classification question and stimulus word were never viewed simultaneously. The subject's task was to classify the word in terms of the associated question by responding "yes" or "no" into the microphone. Subjects were encouraged to respond as quickly and as correctly as possible. After each response, the experimenter removed the card and recorded the question-word pair shown and type of decision made. No actual response time measures were obtained.

After the last trial the subject was given a short questionnaire that required the subject's signature, class, and the names and numbers of all psychology courses taken. The purpose of the questionnaire was to occupy the subject for 30 sec to act as a displacement for any items held in immediate memory. After the questionnaire was completed, the subject was asked for free recall of as many list words as possible on the reverse side of the questionnaire. Approximately 5 min was given for free recall.

Results

The percentage of words recalled for each classification question was obtained for each subject and then averaged over subjects. Figure 2 presents the mean percent recall and standard errors for Questions 1-8 and 10-13. Very clearly, the questions differed in terms of their associated recall probabilities; a general increase in recall was observed with an increase in subject-defined depth of processing. This relationship was supported by the finding of a significant Kendall rank-correlation coefficient [$\tau(\text{tau}) = .64$, $Z = 2.90$, $p < .002$] between the median ranks for the classification questions in Part 1 and their associated recall probabilities in Part 2. Virtually the same results were observed when the correlation was obtained in each of the four blocks of 12 trials for Serial Positions 1-12, 13-24, 25-36, and 37-38 (all $ps < .001$).

The present data demonstrate that a positive relationship exists between subject-defined depth of processing and recall performance. This observation would be strengthened by the finding that the classification order obtained in Part 1 is generally independent of the particular verbal stimulus employed (e.g., FROG) and the type of response made to each classification question (e.g., yes). To determine the effect of these variables upon the subject-generated classification order, additional groups of subjects were asked to construct orders based on concrete or abstract stimulus words and positive or negative responses to each question.

PART 3

Method

Subjects. The subjects were 40 Wesleyan undergraduates who served as paid volunteers. The subjects were enrolled in an

Figure 2. The mean percent recall and plus and minus one standard error of the mean for each classification question used in Part 2. The numbers along the abscissa correspond to the same questions along the abscissa of Figure 1 from Part 1.

introductory psychology course and had not yet been exposed to any concepts in human information processing.

Materials. With the exception of the anagram rule from Part 1 (Rule 9 in Table 1), each of the remaining 12 questions was used again in either the same or varied form. As before, the questions were typed separately on paper, shuffled, and put in an envelope for each subject. Each envelope contained one of two words printed in capital letters; half of the subjects received the concrete stimulus word FROG and half received the abstract stimulus FATE. Within each subset of 20 subjects, half again received questions that would correctly be answered affirmatively and half received questions that should be answered negatively. Subjects in the FROG/positive-response condition, for example, saw the questions listed in Table 1, while those in the FROG/negative-response condition saw the following questions: "Is it printed in lowercase letters? Does it contain the letter M? Does it contain an m sound? Does it rhyme with BOOK? Does it contain more than one syllable? Can you image it? Is it a verb? Is it blue? Is it a vegetable? Is it an associate of READ? Can you use it in a sentence 'A _____ was spelled? Is it bad?'" Comparable sets of positive and negative questions were employed for the stimulus FATE. Of special note is the question on imagery. Since there is no way to manipulate the response variable for this question, it was necessary to have a positive response to both question sets for FROG and a negative response to both question sets for FATE.

Procedure. The design of this portion of the experiment was a between-subjects 2 by 2 factorial that manipulated stimulus word by response type. The subjects were tested in two large group settings of approximately 20 subjects each. Assignment of 10 subjects to each of the four experimental conditions was done randomly within each testing setting. The procedures were identical to those employed in Part 1 with one exception. Since the subjects did not know what depth or amount of processing meant, they were told that processing depth could be defined as the amount of mental effort necessary to answer each question. Although previous research has shown mental effort to be unrelated to recall performance when effort is defined in terms of the number of necessary decisions for each stimulus in incidental learning (Walsh & Jenkins, 1973), the

Table 2
Median Ranks for Each Question as a Function
of Stimulus Word and Correct Response

Classification Question	Stimulus Word and Response			
	FROG		FATE	
	Yes	No	Yes	No
1 Capitals	1.75	3.50	2.00	5.00
2 Spelling	4.50	3.75	1.75	4.25
3 Sound	5.50	3.75	4.00	4.50
4 Rhyme	5.00	4.50	4.00	4.25
5 Syllables	6.00	4.75	4.00	4.75
6 Category	4.25	5.50	9.50	4.25
7 Color	6.25	4.25	9.75	7.00
8 Speech	7.00	8.00	6.00	7.75
9 Associate	9.00	11.00	7.50	8.25
10 Sentence	9.25	10.75	7.00	8.75
11 Image	7.75	9.25	10.50	11.25
12 Affect	12.00	10.50	12.00	11.75

Note—Ranks are based on a scale of 1-12. Response type was not varied within words for the imagery question; both responses were "yes" for FROG and "no" for FATE.

present experiment uses the term in the sense of the amount of attention (Craik & Lockhart, 1972) devoted to a given stimulus when only a single decision is being made.

Results

Median ranks were obtained as before for each question in each experimental condition and the results, which are shown in Table 2, were very similar to those obtained in Part 1, except that the imagery question was rated much higher in Part 3. In the main, stimulus word and response type had little effect upon the observed orders of classification questions. Consistent with this were the findings of (1) a significant coefficient of concordance (Kendall's $W = .57$, $\chi^2 = 24.94$, $p < .01$) over the four classification orders, which indicates that similar standards were used in each condition; and (2) a failure to find significant effects of response type for either stimulus word (both $ps > .05$) or effects of stimulus word for either response (both $ps > .05$) in pairwise comparisons based on the Wilcoxon Match-Pairs Signed-Ranks test. One notable exception within each order concerned the effect of stimulus word on the imagery question. Based on the results of the Mann-Whitney U test, subjects who received FATE as a stimulus rated the imagery question significantly higher for both positive and negative response decisions than those who received FROG (both $ps < .001$). Differences of this type, however, are not unexpected, as subjects should experience more difficulty in trying to imagine an abstract stimulus than a concrete item.

To determine the relationship between the classification orders of Part 3 and the recall performance of Part 2, the four orders were collapsed into a single order based on the median of the medians for each question. Since different stimulus words and responses were used in Part 2, an order based on these differences was obtained in Part 3. The order that is nominally shown

in the leftmost column of Table 2 is more highly correlated with probability of recall than that obtained from Part 1 [Kendall's $\tau(\text{tau}) = .70$, $Z = 3.16$, $p < .0008$]. This improvement would seem to be largely due to the present subjects' ranking the imagery question, which produced high recall, higher in the classification order than subjects had done previously. The reason for this difference may be partially due to the inclusion of an abstract stimulus in Part 3.

A reexamination of Figure 1 and, to a lesser extent, Table 2 suggests that subjects in Parts 1 and 3 may have viewed the 12 classification questions listed in Table 2 as consisting of only four different general categories: orthographic classifications (Questions 1 and 2), acoustic-phonological classifications (Questions 3, 4, and 5), classifications based on various types of category dimensions (Questions 6, 7, and 8), and object or meaning referent classifications (Questions 9, 10, 11, and 12). Table 3 shows mean recall over subjects as a function of this general scheme for each of the four blocks of serial positions. In each category, probability of recall changed little over Trial Blocks 1-3, but showed a sharp upturn on Block 4. Most importantly, the ordering of recall among the four general categories remained constant over each block of serial positions. This simple four-category classification, however, has serious shortcomings. Such a scheme would predict that questions that did not differ greatly in median rank should yield comparable results for recall. This is not always the case. Large differences were found in probability of recall in Part 2, for example, for the part-of-speech and color questions, yet little differences were observed in their median ranks in either Part 1 or Part 3.

GENERAL DISCUSSION

The present experiment showed that subjects can order a series of classification questions in terms of their relative processing depth on comparable bases. This subject-defined order-of-processing depth was then used as a basis for predicting relative free recall performance in an incidental learning experiment that employed the same classification questions with different subjects. Significant rank correlations were obtained, indicating that probability of recall increased as subject-defined

Table 3
Probability of Recall for Each Trial Block and Coding Condition

Coding Condition	Trial Block			
	1	2	3	4
Orthographic	0	8	6	16.5
Acoustic/Phonological	9.3	12.3	14	30.7
Categorical	29	16.7	27.7	41.7
Semantic Meaning	39.4	36.5	42.8	62.5

Note—Data are expressed as percentages. See text for questions associated with particular codes.

depth of processing increased. These data appear to satisfy Nelson's (1977) requirement for an independent assessment of memory performance and processing depth. The possibility of the data falsifying the principle was obviously present in this task.

The present data are clearly in accord with the depth-of-processing hypothesis (Craik & Lockhart, 1972). They do, however, provide only weak support for the theory as they satisfy a necessary, but not sufficient condition for its acceptance. Several problems are apparent. First, what is meant by depth of processing? Depth was operationally defined in this experiment in terms of the subject-generated ordering of classification questions. While there was strong agreement among the subjects, the basis for the agreement is not clear. Second, causality may not be inferred from correlational data. Variations in depth of processing may produce changes in memory performance either directly or indirectly through an unobserved factor. One possibility is that memory performance increases directly with the degree of semantic analysis given to a stimulus (Craik & Lockhart, 1972); this implies a direct relationship between processing depth and performance. Alternatively, depth of processing may be inversely related to the degree of automaticity of the pattern-recognition operations during stimulus encoding. Automaticity has been shown to be inversely related to memory performance (Kolers, 1975). Other than noting that encoding processes have important consequences for memory performance, no specific conclusions may yet be drawn.

Finally, assuming a relationship between depth of processing and memory performance as has been demonstrated in this experiment, what is the nature of the relationship between the predictor and criterion variables? Obviously, until the factor or factors responsible for the effect on memory performance have been isolated, no such relational scheme can be made apparent. The simple four-category coding order

examined in this experiment failed in several specific instances. Several questions thus become readily apparent. Should the order be expanded to include more than four discrete categories or redefined as a continuous variable? Might a multidimensional structure provide a better account of the data than a unidimensional scale? While no specific scheme has yet been found, the finding of a strong correlation in the present data suggests that the search should be continued.

REFERENCES

- CRAIK, F. I. M., & LOCKHART, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, 11, 671-684.
- ELIAS, C. S., & PERFETTI, C. A. Encoding task and recognition memory: The importance of semantic encoding. *Journal of Experimental Psychology*, 1973, 99, 151-156.
- HYDE, T. S., & JENKINS, J. J. Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, 1969, 82, 472-481.
- KOLERS, P. A. Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, 1, 689-701.
- NELSON, T. O. Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior*, 1977, 16, 151-171.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement*, 1968, 76(1, Part 2).
- SEAMON, J. G., & MURRAY, P. Depth of processing in recall and recognition memory: Differential effects of stimulus meaningfulness and serial position. *Journal of Experimental Psychology: Human Learning and Memory*, 1976, 2, 680-687.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- WALSH, D. A., & JENKINS, J. J. Effects of orienting tasks on free recall in incidental learning: "Difficulty," "effort," and "process" explanations. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 481-488.

(Received for publication October 25, 1977;
accepted January 11, 1978.)