The effects of recall and recognition test expectancies on the retention of prose

STEPHEN R. SCHMIDT
Purdue University, Lafayette, Indiana 47907

The hypothesis that people expecting recall and recognition employ different encoding processes was tested in two experiments using prose materials. In Experiment 1, unrelated sentences were used, and in Experiment 2, a short essay was used. The results indicated that a recall test expectancy led to greater sentence recall than a recognition test expectancy. No evidence was found to support the hypothesis that people expecting recall and recognition retained different types of information contained in sentences. In Experiment 2, the effects of test expectancy were analyzed as a function of the structural importance and rated comprehensibility of sentences. A main effect of test expectancy was found in sentence recall, replicating the results of Experiment 1. Also, people expecting recall tended to remember greater detail than did people expecting recognition. The results suggested that encoding processes vary as a function of test expectancy and that the appropriateness of encoding depends on the type of test received.

The relation between the type of test students expect and the study strategies they employ in preparation for that test has been investigated in many experiments (see Neely, Balota, & Schmidt, Note 1, for a recent review). Generally, when lists of words were the to-be-remembered material, people expecting recall tests did better on both recall and recognition than people expecting recognition tests (Balota & Neely, 1980; Neely & Balota, 1981). However, several experimenters have shown that the effects of test expectancy varied with the type of materials employed. For example, Wnek and Read (1980) found a larger effect of test expectancy for high- than for low-imagery words, and Balota and Neely (1980) found larger effects of test expectancy for high-frequency than for low-frequency words. Connor (1977, Experiment 2) found effects of test expectancy on the retention of categorized word lists when the words from the same category were blocked during presentation. With random presentation, the effects of test expectancy were reduced (but see Neely & Balota, 1981). Together, these results demonstrate that the effects of test expectancy obtained with one set of materials often will not generalize to other materials. Thus, if one is interested in generalizing the effects of test expectancy to settings outside the laboratory, it is important to determine directly the effects of

test expectancy on the retention of prose and other naturally occurring materials.

There are three lines of research concerning the effects of test expectancy on the processing of prose. First, experimenters have asked students to report the study strategies they employ in preparation for essay and multiple-choice tests. Students report that they are more likely to study general trends, to draw conclusions, and to organize related material in preparation for an essay test than for a multiple-choice test (Douglass & Tallmadge, 1934; Terry, 1933). In contrast, students studying for a multiple-choice test report trying to remember the specific wording of the text, to memorize details, and to underline important sentences (Douglass & Tallmadge, 1934; Meyer, 1934; Terry, 1933).

In the second line of research, the notes students take in preparation for different types of tests have been analyzed. While main effects of test expectancy on the total number of notes taken have not been observed (Hakstian, 1971; Rickards & Friedman, 1978; Weener, 1974), the content of notes does vary as a function of test expectancy. Investigators have found that students preparing for an essay test, when compared to students preparing for a multiple-choice test, were more likely to make an outline of the material, and their notes included a greater number of items high in structural importance (Meyer, 1936; Rickards & Friedman, 1978).

In the third line of research concerning the effects of test expectancy on prose processing, retention as a function of test expectancy has been directly measured. Unfortunately, consistent effects of test expectancy on prose retention have not been found. A recall test expectancy has been found to lead to better retention than a recognition test expectancy as measured by both recall and recognition (Meyer, 1934, 1936). However, in

The research reported herein was conducted as partial fulfillment of the PhD requirements of Purdue University. I would like to extend special thanks to Henry L. Roediger III (committee chairman), Harley A. Bernbach, James H. Neely, and Howard R. Ranken for their critical comments and moral support. Requests for reprints should be sent to S. R. Schmidt, Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

several more recent investigations, effects of test expectancy on prose recall or recognition have not been found (Kulhavy, Dyer, & Silver, 1975; Rickards & Friedman, 1978). In contrast, several experimenters have found that a multiple-choice test expectancy led to better multiple-choice performance than an essay test expectancy (Sax & Collet, 1968; Schulz, 1977).

Several factors may be responsible for the lack of consistent effects of test expectancy on prose retention. First, several of the "experiments" reported above were conducted by the instructors of introductory courses, with the information presented in the course serving as the to-be-remembered material (Hakstian, 1971; Sax & Collet, 1968; Schulz, 1977). Under these conditions, it is unclear to what extent several experimental factors were controlled. For example, did the essay and multiple-choice tests require the retention of the same information? How many times were the students tested on the same information? Given possible contact between students in different sections of the course, how successful was the manipulation of test expectancy?

The inconsistent effects of test expectancy on prose retention may also stem from possible differences in the types of foils used in the construction of the recognition and multiple-choice tests. Researchers employing prose materials generally have not described the relation between the targets and lures on the recognition tests they employed (Kulhavy et al., 1975; Meyer, 1934, 1936; Schulz, 1977). The effects of test expectancy on recognition performance may be partially determined by this relation between the targets and lures (Schmidt, Note 2).

Despite the lack of conclusive evidence concerning the main effect of test expectancy, test expectancy may have an effect on what people remember from a prose passage. Rickards and Friedman (1978) found that students were more likely to remember highly central material from a passage when they prepared for an essay test than when they prepared for a multiple-choice test. However, they did not report free recall data to support their conclusions, relying instead on cued recall of information contained in notes the students took in preparation for the memory test. Since test expectancy was found to have an effect on note taking, the measure of cued recall from the notes does not provide a clear picture of the effect of test expectancy on recall per se. A more appropriate measure of memory performance would be the conditional probability of recalling items of high and low importance given that the items were contained in the students' notes.

In the Rickards and Friedman (1978) study, and in most of the other investigations into the effects of test expectancy on prose retention (e.g., Kulhavy et al., 1975; Meyer, 1934, 1936; Rickards & Friedman, 1978; Sax & Collet, 1968; Schulz, 1977), study time has not been controlled. Other investigators have found that students expecting recall study the material for a longer period of time than students expecting recognition (Kulhavy et al., 1975). While this result is of some

interest in itself, study time must not be confounded with test expectancy if one wishes to infer qualitative differences in retention resulting from the manipulation of test expectancy.

In summary, previous research has not provided conclusive evidence concerning the effects of test expectancy on prose retention. However, some evidence suggests that students expecting an essay test emphasize general trends or higher order units, whereas students expecting multiple-choice tests emphasize detailed information. The experiments reported below were designed to test these hypotheses, as well as to provide a well controlled test for the main effect of test expectancy on prose retention.

EXPERIMENT 1

Experiment 1 was designed to test the hypothesis that on a test requiring memory for the exact wording or syntax of a sentence, a recognition expectancy would lead to better performance than a recall expectancy. In contrast, on tests for the retention of meaning, a recall expectancy should lead to better performance than a recognition expectancy. To test these hypotheses, recall protocols were scored for recall of any identifiable part of a sentence (memory for gist) and for recall of the exact wording of the sentence (memory for detail). In addition, three types of recognition tests were devised to obtain separate measures of retention of meaning and structure.

Method

Subjects. Three hundred and thirty-six subjects participated in this experiment as partial fulfillment of an introductory psychology course requirement.

Materials. One hundred and twenty sentences were selected as materials for Experiment 1. Half of the sentences were selected from classic novels, and the other half were selected from *Scientific American*. The 60 fiction and 60 nonfiction sentences were selected to fit the following criteria: self-contained in meaning, void of proper nouns, and not part of a direct quotation from a character. Further, the selection of related sentences was avoided. Following selection, half of the sentences of each type were randomly assigned to one list (List 1), and the other half were assigned to another (List 2).

Each of the sentences selected in the above manner was rewritten to form an alternate version of the sentence. The alternate version of each sentence retained the meaning of the sentence, but synonyms were substituted for some of the words and the syntax of the sentence was altered. For example, the original sentence "The biological value of protein depends on its content of essential amino acids" was rewritten to create the following sentence: "The presence of necessary amino acids determines the biological usefulness of protein." One member of each sentence pair was randomly selected to construct an alternate list, creating two forms of each list of sentences (Forms A and B). Thus, there were a total of four lists of sentences constructed (1A, 1B, 2A, 2B).

Three different types of recognition tests were constructed from the lists described above. On each recognition test form, target sentences were intermixed with distractor sentences, creating a "yes/no" recognition test. In the old/new recognition test, sentences from List 1 were randomly ordered with sentences from List 2. The A forms of each list were paired to construct one test form (1A/2A), and the B forms of the lists

were used to construct another test form (1B/2B). Each of these randomly ordered recognition tests was split in half, and the first and second halves were reversed to construct two more test forms. This allowed for partial counterbalancing of sentences with test positions. Thus, a total of four old/new recognition test forms were employed.

Four old/reworded test forms were constructed using the same logic as was used to construct the old/new forms. However, the old/reworded tests contained both forms of sentences from a given list. Thus, as an example of an old/reworded test form, sentences from List 1A were intermixed with sentences from List 1B. Correct responses on old/reworded tests required retention of the exact wording and/or syntax of the to-be-remembered sentences.

Four forms of reworded/new recognition tests were also constructed. These recognition test forms were identical to the old/new recognition test. However, these tests were paired with different acquisition sentences to form the reworded/new condition. For example, if subjects studied List 1A, they would then be given recognition test Form 1B/2B. Their task would be to select sentences similar in meaning to the acquisition sentences. Thus, in this example, the 1B sentences are the correct responses.

Design. The design was a 2 (type of test: recall vs. recognition) by 2 (type of expectancy: recall vs. recognition) by 2 (materials: fiction vs. nonfiction) by 4 (lists: 1A, 1B, 2A, 2B) factorial with repeated measures on the materials factor. In addition, within the recognition half of the experiment, there were three types of recognition tests (old/new, old/reworded, reworded/new). Ninety-six subjects were given the recall test, and each type of recognition test was given to 80 subjects.

Procedure. Subjects were tested in small groups. All the subjects in each group were given the same test expectancy and the same type of final retention test.

An initial set of instructions determined the test expectancy for each group of subjects. Subjects were led to expect a free recall test by the instructions that they would have to recall, without any aids or prompts, the acquisition sentences. Recognition expectancy was induced with the instructions that they would be asked to choose "the sentences you have read from a group of sentences which will include some sentences you have read and some sentences you will not have read." They were also told that the test would be similar to a multiple-choice test.

Each subject received a booklet containing an acquisition list of sentences. Each sentence was printed on a separate page of the booklet. A tone sounded every 10 sec, signaling the subject to turn the page and read the next sentence.

Following acquisition, response booklets were distributed. On the first page of the booklets, several addition problems were printed. The subjects were asked to solve as many of these problems as they could in 1 min. They were then given instructions describing the nature of their retention test. The recall subjects were told to try to remember as many of the sentences as they could and to try to remember the exact wording of the sentences. They were further instructed that if they could not remember a sentence in its entirety, they should write down as much of the sentence as they could remember. The recognition subjects were told about the nature of their recognition test,

including the relation between the distractors and the target items. The subjects were instructed to make a "same" or "different" response to each sentence and then to rate their confidence in that response on a scale of 1 to 5. Following this second set of instructions, the subjects were allowed to begin the retention test. Approximately 7 min elapsed between presentation and test. Recall and recognition groups were both given 20 min to complete their tests.

Results and Discussion

Recall. The recall protocols were scored for five dependent variables. Sentences were scored as recalled by a lenient criterion or a strict criterion. A sentence was scored as recalled by the lenient scoring if any identifiable part (e.g., a word or phrase unique to a sentence) was recalled. The strict scoring required recall of the subject, verb, and object of a sentence or synonyms for any of these sentence parts. Recalled sentences were also scored for the number of words recalled that matched words in the acquisition sentence. Each subject, then, was given a score for the number of sentences recalled—strict, the number of sentences recalled—lenient, and the number of words recalled. Analyses were also performed on the probability of recall—strict given that a sentence was recalled—lenient, and on words per sentence recalled—lenient. The last two measures provide estimates of memory for sentence detail given that some part of the sentence was recalled. A summary of the means for these five dependent variables is presented in Table 1.

An initial multivariate analysis of variance on the five dependent variables indicated a main effect of test expectancy [$F(5,90) = 4.13$] (except as otherwise noted, a p of .05 was required for all tests reported) and a main effect of materials [$F(5,90) = 17.74$]. The Expectancy by Materials interaction was not significant [$F(5,90) = 1.42$]. From these statistics, one can conclude that recall expectancy led to greater recall than recognition expectancy and that memory for fiction sentences (mean = 3.36 sentences—strict) exceeded memory for nonfiction sentences (mean = 2.91 sentences—strict). Since the fiction and nonfiction sentences may vary along a number of uncontrolled dimensions, no attempt will be made to interpret the main effect of materials. The absence of an interaction of test expectancy with materials allows one to generalize the effect of test expectancy to both fiction and nonfiction materials.

Univariate analyses indicated that the main effect of test expectancy was reliable for sentences recalled—strict

Table 1
Summary of Results From Experiment 1

Expectancy	Recall					Recognition (Probability)		
	Sentences— Strict	Sentences— Lenient	Words	p(S/L)	Words/ Sentence	Old/New	Old/ Reworded	Reworded/ New
Recall	3.63	5.44	34.64	.63	6.23	.94	.80	.92
Recognition	2.75	4.40	26.39	.60	5.57	.95	.80	.92

Note— $p(S/L)$ = probability of recalling a sentence by the strict criterion, given that it was recalled by the lenient criterion. Words/sentence = number of words recalled per recalled sentence.

[$F(1,94) = 5.39$, $MSe = 6.66$], for sentences recalled—lenient [$F(1,94) = 7.07$, $MSe = 7.36$], and for words recalled [$F(1,94) = 6.24$, $MSe = 523.63$]. However, there was not a reliable effect of test expectancy on within-sentence recall as measured by words per sentence recalled—lenient [$F(1,94) = 2.47$, $MSe = 8.44$] or as measured by the conditional probability of sentence recall—strict given sentence recall—lenient [$F(1,94) = .38$, $MSe = .11$].

These results indicated that recall expectancy leads to memory for a greater number of sentences than does recognition expectancy. As noted in the introduction, when lists of words are the to-be-remembered material, greater recall is usually found when people expect recall than when they expect recognition. The results reported above extend this effect to sentential materials. There was no evidence to suggest that a recognition expectancy leads to greater recall of sentence detail. Memory for the exact wording of the sentence did not significantly differ between test expectancy groups. In fact, the nonsignificant trends in the data suggest greater within-sentence recall by recall expectancy subjects than by recognition expectancy subjects.

Recognition. For each subject, an R score (Brown, 1976) was calculated. The R score is an estimate of the area under the memory-operating characteristic obtained from confidence ratings assigned to “old” and “new” recognition test items. The R scores were analyzed by a single univariate analysis of variance in which type of test (old/new, old/reworded, reworded/new), test expectancy, and materials were treated as factors. The mean R scores, as well as hit and false alarm rates, are presented in Table 2. A briefer summary of the recognition data appears in Table 1.

The type of recognition test the students received had a significant effect on recognition [$F(2,234) = 84.76$, $MSe = .0111$]. The major source of this effect was the difference in performance between the old/new test and the old/reworded test. Thus, the finding indicates that the nature of the distractors had an effect on recognition performance. A significant effect of materials was also observed [$F(1,234) = 140.28$, $MSe = .0021$], replicating the effect obtained in recall. Recognition was

better for fiction than for nonfiction sentences. This main effect of materials was compromised by an interaction between materials and type of recognition test [$F(2,234) = 31.02$, $MSe = .0021$].

The factor of most interest, test expectancy, did not have a significant main effect [$F(1,234) = .17$, $MSe = .0111$]. Also, the interaction of expectancy with recognition test did not approach significance [$F(1,234) = .86$, $MSe = .0021$]. The absence of significant effects of test expectancy was probably not due to a lack of statistical power. With respect to a hypothesized main effect of test expectancy of .05 (which would account for approximately 5.4% of the variance) and with α set at .05, the power of the statistical test was approximately .91. Thus, it is reasonably safe to conclude that there were no reliable effects of test expectancy on recognition.

In summary, a main effect of test expectancy was found on the number of sentences recalled. However, effects of test expectancy were not found on measures of within-sentence memory or recognition performance. Thus, the recognition data support the findings of Kulhavy et al. (1975) and Rickards and Friedman (1978) in that no effects of test expectancy were obtained. However, unlike these previous studies, effects of test expectancy on sentence recall were clearly obtained.

The overall pattern of results from Experiment 1 can be explained if it is assumed that subjects perform some additional process in preparation for recall but not in preparation for recognition. This additional process must affect recall performance but not recognition performance. In several theories of recognition and recall (Anderson & Bower, 1972; Kintsch, 1970), organizational processes are hypothesized to affect recall but not recognition performance. Thus, one explanation of the results of Experiment 1 is that subjects preparing for the recall test attempted to detect relations between sentences and to organize the to-be-remembered material to a greater extent than did subjects preparing for the recognition test. This conjecture is supported by students' reports that they are more likely to organize material in preparation for an essay test than in prepara-

Table 2
Summary of the Recognition Data From Experiment 1 as a Function of Test Expectancy and Materials

	Old/New Test		Old/Reworded Test		Reworded/New Test	
	Recall Expectancy	Recognition Expectancy	Recall Expectancy	Recognition Expectancy	Recall Expectancy	Recognition Expectancy
Fiction Materials						
Hits	.89	.92	.79	.75	.89	.91
False Alarms	.04	.05	.19	.18	.08	.11
R Score	.95	.96	.86	.84	.93	.94
Nonfiction Materials						
Hits	.86	.86	.69	.68	.84	.87
False Alarms	.06	.07	.29	.28	.14	.20
R Score	.93	.94	.75	.76	.90	.90

tion for a multiple-choice test (Douglass & Tallmadge, 1934). Similarly, Connor (1977) found a larger effect of organization on the memory of the subjects expecting recall than on the memory of subjects expecting recognition (but see Neely & Balota, 1981). Experiment 2 was designed to test the hypothesis that subjects expecting recall are more sensitive to the organization of material than are subjects expecting recognition.

EXPERIMENT 2

In Experiment 2, the memory performance of subjects expecting recall and recognition was compared as a function of two attributes of sentences embedded in a short essay. First, the role of each sentence in the overall structure of the essay was determined by asking subjects to rate the centrality of each sentence (Johnson, 1970). In terms of a hierarchical analysis of prose structure (e.g., Johnson, 1970; Thorndyke, 1977), increased emphasis on the organization of a prose passage should entail a "highlighting" of high-level, structurally important material. Thus, when subjects expect a recall test, they should remember a larger number of important sentences than subjects expecting recognition. This hypothesis has received weak support from studies on note taking (Meyer, 1936; Rickards & Friedman, 1978) and from the analysis of prose retention (see the earlier discussion of Rickards & Friedman, 1978).

Memory performance was also assessed as a function of ratings of sentence comprehensibility. Sentence comprehensibility should be less closely related to the organization of an essay than is structural importance. If subjects expecting recognition are less sensitive to the organization of prose, then the individual attributes of the sentences (e.g., comprehensibility, concreteness, etc.) should influence retention.

Method

Subjects. Three hundred and sixty subjects participated in the experiment as partial fulfillment of an introductory psychology course requirement. Two hundred subjects provided normative data on the memory materials, and the other 160 participated in the memory part of the experiment.

Materials. An essay titled "The Laws of Looking" (Argyle, 1978) was selected from *Human Nature*. Subsections of the essay were selected to form two self-contained essays. The first two paragraphs (seven sentences) and final paragraph (four sentences) of the original essay were used as an introduction and conclusion to both the constructed essays. The intervening 54 sentences, however, were not shared. In this manner, two essays were constructed (Essay 1 and Essay 2) that were on the same topic, written by the same author, but covered somewhat different material. The central, or critical, 54 sentences from each essay were then rewritten to form alternate versions. The sentences were rewritten in the same manner as were the sentences in Experiment 1. From each essay, two versions were constructed by replacing every other sentence by a reworded sentence. For example, in Essay 1A, the odd sentences appeared in their original form and the even sentences were reworded versions of the original form. In Essay 1B, the even sentences appeared in their original form and the odd sentences were

reworded versions. In a similar manner, Essay 2 was rewritten to create Essay 2A and Essay 2B. The first seven and last four sentences always appeared in the same form across all four essays.

For the acquisition phase of the experiment, sentences were printed individually on the pages of booklets. The booklet method of presentation gave the experimenter control over rate of sentence presentation. This provided continuity with the procedure employed in Experiment 1 and prevented a confounding of test expectancy with study time. Sentences from a given version of a given essay appeared in proper order in the booklets.

From the four essays, a four-alternative forced-choice recognition test on the critical 54 sentences was constructed. Both versions of each sentence from Essay 1 were randomly paired with two versions of a sentence from Essay 2 to produce a single test item. Thus, for each test item, the subjects were required to make a meaning discrimination to determine which pair of sentences was congruent with the essay they read. Each test item also required a structural discrimination to determine which version of the congruent sentences was actually read. Across test items, the order of sentences within the items was completely counterbalanced. In addition, test items were randomly ordered and divided into two halves, A and B. Two test-item orders were constructed, one with an A-B ordering of test halves and the other with a B-A ordering.

Design. Normative data were collected on all four essays. Separate groups of subjects rated the centrality and comprehensibility of each sentence. Also, different subjects rated each of the four essays.

In the memory part of the experiment, test expectancy (recall vs. recognition) was crossed with test received (recall vs. recognition). In addition, the four essays (1A, 1B, 2A, 2B) were factorially combined with the four experimental conditions.

Normative analysis of materials. The normative data were collected in two large groups. Within these groups, equal numbers of subjects received one of each of the four essays. One group of subjects was asked to rate the importance of each sentence in an essay. They were asked to read carefully the whole essay and then return to the beginning and rate the degree to which each sentence was important, or central, to the essay as a whole. A second group of subjects was asked to rate how easily each sentence in the essay could be understood. The group rating comprehensibility was not instructed to read the essay prior to rating the sentences. Both groups rated the sentences on a scale from 1 to 4, and each group was asked to place approximately equal numbers of sentences in each of the four rating categories.

Analysis of retention. The procedure for the memory part of the experiment was essentially identical to the procedure used in Experiment 1, with the following exceptions: The subjects were told that the sentences were part of an essay. Subjects that were given the recognition tests were told about the four-alternative forced-choice test and the nature of the three distractors.

Results and Discussion

Sentence analysis. Responses on the normative part of the experiment indicated that sentences within the essays varied considerably in terms of both rated centrality and comprehensibility. The mean ratings for sentences ranged from 1.36 to 3.52 for centrality and from 1.32 to 3.24 for comprehensibility. Ratings of sentence centrality and sentence comprehensibility were not significantly correlated [$r(52) = .08$]. For each rating system, the 54 critical sentences were divided into thirds based on their mean scores. These divisions were made separately for each of the four essays. Memory performance on each essay was then calculated as a

function of the sentence groups based on centrality and based on comprehensibility. Except where otherwise noted, memory performance was evaluated only on the critical 54 sentences.

Recall. Sentence recall was scored for the same five dependent variables analyzed in Experiment 1. Mean performance for these five dependent variables is summarized in Table 3. As in Experiment 1, the same pattern of results was found when the strict and lenient scoring procedures were employed. In the interest of clarity, the discussion of the recall data will focus on performance as measured by the number of sentences recalled by the strict criterion. The effects of test expectancy will be discussed first as a function of rated centrality and then as a function of rated comprehensibility.

Recall as a function of rated centrality is summarized in Figure 1. A multivariate analysis with centrality treated as a factor revealed significant effects of test expectancy [$F(5,152) = 3.58$] and centrality [$F(10,304) = 4.0$]. The Test Expectancy by Centrality interaction was not significant ($F < 1.0$). These results indicate that a recall expectancy led to greater recall than a recognition expectancy. Furthermore, sentence recall declined as sentence centrality increased. This finding may be a result of the specific materials used in this experiment. In particular, with the nonfiction materials employed, sentences that were central to the essay were typically abstract generalizations. Concrete examples, which were easy to remember, tended to be less central. Thus, with nonfiction materials, there is not always a positive relation between centrality and recall (see also Johnson, Note 3).

Separate univariate analyses of the five dependent variables indicated significant effects of centrality for all five measures [smallest $F(2,156) = 4.16$, $MSe = 3.45$]. However, the effect of test expectancy was not obtained for all the dependent variables (see Table 3). Significant effects of expectancy were found for sentences—strict [$F(1,78) = 4.81$, $MSe = 6.25$], sentences—lenient [$F(1,78) = 3.32$, $MSe = 9.51$], and for words recalled [$F(1,78) = 5.00$, $MSe = 391.27$]. Within-sentence recall was not found to vary with test expectancy as measured by words per sentence recalled—lenient or as measured by the probability of recall—strict given recall—lenient ($F_s < 1.0$).

When rated comprehensibility was treated as a factor,

Figure 1. Recall and recognition performance in Experiment 2 as a function of sentence centrality.

a multivariate analysis indicated significant effects of test expectancy [$F(5,152) = 4.25$] and rated comprehensibility [$F(10,304) = 7.75$]. The number of sentences recalled (strict criterion) increased from 1.80 for low-comprehensibility sentences to 3.30 for high-comprehensibility sentences. The Test Expectancy by Comprehensibility interaction was not significant [$F(10,304) = 1.51$].

The recall data from Experiment 2 were very similar to the results of Experiment 1. Test expectancy was found to have an effect on the number of sentences recalled, but no effect of expectancy was found on within-sentence recall. Nonetheless, the data failed to provide evidence for qualitative differences in sentence recall in terms of sentence centrality or comprehensibility. Several additional analyses were performed in an attempt to measure qualitative differences in the recall performance of recall and recognition expectancy

Table 3
Summary of Results From Experiment 2

Expectancy	Recall					Recognition (Probability)		
	Sentences—Strict	Sentences—Lenient	Words	p(S/L)	Words/Sentence	Correct	Structural Errors	Semantic Errors
Recall	8.93	13.51	66.68	.65	4.60	.63	.30	.07
Recognition	6.80	10.98	49.55	.63	4.45	.62	.30	.08

Note— $p(S/L)$ = probability of recalling a sentence by the strict criterion, given that it was recalled by the lenient criterion. Words/sentence = number of words recalled per recalled sentence.

groups. In each of these analyses, performance was measured on all 65 sentences.

The first analysis was designed to test the general hypothesis that recall and recognition expectancy groups remembered different sentences from the essays. An analysis of variance was performed in which test expectancy, essay, and sentences (within essays) were treated as factors. Unfortunately, the interaction between test expectancy and sentence did not approach statistical significance [$F(256,4608) = .90$, $MSe = .14$]. While the test expectancy groups may not have recalled different sentences, the pattern of sentence recall may have differed between the groups. Two analyses were performed to test this hypothesis. First, sequential recall of sentences adjacent in the essays was analyzed. No reliable differences were found in sequential recall as a function of test expectancy [$t(78) = .88$]. Second, recall protocols were scored for the number of paragraphs recalled and the number of sentences recalled per recalled paragraph. Paragraph recall for recall expectancy subjects (mean = 9.00) was greater than paragraph recall for recognition expectancy subjects [mean = 8.08; $t(78) = 1.67$, one-tailed test, $p < .05$]. Recall expectancy subjects also recalled more sentences per paragraph (mean = 1.72) than did recognition expectancy subjects (mean = 1.53; $t(78) = 2.21$, $p < .05$). These results provide additional evidence for greater recall by subjects expecting recall than by subjects expecting recognition, but they do not provide evidence for qualitative differences in retention as a function of test expectancy.

Recognition. Each subject's response to each recognition test item was classified into one of three categories. If the subject selected the exact sentence he/she saw during input, then he/she was given credit for a correct response. Choice of a reworded version of an input sentence constituted a structural error. Choice of either of the remaining two sentences indicated a semantic error. Table 3 includes a summary of recognition performance for these three dependent variables.

Recognition performance was evaluated separately for the classifications of sentences in terms of centrality and comprehensibility. Recognition as a function of centrality is summarized in Figure 1. With centrality treated as the sentence factor, a multivariate analysis revealed no effect of test expectancy [$F(2,155) = .86$]. There was a significant effect of sentence centrality [$F(4,310) = 2.89$], but it did not interact with test expectancy [$F(4,310) = 1.65$]. In univariate analyses of the three dependent variables, there was evidence for a Test Expectancy by Sentence Centrality interaction. With the probability correct as the dependent variable, the Expectancy by Centrality interaction was marginally significant [$F(2,156) = 2.67$, $MSe = .0136$, $p < .07$]. The interaction was significant with the probability of structural error as a dependent variable [$F(2,156) = 3.11$, $MSe = .0118$]. The interaction appears to be due to a slightly larger effect of centrality

on performance following a recall expectancy than following a recognition expectancy (see Figure 1). Recall expectancy subjects were less likely to make structural errors on low-centrality sentences than were recognition expectancy subjects [$F(1,231) = 3.66$, MSe (pooled) = .0129, $p < .06$]. This result is exactly opposite from the anticipated results. It was hypothesized that recall expectancy subjects would concentrate less on detailed information and would perform better on high-centrality sentences than would recognition expectancy subjects. However, the obtained interaction between centrality and test expectancy is supported by several other results. For example, there was also a trend in the recall data toward greater retention of low-centrality sentences by subjects expecting recall than by subjects expecting recognition (see Figure 1). Also, in both Experiments 1 and 2, there was a tendency toward better performance on measures of within-sentence memory by subjects expecting recall than by subjects expecting recognition (see Tables 1 and 3). Thus, contrary to expectation, subjects expecting recall remembered greater sentence detail and a greater number of low-importance sentences than did subjects expecting recognition.

The recognition data were also analyzed as a function of sentence comprehensibility. Once again, no effect of test expectancy was found in a multivariate analysis ($F < 1.0$). Sentence comprehensibility had a reliable effect on recognition [$F(4,310) = 3.34$]. Univariate analyses revealed that this effect was limited to an increase in semantic errors from .06 to .09 as sentence comprehensibility decreased [$F(2,156) = 6.45$, $MSe = .0030$]. Comprehensibility did not interact with test expectancy in any of the analyses [largest $F(2,156) = 1.84$, $MSe = .0030$].

GENERAL DISCUSSION

The experiments reported above were designed to explore the possibility that students employ different types of study strategies to prepare for different types of test. The results of Experiments 1 and 2 consistently demonstrated a recall expectancy superiority on recall tests. However, the effects of test expectancy were found only on the number of sentences recalled, not on measures of within-sentence recall or on the recognition of sentence meaning or structure. These results seemed consonant with the hypothesis that subjects expecting recall organized the sentences to a greater extent than did subjects expecting recognition. The design of Experiment 2 permitted a direct test of this hypothesis. While there was evidence for the retention of more detailed information by subjects expecting recall than by subjects expecting recognition, the results did not support the hypothesis that subjects expecting recall organize the to-be-remembered material to a greater extent.

While the results of studies employing lists of words need not generalize to prose materials, the results

reported above are consistent with some recent research. For example, Neely and Balota (1981) studied the effects of semantic relations among words on retention as a function of test expectancy. They found that the effects of test expectancy and word relatedness were additive, and no differences in clustering were observed as a function of test expectancy. Wnek and Read (1980) found effects of test expectancy on the recall of word lists, but they did not find a main effect of test expectancy in recognition (although test expectancy did interact with imagery value in recognition). Nonetheless, there is ample evidence for qualitative differences in retention of word lists as a function of test expectancy (Balota & Neely, 1980; Connor, 1977; Tversky, 1973; Wnek & Read, 1980). The results reported above suggest that such qualitative differences in what people remember from prose materials as a function of test expectancy are generally lacking.

While the results of Experiments 1 and 2 are consistent with past research, it is difficult to provide a theoretical explanation for these results. The difficulty arises from the interaction of test expectancy with type of test (e.g., test expectancy affected recall but not recognition) in the absence of qualitative differences in what people remember as a function of test expectancy. Let us consider several explanations for this pattern of results. First, perhaps subjects expecting a recall test did organize the material to a greater extent than subjects expecting a recognition test. Perhaps a more exact measure of the structure of the essays would have yielded a pattern of results in which recall was a function of essay structure and structure interacted with test expectancy. This possibility seems remote, given the variety of analyses that failed to indicate differences in the types of sentences recalled as a function of test expectancy.

A second possible explanation for the results is that recall and recognition expectancy groups may not have been equally motivated to perform well when given a recall test (Neely et al., Note 1). Subjects given an unexpected recall test may do worse on that test than subjects expecting recall because they feel "double-crossed." However, in addition to predicting better recall by recall expectancy subjects, the double-crossing hypothesis predicts better recognition performance by subjects expecting a recognition test. In Experiments 1 and 2, no evidence for such an effect was found, seriously damaging a simple motivational hypothesis.

A third potential explanation of the results of Experiments 1 and 2 is based on the notion of "test-appropriate processing" (Morris, Bransford, & Franks, 1977; Stein, 1978). Subjects who prepared for a recall test seem to have encoded the material in a manner that was appropriate for retrieval in a recall test. Subjects expecting recognition, when compared to subjects expecting recall, apparently employed encoding processes that were less appropriate for a recall test. However, the encoding

processes employed by subjects expecting recall and those expecting recognition were apparently equally appropriate for the recognition test. The exact difference between the encoding processes employed by subjects expecting recall and subjects expecting recognition is not yet known. From the evidence reviewed above, one may conclude that differences in encoding resulting from the manipulation of test expectancy do not include differences in the encoding of either within-sentence relations or hierarchical between-sentence relations. Perhaps subjects expecting recall encode a greater number of context-item relations than do subjects expecting recognition. Context-item relations would be important to retrieval processes involved in recall, whereas item-context relations would be important for correct performance on recognition tests (e.g., Lockhart, Craik, & Jacoby, 1976).

In addition to the theoretical importance of the effects of test expectancy, the results reviewed above have several interesting educational implications. First, the research suggests that students may learn more when preparing for a short-answer or essay test than preparing for a multiple-choice test. This implication is based on the superior recall of subjects expecting recall when compared to subjects expecting recognition. Second, tests requiring recall may be more sensitive to differences between students than are multiple-choice tests. This inference is based on the failure to detect effects of test expectancy in recognition, effects that were observed in recall.¹ Taken together, these two factors suggest that whenever practical considerations permit a choice of test format, educators should employ some form of a recall test.

In summary, subjects expecting a recall test recalled more words, sentences, and paragraphs than did subjects expecting recognition. However, effects of test expectancy were not found in recognition. There was no evidence to suggest that subjects expecting recall emphasized general trends at the expense of memory for detailed information. Rather, the results suggested that retention of within-sentence information and low-centrality sentences was greater when subjects were expecting a recall test than when they were expecting a recognition test. The main effects of test expectancy were most easily interpreted within the framework of test-appropriate processing. Encoding processes apparently varied as a function of test expectancy, and the appropriateness of the encoding processes varied as a function of type of test.

REFERENCE NOTES

1. Neely, J. H., Balota, D. A., & Schmidt, S. R. *Test expectancy effects in recall and recognition: A methodological, empirical, and theoretical analysis*. Unpublished manuscript, 1982.
2. Schmidt, S. R. *The effects of test expectancy as a function of the type of distractors used in practice recognition tests*. Manuscript in preparation, 1982.

3. Johnson, R. E. *Dimensions of textual prose and remembering*. Paper presented at the meeting of the American Educational Research Association, New Orleans, 1973.

REFERENCES

- ANDERSON, J. R., & BOWER, G. H. Recognition and retrieval processes in free recall. *Psychological Review*, 1972, 79, 97-123.
- ARGYLE, M. The laws of looking. *Human Nature*, 1978, 1, 32.
- BALOTA, D. A., & NEELY, J. H. Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, 1980, 6, 576-587.
- BROWN, J. An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition*. New York: Wiley, 1976.
- CONNOR, J. M. Effects of organization and expectancy on recall and recognition. *Memory & Cognition*, 1977, 5, 315-318.
- DOUGLASS, H. R., & TALLMADGE, M. How university students prepare for new types of examinations. *School and Society*, 1934, pp. 318-320.
- HAKSTIAN, A. R. The effects of type of examination anticipated on test preparation and performance. *Journal of Educational Research*, 1971, 64, 319-324.
- JOHNSON, R. E. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, 1970, 9, 12-20.
- KINTSCH, W. Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory*. New York: Academic Press, 1970.
- KULHAVEY, R. W., DYER, J. W., & SILVER, L. The effects of notetaking and test-expectancy on the learning of text material. *Journal of Educational Research*, 1975, 68, 363-365.
- LOCKHART, R. S., CRAIK, F. I. M., & JACOBY, L. Depth of processing, recognition and recall. In J. Brown (Ed.), *Recall and recognition*. New York: Wiley, 1976.
- MEYER, G. An experimental study of the old and new types of examination: I. The effects of examination set on memory. *Journal of Educational Psychology*, 1934, 25, 641-661.
- MEYER, G. The effects of recall and recognition of the examination set in classroom situations. *Journal of Educational Psychology*, 1936, 27, 81-99.
- MORRIS, C. G., BRANSFORD, J. D., & FRANKS, J. J. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 1977, 16, 519-533.
- NEELY, J. H., & BALOTA, D. A. Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition*, 1981, 9, 283-300.
- RICKARDS, J. P., & FRIEDMAN, F. The encoding versus the external storage hypothesis in notetaking. *Contemporary Educational Psychology*, 1978, 3, 136-143.
- SAX, G., & COLLET, L. S. An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement*, 1968, 5, 169-173.
- SCHULZ, R. A. Discrete-point versus simulated communication testing in foreign languages. *Modern Language Journal*, 1977, 61, 94-101.
- STEIN, B. S. Depth of processing reexamined: The effects of precision of encoding and test appropriateness. *Journal of Verbal Learning and Verbal Behavior*, 1978, 17, 165-174.
- TERRY, P. W. How students review for objective and essay tests. *Elementary School Journal*, 1933, 33, 592-603.
- THORNDYKE, P. W. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, 1977, 9, 77-110.
- TVERSKY, B. Encoding processes in recognition and recall. *Cognitive Psychology*, 1973, 5, 275-287.
- WEENER, P. Notetaking and student verbalization as instrumental learning activities. *Instructional Science*, 1974, 3, 51-74.
- WNEK, I., & READ, J. D. Recall and recognition encoding differences for low- and high-frequency words. *Perceptual and Motor Skills*, 1980, 50, 391-394.

NOTE

1. It is assumed here that the recall format (essay or short-answer) test samples the same amount of knowledge as does the recognition format (multiple-choice) test. While this was the case in the experiments described above, in a classroom setting this may be difficult to achieve.

(Received for publication May 19, 1982;
revision accepted September 17, 1982.)