Memory & Cognition
1981, Vol. 9 (3), 283-300

Test-expectancy and semantic-organization effects in recall and recognition

JAMES H. NEELY
Purdue University, West Lafayette, Indiana 47907
and
DAVID A. BALOTA
University of South Carolina, Columbia, South Carolina 29208

Two experiments examined whether people expecting recall are, compared with people expecting recognition, more likely to form associations between semantically related words in a list of to-be-remembered words. People were induced to expect either a recall or a recognition test on a critical list that included three conditions of semantic organization. Words in the unrelated (U) condition were semantically unrelated to all other words on the list, whereas words in the two related conditions were semantically related to one other list word. In the related-spaced (R-S) condition, the two related words appeared in input positions separated by 5-11 other items, whereas in the related-massed (R-M) condition, they appeared in adjacent input positions. Different groups received either an expected or unexpected recall (Experiment 1) or recognition (Experiment 2) test on the critical list. In both recall and recognition, (1) people expecting recall did better than those expecting recognition, (2) memory was worst for U words, next best for R-S words, and best for R-M words, and (3) the test-expectancy and semantic-organization effects were additive. A standardized (z-score) measure of category dependency in memory indicated that (1) people expecting recall were not more likely than those expecting recognition to form interitem associations between the related words and (2) recognition was category dependent, but less so than recall. Within the framework of Anderson and Bower's (1972, 1974) theory, these data indicate that, compared with people expecting recognition, those expecting recall are not more likely to form interitem associations by tagging more pathways connecting semantically related nodes but, rather, are more likely to tag the nodes themselves. The implications that semantic-organization effects in recognition have for the Anderson-Bower theory were also discussed.

During the past decade, episodic memory researchers have found that certain experimental variables have differential effects upon recall and recognition performance (see Brown, 1976). Proponents of generate-recognize theories of recall originally accounted for the differential effects that experimental manipulations have upon recall and recognition performance by arguing that these manipulations differentially affect only retrieval processes (e.g., Kintsch, 1970). However, it soon became apparent that an adequate theory of recall and recognition needed to give encoding processes their due as well. This shift in emphasis to encoding processes occurred primarily because (1) recognition performance depends on the similarity of the semantic contexts in which an item is embedded during input and at the time of the recognition test (e.g., Light & Carter-Sobell, 1970) and (2) an item not recognized in isolation can be recalled in the presence of a contextual cue present during input (e.g., Tulving & Thomson, 1973). To handle these contextual effects, Anderson (1972) and Anderson and Bower (1972, 1973, 1974) added to generate-recognize theory a relatively explicit set of theoretical assumptions concerning how encoding processes influence recall and recognition performance.

In the present experiments, we were interested in determining whether people expecting a recall test encode the to-be-remembered materials differently from people expecting a recognition test and, if so, whether these encoding differences represent a differential utilization of those encoding processes that should, according to the Anderson-Bower theory, differentially benefit recall and recognition performance. To address these issues, we must first examine in some detail the Anderson-Bower theory's assumptions about how encoding processes affect recall and recognition performance.

Because both authors contributed equally, the order of authorship counterbalances the randomly determined order of authorship on another article (Balota & Neely, 1980) to which the authors also made equal contributions. This research was conducted as D. A. Balota's master's thesis at the University of South Carolina under J. H. Neely's supervision. Portions of this research were reported at the 1978 meeting of the Southeastern Psychological Association in Atlanta. We thank F. Gregory Ashby for his assistance in deriving the formulas in the appendix, Randall W. Engle and Henry L. Roediger III, for their helpful comments on earlier drafts of this article, and Richard Schweickert for statistical advice. Requests for reprints should be addressed to James H. Neely, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

0090-502X/81/0283-18$02.05/0

Anderson and Bower's Distinction Between Node Tagging and Pathway Tagging

According to the Anderson-Bower theory, during encoding in list learning experiments, both the memory nodes corresponding to a particular meaning of each to-be-remembered word and the pathways that connect these nodes are separately associated with contextual list-marker elements via paired associate-like processes called, respectively, node tagging and pathway tagging. Anderson (1972) suggests that a memory node can be tagged with list-marker elements only while it resides in a limited-capacity short-term store (STS). When a to-be-remembered word is presented, a node corresponding to one of its meanings assumes residence in the STS. Nodes associatively and/or semantically related to that node are also entered into the STS to determine if they can be recognized as having represented one of the prior words in the current to-be-remembered list. If any of these related nodes are so recognized, they remain in STS so that the pathways that connect them to the node corresponding to the current to-be-remembered word can also be tagged with list-marker elements.

Presumably, a word presented in a recognition test activates a node corresponding to one of its meanings. If this node currently resides in the STS, the word will always be correctly recognized; if it does not, a more elaborate recognition process occurs in which list-marker elements associated with the activated node are retrieved and "counted"; as this count increases, recognition becomes more likely. Context effects in recognition occur during the operation of this more elaborate recognition process. The more similar are the semantic contexts surrounding the to-be-remembered item at input and at the time of the recognition test, the more likely it is that the meaning node interrogated for the presence of associated list-marker elements in the recognition test will be the same meaning node as the one tagged with list-marker elements at input (but see Tulving & Watkins, 1977, for a critique of this analysis).

In recall, an extra process of generating items, namely, searching for and finding nodes, is necessary. This search originates in the nodes in STS and the nodes in an ENTRYSET (see Anderson, 1972, for details) and follows those pathways that were tagged during input with list-marker elements. Once a node is accessed during this generation process, it is submitted to the same recognition test that occurs when a node is accessed by the presentation of a word in a recognition test.

The encoding assumptions made by the Anderson-Bower theory not only permit an accommodation of context effects in recognition but permit, as well, a rich set of deductions from the generate-recognize model. For example, since node tagging and pathway tagging are partly reciprocal and differentially important in mediating recall and recognition performance, one should be able to affect recall and recognition performance differentially by inducing a person to emphasize differentially node tagging or pathway tagging. Support for this inference comes from two sources. The first source of support is that intentional learning instructions can produce better recall, but poorer recognition, performance than can incidental learning instructions (Eagle & Leiter, 1964), and an incidental learning task involving categorization of the to-be-remembered words produces better recall, but poorer recognition, performance on a list of categorically related words than does an incidental task involving imagery instructions (Griffith, 1975). To account for these effects, one need only argue (cf. Anderson, 1972, p. 370) that people receiving the intentional (or categorization) instructions were more likely to produce pathway tags, which play a crucial role in facilitating recall, than they were to produce node tags. Likewise, it can be argued that those receiving the incidental (or imagery) instructions were more likely to produce node tags, on which recognition performance is based, than they were to produce pathway tags. A second source of support for recall and recognition being differentially affected by certain encoding processes is that increased maintenance-rehearsal duration leads to better recognition performance but has little or no effect on recall performance (e.g., Glenberg, Smith, & Green, 1977; Woodward, Bjork, & Jongeward, 1973; but see also Maki & Schuler, 1980). To account for this, one can assume that maintenance rehearsal on a particular item increases only the number of node tags associated with that item's node, thus facilitating its recognition, and does not increase the number of pathway tags that, if increased, would facilitate the search process in recall (cf. Crowder, 1976, pp. 386-387).

An Account of Test-Expectancy-Induced Encoding Differences

What does the Anderson-Bower theory predict about the effects of test-expectancy-induced encoding differences on recall and recognition performance? In line with the theoretical analyses given in the preceding paragraph, one might conclude that people expecting recall should do more pathway tagging than those expecting recognition. Unfortunately, such a conclusion receives only indirect support from the test-expectancy literature. This support is in the form of the finding that those expecting recall do better in recall than those expecting recognition (e.g., Hall, Grossman, & Elwood, 1976; Poltrock & MacLeod, 1977, Experiment 1; Miller, Maisto, Fleming, & Rosinsky, Note 1; also, see Neely, Balota, & Schmidt, Note 2, for an extensive review of such effects). However, this is only indirect support because this finding would also occur if those expecting recall facilitated the recognition stage of their recall by doing more node tagging than those expecting recognition.

To determine whether a recall expectancy produces more pathway tagging than a recognition expectancy,

one must devise a more direct test for determining when a greater amount of pathway tagging has occurred. The rationale for such a test can be found in the Anderson-Bower theory's account of the finding that recall performance is better for lists containing semantically related words than for lists containing semantically unrelated words, whereas recognition performance is equivalent for these two kinds of lists (e.g., Bruce & Fagan, 1970). The account proceeds as follows: The greater the semantic organization among the to-be-remembered words is, the more likely it is that the pathways connecting their nodes will have been tagged during encoding (see earlier discussion on pathway tagging) and, hence, the more likely it is that these nodes will be generated during the first stage of recall. However, once generated, these nodes are submitted to a recognition test, the performance on which, like on the experimenter-generated recognition test, is unaffected by the semantic relationships among the to-be-remembered words. If this analysis is correct and if people expecting recall do more pathway tagging than those expecting recognition, it is predicted that those expecting recall should, relative to those expecting recognition, do progressively better in recall as the semantic organization among the to-be-remembered words increases. The most straightforward way to test this prediction would be to examine the joint effects of test expectancy and semantic organization on recall performance. The results of one such examination have been reported by Connor (1977).

Connor (1977) induced recall or recognition expectancies by administering three practice lists involving either only recall or only recognition tests, respectively. These practice lists had the same level of semantic organization as the critical list, which was tested by either an expected or unexpected recall or recognition test. In Experiment 1, lists consisted of either 8 four-word categories (random-categorized lists) or 32 one-word categories (noncategorized lists). In Experiment 2, the 8 four-word categories were presented either randomly (random-categorized lists) or blocked by category (blocked-categorized lists). In Experiment 1, recall was better for those expecting recall than for those expecting recognition and better to the same degree for the noncategorized and random-categorized lists. This additivity of the test-expectancy and semantic-organization effects runs counter to the interpretation that those expecting recall do more pathway tagging than those expecting recognition. On the other hand, the recall results of Experiment 2 support the additional pathway-tagging interpretation, since the facilitatory effect of the recall expectancy was large for the "organized" blocked-categorized lists and nonexistent for the "unorganized" random-categorized lists.

Unfortunately, there are three problems with Connor's (1977) study that becloud the interpretation of her results. Perhaps the most obvious problem is that the results of Experiment 2 indicated that test expectancy and semantic organization have interactive effects in recall, whereas the results of Experiment 1 indicated that these two variables have additive effects in recall. A second, and more profound, problem is that a test-expectancy effect was obtained for the random-categorized lists in Experiment 1 but not in Experiment 2. This casts doubt on the reliability of her results. But even if the results had been reliable and consistent, they would still be difficult to interpret because of a third methodological problem. As has been noted by Hall et al. (1976), when test expectancies are induced by practice lists tested by only one type of test, test-expectancy effects are necessarily confounded with retrieval practice and proactive interference (PI) effects. Thus, retrieval practice effects were confounded with test-expectancy effects in Connor's experiments because subjects given the expected recall test had been given practice on recall retrieval operations, whereas those given the unexpected recall test had not. Because of this, the obtained test-expectancy effects could have been due to differences in test-specific retrieval practice rather than to test-expectancy-induced encoding differences. In addition to being confounded with retrieval practice effects, Connor's test-expectancy effects were also confounded with PI effects, because those expecting recognition were exposed to the lures appearing in the practice recognition tests, whereas those expecting recall were not. Perhaps these lures produced additional PI with performance on the critical list recall test. If so, those expecting recall may have done better than those expecting recognition because they suffered from less PI rather than because they did more pathway tagging. These confoundings are even more problematic when semantic organization is a between-subjects variable and the practice lists have the same organization as the critical list, as was the case in Connor's experiments. For example, not only were those expecting recall given more practice on recall-specific retrieval operations, but they were also given more practice on retrieval operations specific to the recall of "organized" or "unorganized" lists. (In fairness to Connor, it should be noted that it is a relatively standard practice to confound test-expectancy effects with retrieval practice and PI effects. See Neely et al., Note 2, for a review.)

EXPERIMENT 1

Experiment 1 was conducted to determine whether, compared with people expecting recognition, people expecting recall are more likely to form interitem associations between semantically related words in a to-be-remembered list. Like Connor's (1977) experiments, Experiment 1 examined the joint effects of test expectancy and semantic organization in recall. However, there are two important design features of Experiment 1 that contrast with those employed by Connor. First, semantic organization was a within- rather than a

between-list variable. Words in the unrelated (U) condition were semantically unrelated to all other list words, whereas words in the two related conditions were semantically related to one other list word. In the related-spaced (R-S) condition, the two related words appeared in input positions separated by 5-11 other items, whereas in the related-massed (R-M) condition, they appeared in adjacent input positions. Of course, the U, R-S, and R-M conditions correspond, respectively, to Connor's noncategorized, random-categorized, and blocked-categorized lists. This within-list manipulation of semantic organization isolates test-expectancy effects for semantically related and unrelated words from other effects caused by differences in encoding or retrieval strategies that might be produced by differences in (1) the words each person must remember and/or (2) the overall list organization. Organization-dependent test-expectancy effects were isolated from these other effects in Experiment 1 because the semantically related and unrelated words were embedded in a list consisting of exactly the same to-be-remembered words with exactly the same overall list organization. The isolation of test-expectancy effects from list-organization effects is important because test-expectancy-induced differences in pathway tagging should, according to the Anderson-Bower theory, occur at the level of specific semantic associations rather than at the level of the overall organization of the list.

A second feature of the design of Experiment 1 was that it examined "pure" test-expectancy effects in the absence of the confounded effects of retrieval practice and PI. This was accomplished in a balanced-practice (BP) condition, in which those expecting recall and those expecting recognition received exactly the same practice lists and exactly the same practice test sequences (i.e., three practice lists tested by recall and three tested by recognition). Each practice list in the BP condition was preceded by a prelist cue that validly indicated the type of test the person would receive on that list. Such a prelist cue was then used to induce a recall or recognition expectancy for the critical list. To make our test-expectancy manipulation analogous to that of previous research, we also included an unbalanced-practice (UP) condition, in which test-expectancy effects were confounded with retrieval practice and PI effects.

The results of Experiment 1 should help elucidate the mechanisms that mediate test-expectancy effects in free recall. First of all, if a recall superiority of those expecting recall over those expecting recognition is obtained in the BP condition, it can be attributed to test-expectancy-induced differences in encoding rather than to differences in retrieval practice or PI. If, on the other hand, the recall superiority of those expecting recall over those expecting recognition is larger in the UP condition than in the BP condition, one would need to exercise caution in interpreting test-expectancy effects obtained with the UP procedure, since part or all of these "test-expectancy" effects may in actuality be PI or retrieval practice effects. Second, and more important, if in the BP condition the recall superiority of those expecting recall over those expecting recognition is smallest for the U condition, larger for the R-S condition, and largest for the R-M condition, the results would, within the framework of the Anderson-Bower theory, support the view that people expecting recall do more pathway tagging (form more interitem associations) than do people expecting recognition.

Method

Design. Two between-subjects factors (test expectancy, recall vs. recognition, and balancing of practice, BP vs. UP) and one within-subjects factor (degree of semantic organization, U vs. R-S vs. R-M) were crossed to produce a 2 by 2 by 3 mixed-factors design. To determine if the order of presentation of the recall and recognition tests in the BP groups influences performance on the critical test, two different practice-test orders were tested under the BP factor. In the "recognition-last" ordering, the order of the practice tests was recall, recognition, recognition, recall, recall, recognition; in the "recall-last" ordering, the order of practice tests was recognition, recall, recall, recognition, recognition, recall.

Materials. Six different 20-word lists were constructed to serve as practice lists. The to-be-remembered words (targets) for these practice lists and the lures in the recognition tests for these practice lists were obtained from a pool of 240 unrelated words with frequency counts from 10 to 30 per million (Kučera & Francis, 1967). For the practice list recognition tests, a 5-point confidence rating scale appeared at the top of each page along with the 20 targets and the 20 lures, which were randomly arranged in two 20-word columns. In the rating scale, 5 meant "absolutely certain the word occurred on the most recently presented list," 1 meant "absolutely certain the word did not occur on the most recently presented list," and 3 meant "just guessing."

Each 100-word critical list consisted of three different types of targets: 40 buffer targets, 40 critical targets, and 20 related-pairmate targets. The 40 buffer targets were unsystematically selected from the Kučera and Francis (1967) and Palermo and Jenkins (1964) norms and, as judged by the authors, were only minimally semantically related to other targets in the critical list. The 40 critical targets and the 20 related-pairmate targets were selected from a pool of 80 words representing two high-dominance exemplars from each of 40 semantic categories in the Battig and Montague (1969) norms. One exemplar from each category was arbitrarily designated a critical target, and the remaining exemplar was designated a related-pairmate target.

The same 40 critical targets, representing 40 different semantic categories, appeared in all critical lists. Twenty of these 40 critical targets were randomly assigned to the U condition, 10 were assigned to the R-S condition, and the remaining 10 were assigned to the R-M condition. For each of the 20 critical targets in the U condition, a corresponding related-pairmate target did not occur anywhere in the critical list. Each of the 10 critical targets in the R-S condition was placed in a position in the critical list that was separated from the position occupied by its related-pairmate target by 5-11 words that were buffer targets, other critical targets, and/or other related-pairmate targets. Each of the 10 critical targets in the R-M condition was placed in a position in the critical list that was adjacent to the position occupied by its related-pairmate target.

Four different critical lists were constructed. List 2 was derived from List 1 by interchanging the critical targets that had been assigned to the R-M and R-S conditions in List 1. In this exchange, the critical targets retained their positions in the list, whereas the related-pairmate targets were moved to new

positions in the list that were appropriate to their serving in the massed or spaced condition. Lists 3 and 4 corresponded to Lists 1 and 2, respectively, with the 20 critical targets in the two related conditions being interchanged with the 20 critical targets in the U condition. This exhange necessitated introducing 20 "new" related-pairmate targets into Lists 3 and 4 and dropping out the 20 "old" related-pairmate targets that appeared in Lists 1 and 2. Thus, across the four critical lists, (1) each critical target appeared twice in the U condition, once in the R-S condition, and once in the R-M condition, (2) each related-pairmate target appeared once in the R-S condition and once in the R-M condition, and (3) each buffer target appeared four times.

Procedure. All subjects received six practice lists. In the BP groups, three lists were tested by recall and three by recognition, with two different test orderings ("recognition last" or "recall last"). BP subjects were given both recall and recognition instructions at the beginning of the experiment. A brief review of the instructions appropriate to the test the person was to expect to receive for that particular list was given prior to the presentation of each practice list and the critical list. During practice, the person always received the expected test. After the critical list was presented, BP subjects received a brief review of the instructions appropriate to the recall test they would actually receive on the critical lists.

In the UP groups, the practice lists were tested only by recall or only by recognition tests. UP subjects were initially given only the test instructions appropriate to the practice list tests. A brief review of these instructions preceded the critical list. UP subjects who received the expected recall test on the critical list received these same instructions before the critical test was administered; UP subjects who received the unexpected recall test on the critical list did not receive instructions appropriate to the unexpected recall test until just before the critical test.

Test instructions conveyed only information about the mechanics of the tests. People who received recognition instructions were told to rate how sure they were that a particular word actually occurred on the most recently presented list. People who received recall instructions were told to write down, in any order, as many words as they could remember from the most recently presented list. All people were given a test booklet appropriate to the condition to which they had been assigned. Blank sheets separated the test sheets so that the type of test to be given next was unknown.

Each word was presented via a Carousel slide projector at a 3-sec rate, and 2 min were allowed for each practice list test. After the six practice lists, the people were informed that the next list would be much longer than the previous lists. Following the presentation of the critical list, 1 min of test instructions intervened before the 10-min critical list test. Only one cell of the design was tested in any given session, and each session tested from two to eight people.

Subjects. Two hundred and ninety-four male and female introductory psychology students participated in the experiment in partial fulfillment of a course requirement. They were assigned to one cell of the design in the order in which they signed up, such that n + 1 sessions were not conducted for a particular cell until all cells had n sessions. The number of people tested in each cell is given in Table 1. In the BP groups, nearly equal numbers of people received the "recognition-last" and "recall-last" orderings of the practice tests.

Results

For purposes of statistical analyses, 10 of the 20 critical targets appearing in the U condition were randomly designated as "related-pairmate" targets, such that type of target (critical vs. "related pairmate") could be treated as a factor crossed with the level-of-semantic-organization (U vs. R-S vs. R-M) factor. A preliminary 2 (recall vs. recognition expectancy) by 2 ("recognition-last" vs. "recall-last" ordering of practice tests) by 2 (critical vs. "related-pairmate" targets) by 3 (level of semantic organization, U vs. R-S vs. R-M) mixed-factors analysis of variance was performed on the percentage correct recall data from only the BP groups to determine if the nested ordering-of-practice-tests variable participated in any statistically significant effects. Since it did not, this variable was ignored in all subsequent analyses.

Since the type-of-target (critical vs. "related pairmate") variable did not participate in any statistically significant interactions, Table 1 presents the mean percent correct recall averaged across the two types of targets. Each cell in Table 1 is based on at least 1,280 "observations" (64 people × 20 items). There are four general points to be made about the data displayed in Table 1: (1) The corresponding BP and UP means never differed by more than 4%, and, more important, the patterns of data obtained within the BP and UP groups were the same. (2) The average superiority in the recall of those expecting recall over those expecting recognition was 9% (i.e., 8% in the BP group and 10% in the UP group). (3) Recall was worst in the U condition, was considerably better in the R-S condition, and was somewhat better yet in the R-M condition. (4) The superiority in recall of those expecting recall over those expecting recognition was of comparable magnitudes in the U, R-S, and R-M conditions.

These four observations were supported by a 2 (recall vs. recognition expectancy) by 2 (BP vs. UP testing on practice tests) by 2 (critical vs. related-pairmate targets)

Table 1
Mean Percent Correct Recall in Experiment 1 as a Function of the Level of Semantic Organization, Balancing of Practice, and Test Expectancy

	Level of Semantic Organization and Balancing of Practice								
	Unrelated			Related-Spaced			Related-Massed		
Test Expectancy	BP	UP	Mean	BP	UP	Mean	BP	UP	Mean
Recall (84, 64)	21	21	21	36	36	36	38	38	38
Recognition (82, 64)	16	12	14	26	25	26	30	28	29
Mean Expectancy Difference			7			10			9

Note–BP = balanced practice; UP = unbalanced practice. The numbers in parentheses indicate the number of subjects in the BP and UP groups, respectively.

by 3 (level of semantic organization, U vs. R-S vs. R-M) mixed-factors analysis of variance on the mean percent recall. (Unless otherwise specified, all differences referred to as statistically significant have p values less than .01.) This analysis indicated that the recall of the related-pairmate targets (28%) was slightly, although significantly, higher than the recall of the critical targets (26%) [$F(1,290) = 9.94$, $MSe = 105.92$]. (This slight difference is probably due to item effects.) However, as noted earlier, the type-of-target variable did not participate in any statistically significant interactions (all $ps > .10$). Neither the main effect of the balancing-of-practice variable nor any interactions in which this variable participated approached statistical significance (all $Fs < 1$). However, both the effect of test expectancy [$F(1,290) = 43.65$, $MSe = 777.78$] and the effect of level of semantic organization [$F(2,580) = 145.54$, $MSe = 288.56$] were highly significant. More important, the Test Expectancy by Semantic Organization interaction did not approach statistical significance [$F(2,580) = 1.39$, $MSe = 288.56$].

The conclusion that the effects of the test-expectancy and semantic-organization variables did not interact cannot be attributed to a lack of statistical power. With no other information available, we let Connor's (1977) Experiment 2 finding of a 12% recall expectancy superiority in the recall of blocked-categorized lists serve as a basis for the research hypothesis that the superiority of those expecting recall over those expecting recognition would be 0%, 6%, and 12% in the U, R-S, and R-M conditions, respectively. In the present experiment, the probability that we would have detected an interaction of this magnitude (i.e., one accounting for only 2% of the variance) was .91 for $p < .05$. (See Cohen, 1977, Chapter 8.)

A 2 (UP vs. BP) by 2 (recall vs. recognition expectancy) analysis of variance performed on the percent correct recall for the buffer targets indicated that those expecting recall recalled more buffer targets (12%) than did those expecting recognition (7%) [$F(1,290) = 40.39$, $MSe = 46.79$]. Recall for buffer targets was considerably lower than recall for words in the U condition, probably because the buffer targets had considerably lower imagery values (as judged by the authors).

Discussion

The results of Experiment 1 are relatively clear-cut. The fact that the test-expectancy variable had an effect in the BP groups in which test-expectancy effects were not confounded with retrieval practice or PI effects indicates that this test-expectancy effect was based on test-expectancy-induced differences in encoding. Furthermore, the fact that the recall superiority of those expecting recall over those expecting recognition was of equivalent magnitude in the BP and UP conditions indicates that retrieval practice and PI effects were inconsequential under the conditions employed in the present experiment.

Although both the test-expectancy and semantic-organization variables had large effects on recall performance, there was no evidence of an interaction between the effects of these two variables. This indicates that the test-expectancy-induced encoding differences were not based on differences in the amounts of pathway tagging performed by people expecting recall and those expecting recognition. Thus, if one wishes to interpret the test-expectancy-induced encoding differences obtained in Experiment 1 without adding new encoding processes to those already postulated by the Anderson-Bower theory, one must conclude that those expecting recall merely do more node tagging than do those expecting recognition. If this conclusion is correct, recognition performance should be facilitated by a recall expectancy. Such a result has been obtained when test-expectancy effects have been confounded with PI effects (e.g., Hall et al., 1976, Experiment 1; Maisto, DeWaard, & Miller, 1977, silent rehearsal condition; Poltrock & MacLeod, 1977, Experiment 1; Toglia, Barrett, & Crothers, Note 3, Experiments 1 and 2); however, this result has not been obtained when these effects have not been confounded (e.g., Hall et al., 1976, Experiment 3; Hall, Miskiewicz, & Murray, 1977; Naus, Ornstein, & Kreshtool, 1977; Miller et al., Note 1, Experiments 1, 2, and 4). Thus it is possible that in the former set of studies those expecting recall did better in recognition than those expecting recognition not because they were doing more node tagging but because they were not suffering PI from lures in practice list recognition tests given to those expecting recognition. Experiment 2 tests for this possibility by giving a recognition test on the critical list in a design analogous to that used in Experiment 1.[1]

EXPERIMENT 2

Because Experiment 2 manipulates semantic organization in the same manner as did Experiment 1, its results will be relevant to the issue of whether semantic organization affects recognition performance. Although our discussion to this point has assumed that semantic-organization manipulations have large effects on recall performance and no effect on recognition performance, the empirical evidence on the latter half of this assumption is mixed.[2] For example, comparisons of recognition memory for lists of semantically related words and for lists of unrelated words have yielded no effect of semantic organization (e.g., Kintsch, 1968) or have shown that recognition memory is better for lists of related words (e.g., Kinsbourne & George, 1974; Slamecka, 1975). However, each of these findings is ambiguous because in each of these studies the targets and lures for the related and unrelated lists were different sets of words. Hence, the similarity between the targets and lures was probably not equated for the related and unrelated lists.

The similarity between targets and lures for related

and unrelated lists can be equated by counterbalancing across lists the targets and lures used for the two different kinds of lists. With such a design, both Bruce and Fagan (1970) and Connor (1977) have found that recognition performance is equivalent for the related and unrelated lists. However, this type of design has its own methodological flaw. When equated in terms of list length, the related list must contain fewer categories but more items per category than the unrelated list, which of necessity can contain only one item per category. If all targets are tested and equal numbers of targets and lures appear in the recognition tests for both the related and unrelated lists, it must be the case that each semantic category will be tested more often in the related lists than in the unrelated lists. If recognition worsens the more times a category is "entered" in a recognition test (and there is some evidence that it does; Todres & Watkins, 1981; Neely, Schmidt, & Roediger, Note 4), a facilitatory effect of semantic organization on recognition could be obscured by the recognition test's being more difficult for the related list than for the unrelated list.

When one compares recognition memory for lists of semantically related words that are presented blocked by semantic category or randomly, one avoids the methodological problems associated with comparisons of recognition performance on related and unrelated lists. These problems are avoided because all subjects are tested on the same set of target items interspersed among the same set of lures. From the present perspective, it is interesting that in the three studies employing the blocked-random paradigm in recognition memory (Connor, 1977; D'Agostino, 1969; Jacoby, 1972), a recognition superiority for blocked lists over random lists has been most clearly obtained when subjects explicitly expected a recall test either on the basis of having received practice lists involving only recall tests (Connor, 1977) or on the basis of instructions (D'Agostino, 1969). When subjects expected a recognition test (Connor, 1977) or did not know what type of test they would receive (Jacoby, 1972), the results were less clear-cut. In the former case, the results were ambiguous because of unreliable recognition performance on the random lists. In the latter case, Jacoby (1972) obtained a blocked-random effect only if the blocked-list recognition tests preserved the input order of the blocked list. Thus, when people do not expect recall, failures to obtain a blocked-random effect with randomly ordered recognition tests could be due to a facilitatory effect of blocking at input being offset by a reduction in blocked-list performance produced by the incongruity between the overall organizational structures of the blocked input list and the randomly ordered recognition test. To account for why those expecting recall show a blocked-random effect with randomly ordered recognition tests (e.g., Connor, 1977; D'Agostino, 1969), it must be assumed that the facilitatory effect on recognition produced by blocking at input is larger for those expecting recall than for those expecting recognition. Experiment 2 tests this assumption.

Experiment 2 has two important design features. The first is that it includes a BP condition that unconfounds test-expectancy and PI effects. If recognition performance in the BP condition is better for those expecting recall than for those expecting recognition, the results will support the view that those expecting recall do more node tagging than those expecting recognition. If, on the other hand, the recognition superiority by those expecting recall is obtained only in the UP condition, the results could be attributed to those expecting recall having suffered less PI than those expecting recognition, rather than being attributed to test-expectancy-induced encoding differences.

The second important design feature of Experiment 2 pertains to the semantic-organization manipulation. The three organizational conditions, U, R-S, and R-M, were all embedded in the same critical list, and all subjects received the same critical list recognition test, in which each critical category was tested with the same three exemplars. Thus, Experiment 2 equated for Conditions U, R-S, and R-M (1) the semantic similarity between the targets and lures, (2) the congruity between the overall organization of the to-be-remembered list and the overall organization of the recognition test, and (3) the number of times each category was "entered" in the recognition test. Experiment 2 should therefore provide some relatively unambiguous data concerning whether semantic-organization affects recognition performance and, if so, whether these semantic-organization effects in recognition occur only when people expect recall or also when people expect recognition.

Method

Design and Procedures. The design and procedures for Experiment 2 were identical to those of Experiment 1, with two exceptions. The test on the critical list was a recognition test rather than a recall test, and the instructions given prior to the critical list test were recognition rather than recall instructions.

Materials. The practice lists and critical lists used in Experiment 1 were also used in Experiment 2.

The 200-word recognition test for the critical list was contained on the last two pages of the test booklet. The 5-point confidence rating scale appeared at the top of each of these two pages with the 50 targets and 50 lures randomly interspersed in four 25-word columns on each page. Of the 200 words in the recognition test, 120 consisted of three Battig and Montague (1969) high-dominance exemplars from each of the 40 critical semantic categories. (Two of these exemplars were the critical and related-pairmate targets in Experiment 1.) For those 20 categories assigned to the U condition, the three test words consisted of the critical target and two related lures. (One of these related lures was a related-pairmate target when that category was assigned to the R-S or R-M condition.) For those 10 categories assigned to each of the R-S and R-M conditions, the three test words consisted of the critical target, the related-pairmate target, and one related lure. The remaining 80 words in the recognition test were 40 buffer targets and their associated

40 related lures, which were obtained from either the Palermo and Jenkins (1964) norms or a thesaurus.

Subjects. Two hundred and seventy-two males and females drawn from the same source as those tested in Experiment 1 were assigned to the various cells of the design in the same way as those tested in Experiment 1.

Results

For each subject, a mean percentage correct recognition score was computed, using the high-threshold correction procedure based on the separate hit and false alarm rates for the U, R-S, and R-M conditions. That is, percent correct recognition = (percent hits – percent false alarms) ÷ (100% – percent false alarms), with targets and lures receiving confidence ratings of 4 or 5 being counted as hits and false alarms, respectively. These percentage correct recognition scores were then submitted to the same analyses of variance as those used to analyze the recall results of Experiment 1. As in recall, the order-of-practice variable did not participate in any statistically significant effects for the recognition data in the BP groups.

Since the type of target (critical vs. related pairmate) did not participate in any statistically significant interactions nor have a main effect on recognition performance, Table 2 presents the mean percent correct recognition averaged across the two types of targets. As can be seen from Table 2, the recognition results are similar to the recall results displayed in Table 1: (1) The corresponding BP and UP means never differed by more than 5%. (All effects in which the balancing-of-practice variable participated yielded $Fs < 1.6$.) (2) Those expecting recall did better than those expecting recognition (70% vs. 65%) [$F(1,268) = 4.77$, $MSe = 2{,}102.17$, $p < .05$]. Although it appears that this test-expectancy effect was larger for the BP groups than for the UP groups (8% vs. 2%), the Balancing of Practice by Test Expectancy interaction was not statistically reliable [$F(1,268) = 1.58$, $MSe = 2{,}102.17$]. (3) Recognition was worst for the U condition, somewhat better for the R-S condition, and considerably better for the R-M condition [$F(2,536) = 43.40$, $MSe = 299.20$]. (4) The superiority in recognition of those expecting recall over those expecting recognition was of comparable magnitudes for the U, R-S, and R-M conditions. Although it appears that the test-expectancy effect was somewhat larger in the R-M condition than in the U and R-S conditions and perhaps particularly so for the UP group, neither the Test Expectancy by Semantic Organization interaction nor the Test Expectancy by Semantic Organization by Balancing of Practice interaction reached conventional levels of statistical significance [$F(2,536) = 2.82$ and .25, respectively, $MSe = 299.20$].

As in recall, the absence of a Test Expectancy by Semantic Organization interaction cannot be attributed to a lack of statistical power. To test the power of Experiment 2, we let Connor's (1977) findings of a 6% recall expectancy inferiority in the recognition of unrelated lists and an 8% recall expectancy superiority in the recognition of blocked related lists serve as a basis for the research hypothesis that the "superiority" of those expecting recall over those expecting recognition would be –6%, 0%, and 8% in the U, R-S, and R-M conditions, respectively. The probability that the present experiment would have detected an interaction of this magnitude (i.e., one accounting for only 2.7% of the variance) was .89 for $p < .05$. (See Cohen, 1977, chapter 8.)

A closer inspection of the effects of semantic organization indicates that the recognition results are somewhat different from the recall results. Whereas recall depended more upon the mere occurrence of semantically related words in the list than upon how far apart the semantically related words appeared in the list (see Table 1), the opposite was true in recognition. That is, as shown in Table 2, the difference between recognition in the U and R-S conditions was only 3%, whereas the difference between recognition in the R-S and R-M conditions was 7%. A post hoc t test, based on the MSe for the semantic-organization effects in the analysis of variance, indicated that the 7% difference in recognition in the R-S and R-M conditions was reliably larger than the 3% difference in recognition in the U and R-S conditions [$t(536) = 3.94$]. This contrasts with the finding that the 3% difference in recall in the R-S and R-M

Table 2
Mean Percent Correct Recognition in Experiment 2 as a Function of the Level of Semantic Organization, Balancing of Practice, and Test Expectancy

	Level of Semantic Organization and Balancing of Practice								
	Unrelated			Related-Spaced			Related-Massed		
Test Expectancy	BP	UP	Mean	BP	UP	Mean	BP	UP	Mean
Recall (72, 64)	66	65	66	69	67	68	79	75	77
Recognition (72, 64)	60	63	61	62	67	64	68	70	69
Mean Expectancy Difference			5			4			8

Note–BP = balanced practice; UP = unbalanced practice. The numbers in parentheses indicate the number of subjects in the BP and UP groups, respectively. The percent correct is based on the high-threshold correction procedure. For the recall expectancy group, the percent hits and percent false alarms for the U, R-S, and R-M conditions were as follows: BP–70, 10; 73, 12; 82, 13. UP–68, 10; 71, 10; 78, 13. For the recognition expectancy group, the corresponding data were as follows: BP–62, 9; 66, 11; 73, 11. UP–68, 10; 72, 11; 75, 13.

conditions was reliably smaller than the 13% difference in recall in the U and R-S conditions [$t(580) = 10.64$].

Although the 3% difference in recall between the R-S and R-M conditions and the 3% difference in recognition between the U and R-S conditions were both statistically reliable [$t(580) = 2.72$ and $t(536) = 2.55$, respectively], the latter difference in recognition may overestimate the "true difference." The reason for this is that two of the three entries into each category in the recognition test were lures and only one was a target in the U condition, whereas the opposite was the case in the R-S and R-M conditions. Thus, recognition performance in the U condition may have been artificially low if a lure produces more intracategorical output interference than does a target (see Todres & Watkins, 1981). Three points are important in this regard: (1) This confound is unavoidably tied to the unrelated vs. related manipulation when one tests all targets and equates for the unrelated and related conditions the number of times a category is entered during a yes-no recognition test. (2) The recognition performance difference in the R-S and R-M conditions is not contaminated by this confound. (3) The difference in the pattern of effects that semantic organization had upon recall and recognition performance is, if anything, diminished (not produced) by this confound.

None of the effects was statistically reliable for the analysis performed on buffer target recognition. However, as in recall, buffer target performance (47% correct recognition) was considerably lower than performance in the U condition. Once again, we attribute this to the low imagery values of the buffer targets. Also, the absence of a test-expectancy effect for the buffer targets is not all that surprising, since test-expectancy effects are smaller for abstract words than for concrete words (Wnek & Read, 1980; Miller et al., Note 1; Toglia et al., Note 3, Experiment 2 vs. Experiment 1).

A measure of category dependency in recall and recognition. Although there was no Test Expectancy by Semantic Organization interaction either in overall recall or in overall recognition performance, it could still be the case that recall and/or recognition memory was more "category dependent" for subjects expecting recall than for those expecting recognition. For example, consider imaginary protocols for the R-M condition, in which those expecting recall recalled six items (viz., both the critical target and the related-pairmate target from each of three different categories), whereas those expecting recognition recalled four items (viz., only one critical target or only one related-pairmate target from four different categories). Obviously, in this example, recall was more category dependent for those expecting recall than for those expecting recognition. This would support the view that subjects expecting recall are more likely to form interitem associations between related words (viz., do more pathway tagging) than are those expecting recognition. However, this would have gone undetected in our analyses on overall recall and recognition performance. To test for this possibility and to compare directly the amounts of category dependency in recall and recognition memory, we developed a measure of category dependency in memory that partials out (is independent of) the overall levels of recall and recognition performance.

What we sought to determine was the mean number of times a person would recall or recognize pairs of items from the same category, given that the person remembered at random exactly c critical targets and exactly r related-pairmate targets. Of course, this mean-number-of-pairs expected measure is highly dependent on the overall level of memory performance, since the greater are c and r, the greater is the mean number of pairs expected. To partial out this effect of overall performance level, we compared the standard deviation of the probability distribution of all possible numbers of pairs that could be obtained for a given (c, r) level of performance and then computed separate z scores for the number of pairs that actually occurred for a particular person at a given (c, r) level of performance in each of the U, R-S, and R-M conditions, where $z = (\text{number of observed pairs} - \text{mean number of pairs expected}) \div \text{standard deviation}$ (see Appendix for details).

How should this z-score measure of category-dependent memory be interpreted? If people randomly remember the critical targets and related-pairmate targets from the different categories, this z-score measure averaged across several people should be very near zero. One would presume that such would be the case for the arbitrarily paired critical targets and related-pairmate targets in the U condition. On the other hand, if there is a very strong category dependency in memory, this z-score measure averaged across several people might approach +2.5 to +3.0, since each person should remember many more pairs of items from the same category than would be expected by chance. Since this category-dependent memory measure partials out differences in overall level of performance produced solely by differences in memory for individual items, if those expecting recall only do more node tagging than those expecting recognition, there should be no test-expectancy effect obtained in any condition with this z-score measure. If, on the other hand, those expecting recall are more likely to form interitem associations between related items (viz., do more pathway tagging) than are those expecting recognition, the Anderson-Bower theory predicts that this z-score measure would in recall, but not recognition, yield larger values for subjects expecting recall than for those expecting recognition. Furthermore, this test-expectancy effect for the z-score measure in recall should be largest for the condition representing the highest degree of semantic organization (i.e., the R-M condition).

The z-score measures of category-dependent memory were submitted to analyses of variance similar in all

respects except one to those used for the recall and recognition scores. The exception was that rather than analyzing the recall and recognition data separately, we treated the type of test received (recall vs. recognition) as a between-subjects factor in the overall analysis. Since neither the order-of-practice variable, nested under the BP condition, nor the balancing-of-practice variable participated in any statistically significant effects, Table 3 presents the means of the category-dependent memory measure averaged across these two variables.

The first point of interest is that for both recall and recognition the mean z scores are near zero for the U conditions (thus validating the measure), whereas the mean z scores are greater than zero for the R-S and R-M conditions, particularly for the R-M conditions. Thus, memory became more category dependent as semantic organization increased [$F(2,662) = 231.98$, $MSe = .94$]. Second, and more important, the test-expectancy main effect and the Test Expectancy by Semantic Organization interaction were both inconsequential (both $Fs < 1$). (As can be seen in Table 3, this was true in both recall and recognition.) Thus, although those expecting recall had higher overall levels of performance in recall and recognition than did those expecting recognition, their memories were not more category dependent. The category-dependent memory measure, like the overall recall and recognition performance measures, therefore provides no support for the view that subjects expecting recall are more likely to form interitem associations (viz., do more pathway tagging) than are those expecting recognition. A third point of interest is that recall memory was much more category dependent than was recognition memory [$F(1,331) = 140.52$, $MSe = 1.05$]. Fourth, a significant Test Received by Semantic Organization interaction [$F(2,662) = 38.42$, $MSe = .94$] merely indicates that recall memory was more category dependent than recognition memory for the R-S and R-M conditions only.

Table 3
Mean z Scores for Category-Dependent Memory as a Function of the Level of Semantic Organization, Type of Test Received, and Test Expectancy

Test Expectancy	Level of Semantic Organization		
	U	R-S	R-M
	Recall Test		
Expecting Recall (105)	−.07	1.43	2.34
Expecting Recognition (70)	.03	1.51	2.10
Mean Expectancy Difference	−.10	−.08	+.24
	Recognition Test		
Expecting Recall (77)	−.08	.45	.98
Expecting Recognition (87)	.02	.29	1.02
Mean Expectancy Difference	−.10	+.16	−.04

Note—U = unrelated, R-S = related-spaced, R-M = related-massed. The numbers in parentheses indicate the number of subjects on which the data are based.

One potential problem arises when our category-dependent memory measure is used to compare directly the category-dependent nature of recall and recognition memory. The problem is that those people whose recall data were excluded from the analyses (see Appendix) were people with particularly poor memories (i.e., people who could not recall at least one critical target and at least one related-pairmate target in each of the U, R-S, and R-M conditions), whereas those people whose recognition data were excluded from the analyses were people with particularly good memories (i.e., people who recognized all 10 critical targets or all 10 related-pairmate targets in at least one of the three conditions, U, R-S, and R-M). If people with better memories are more likely to have a greater category dependency in their memories, it could be that recall memory showed a greater degree of category dependency than recognition memory because the recall data came mostly from people with relatively good (and categorically dependent) memories, whereas the recognition data came mostly from people with relatively poor (and categorically independent) memories.

To test this possibility, the people in each distinct between-subjects cell were assigned to either a "good-subjects" or a "poor-subjects" group. This assignment was determined by whether the person's total recall or recognition performance on critical targets and related-pairmate targets was above or below the median for that person's cell. An analysis of variance containing all of the factors of the previous analysis plus the type-of-subject ("good" vs. "poor") factor revealed that the only significant effects in which the type-of-subject variable participated were the Test Received by Type of Subject interaction [$F(1,323) = 4.48$, $MSe = 1.09$, $p < .05$] and the Test Received by Type of Subject by Semantic Organization interaction [$F(2,646) = 5.11$, $MSe = .94$]. Specifically, "good" subjects' memories were more category dependent than "poor" subjects' memories, but only for the recall test and only in the R-S and R-M conditions. The important point is that even the "poor" subjects receiving the recall test had memories with greater category dependencies than did the "good" subjects receiving the recognition test. Thus, it seems unlikely that the greater category dependency obtained for recall memory as compared with recognition memory was due to a subject-selection artifact.

Discussion

The recognition results of Experiment 2 are consistent with the recall results of Experiment 1: (1) In both recall and recognition there was a beneficial effect of a recall test expectancy and of semantic organization, with little evidence of a Test Expectancy by Semantic Organization interaction. Within the framework of the Anderson-Bower theory, these test-expectancy effects are congruent with the view that subjects expecting recall do not do more pathway tagging than those

expecting recognition but rather do more node tagging. Also congruent with this view is the finding that recall memory was not more category dependent for those expecting recall than for those expecting recognition. (2) Since a semantic-organization effect occurred in recognition under both a recall and a recognition expectancy, the organizational effects found in recognition by Connor (1977) and D'Agostino (1969) cannot be attributed solely to the fact that their subjects expected a recall test. (3) Since the beneficial effect of a recall test expectancy on recognition performance was obtained in the BP groups (in which test-expectancy and PI effects were not confounded), we believe (a) that previous experiments, which confounded test-expectancy and PI effects, obtained a recognition superiority by subjects expecting recall not because of the confounding but because the subjects were given sufficient practice on the recall and/or recognition tests and (b) that previous experiments, which did not confound test-expectancy and PI effects, failed to obtain a recognition superiority by subjects expecting recall not because the studies avoided the confounding but because they gave insufficient practice on the recall and/or recognition tests. Apparently, three practice trials each on recall and recognition tests in a BP design are sufficient to produce test-expectancy-induced encoding differences that will affect recognition performance (i.e., the present results and the Balota & Neely, 1980, results), whereas one practice trial on each type of test is not (Hall et al., 1976, Experiment 3).

GENERAL DISCUSSION

Test-Expectancy Effects in Recall and Recognition

Although up to this point we have discussed our test-expectancy results in terms of only the Anderson-Bower theory, we believe they have three general model-free implications. First, the fact that our test-expectancy effects were obtained in the BP condition, in which test-expectancy effects were not confounded with retrieval practice and PI effects, certifies that these effects were due to test-expectancy-induced encoding differences rather than to retrieval practice or PI differences. Second, the absence of a Test Expectancy by Semantic Organization interaction in recall is problematic for any theory that assumes (1) that recall is more dependent than is recognition on episodically encoded interitem associations among semantically related words, (2) that the formation of such interitem associations is under the subject's strategic control, *and* (3) that test expectancies influence the subject's encoding strategy. Third, the finding that a recall expectancy facilitates both recall and recognition performance indicates that those expecting recall better perform an encoding operation that benefits both recall and recognition.[3]

Given the theoretical importance of the inference that those expecting recall are not more likely than those expecting recognition to form episodically encoded interitem associations, it would be unwise to base this inference on data from only one experiment (i.e., our Experiment 1). For example, one could quite reasonably argue that we might have obtained a Test Expectancy by Semantic Organization interaction in recall had we used a more "powerful" manipulation of semantic organization (viz., had we used more than two items per category in our related conditions). However, one should also be mindful of three other considerations: (1) Our semantic-organization manipulation was powerful enough to produce a rather substantial effect in recognition as well as in recall. (2) Test-expectancy-induced differences in the formation of the interitem associations should manifest themselves at the level of specific pairwise semantic associations as well as at "higher" levels of organization. (3) If one is interested in comparing the test-expectancy and semantic-organization effects obtained in recall with those obtained in recognition, one must consider that comparisons of recognition memory as a function of semantic organization are beset with confoundings when one uses more than two items per category in the related organizational conditions. (See introduction to Experiment 2.) Thus, the important point remains that despite a widely held belief that those expecting recall should be more likely than those expecting recognition to produce episodically encoded interitem associations between semantically related items, there is at present no evidence for this belief in the form of a Test Expectancy by Semantic Organization interaction.

Since a recall expectancy facilitates both recall and recognition performance, one must conclude that those expecting recall emphasize an encoding operation that facilitates both recall and recognition. Within the framework of an unembellished version of the Anderson-Bower theory, that encoding operation is node tagging. Of course, the fact that a recall expectancy facilitates recognition is not in and of itself sufficient evidence for the conclusion that test-expectancy effects are being mediated by differences in node tagging. For example, one might argue, contrary to the Anderson-Bower theory, that pathway tags are used in recognition in some sort of "retrieval-checking" process (see Mandler, 1972, 1979, 1980). Such a process could involve a search along the tagged pathways emanating from the node accessed by the presentation of the item in the recognition test. If this were so, any additional pathway tagging induced by a recall expectancy would facilitate recognition as well as recall. However, if this analysis were correct, a Test Expectancy by Semantic Organization interaction should have been obtained in both recall and recognition. Given that this interaction was obtained in neither recall nor recognition, within the framework of the Anderson-Bower theory, the most prudent and parsimonious conclusion is that subjects expecting recall do more node tagging, but not more pathway tagging, than those expecting recognition.

Unfortunately, this interpretation of test-expectancy

effects seems to be contradicted by self-report data obtained by Hall et al. (1976). In a questionnaire administered after a test-expectancy experiment involving a UP procedure, subjects who had expected recall were more likely than those who had expected recognition to report having used encoding strategies involving cumulative rehearsal and the formation of interitem associations and less likely to report having used an encoding strategy based on single-item rehearsal. However, these self-report data do not necessarily disconfirm our conclusion that those expecting recall do more node tagging, but not more pathway tagging, than those expecting recognition. One possible way to harmonize the Hall et al. data with our conclusion would be to note that people's retrospective introspections about how they perform certain tasks are often inaccurate (Nisbett & Wilson, 1977). But this ignores the considerable independent evidence that people expecting recall are indeed more likely to rehearse cumulatively the to-be-remembered items than are those expecting recognition. The sources of this evidence come from the direct observation of overt rehearsal patterns (Naus et al., 1977) and from the following findings: (1) The requirement to use an overt single-item rehearsal strategy at input has different effects on the memories of people expecting recall and those of people expecting recognition (Maisto et al., 1977). (2) A recall test expectancy has its greatest facilitatory effect on memory performance for primacy items, whether primacy is defined in terms of serial position within a list of unrelated words (Poltrock & MacLeod, 1977, Experiment 1; Miller et al., Note 1, Experiments 1 and 2) or in terms of categorical position (Jacoby, 1973) or intracategorical position (Carey & Lockhart, 1973; but see Footnote 1) in a blocked-categorized list. (3) A recall expectancy has its largest facilitatory effect in recognition tests in which memory for order information is important (Leonard & Whitten, Note 5). Findings 2 and 3 support the notion that people expecting recall are more likely to rehearse cumulatively than are those expecting recognition, because Rundus (1971) has shown that primacy effects are correlated with cumulative rehearsal strategies and because Tzeng, Lee, and Wetzel (1979) have presented data congruent with the proposition that order information is "automatically" stored through cumulative rehearsal. Thus, there is considerable evidence congruent with the Hall et al. self-report data that indicate that subjects expecting a recall test are more likely to rehearse cumulatively the to-be-remembered items than are subjects expecting a recognition test.

A second way to harmonize the Hall et al. (1976) self-report data and our conclusion that subjects expecting recall do more node tagging, but not more pathway tagging, than those expecting recognition would be to argue that people may confuse the act of concurrently rehearsing items, an act that necessarily occurs in cumulative rehearsal, with the act of "associating" these items. That is, the act of concurrently rehearsing words could, in Anderson and Bower's (1974) terminology, represent a "cumulative node-tagging" process that does not necessarily lead to the tagging of the pathways that connect these concurrently rehearsed nodes. Support for this argument comes from the finding by Ambler and Maples (1977, Experiment 1) that although the overall level of recall is higher when people rehearse semantically related items concurrently rather than separately, the amount of semantically organized clustering in recall is equivalent for these two different rehearsal conditions. Indeed, Hall et al. themselves argued that their data indicate that subjects expecting recall and those expecting recognition do not adopt qualitatively different encoding strategies. They based their argument, in part, on two facts. First, regardless of expectancy, people were more likely to report having used cumulative rehearsal and interitem associative strategies than they were to report having used a single-item rehearsal strategy. Second, whenever the correlations between memory performance and the degree to which people reported having utilized a particular strategy were statistically significant for both recall and recognition, they had the same signs. Thus, we believe there is no contradiction between our node-tagging interpretation of test-expectancy effects and Hall et al.'s self-report data.

In short, then, we believe that, within the framework of the Anderson-Bower theory, the existing data nicely converge on the idea that subjects expecting recall do more cumulative node tagging, but not more pathway tagging, than do those expecting recognition. However, this raises the issue of why those expecting recall do more node tagging than those expecting recognition, given that those expecting recognition should ideally be devoting all of their resources to node tagging and none to pathway tagging (because pathway tags are presumably not used in recognition). One possibility is that compared with subjects expecting the "easy" recognition test, those who expect the "difficult" recall test study harder; that is, they allocate a larger total reservoir of processing capacity to the encoding of the to-be-remembered words (Kahneman, 1973). According to this study-harder hypothesis, those expecting recall do indeed do their pathway tagging; however, they still have enough of their expanded processing capacity left over to do more node tagging than those expecting recognition. It should be noted, however, that test-expectancy effects in recall and recognition cannot be accommodated by the most simplistic version of a study-harder hypothesis. If such a hypothesis were correct, one would predict that compared with subjects expecting recognition, those expecting recall should do better to the same degree in remembering both high- and low-frequency words. Contrary to this prediction, Balota and Neely (1980) found that, compared with those expecting recognition, those expecting recall did better at recalling and recognizing high-frequency words but no

better at recalling and recognizing low-frequency words. On the basis of these results, Balota and Neely concluded that the additional node tagging performed by subjects expecting recall is of the nature of their tagging more different meaning nodes corresponding to the to-be-remembered word rather than their laying down more node tags for a single meaning node corresponding to the to-be-remembered word. Since high-frequency words presumably have more different meaning nodes available than do low-frequency words (e.g., Glanzer & Bowles, 1976; Reder, Anderson, & Bjork, 1974), memory performance on high-frequency words benefits more from the greater semantic variability in the episodic encodings of subjects expecting recall. Of course, this interpretation of test-expectancy effects is merely a more molecular account of a more sophisticated version of the study-harder hypothesis, which states that subjects expecting recall study harder in terms of episodically encoding more information about each to-be-remembered word's meaning.

There is another possible reason for why, compared with subjects expecting recognition, those expecting recall have extra processing resources available for additional node tagging. Perhaps those expecting recall, like those expecting recognition, lay down only node tags and perform no pathway tagging at all. Of course, in order to entertain this more provocative possibility seriously, one would need to develop alternative explanations of those effects that have heretofore been explained by an appeal to the pathway-tagging mechanism. The most important of these effects is the differential effect that semantic organization has upon recall and recognition performance, and it is to this effect that we now turn our attention.

Semantic-Organization Effects in Recall and Recognition

Previous experiments have shown that increases in the semantic organization of the to-be-remembered list enhance recall performance substantially but have little or no effect on recognition performance. However, as noted in the introduction to Experiment 2, these previous experiments using between-list manipulations of semantic organization have been plagued by several methodological problems that have precluded unambiguous interpretations of their results. Using a within-list manipulation of semantic organization that avoids these methodological problems, we obtained a substantial semantic-organization effect in recognition as well as in recall.

Of course, a problem is posed when one tries to compare directly the magnitudes of the semantic-organization effects obtained in recall and recognition. The problem is that such comparisons are usually contaminated by recognition performance's being considerably higher than recall performance. Our solution to this problem was to develop a standardized z-score measure of category dependency (semantic organization) in memory that partials out differences in the levels of overall performance in the recall and recognition tests. Using this measure, we found that although the semantic-organization effect we obtained in recognition was substantial, it was smaller than that obtained in recall. Such a result is consistent with the widely held view that recall is more dependent upon semantic organization than is recognition, but it is inconsistent with the view that recognition is completely unaffected by this variable.

The fact that recognition is affected by semantic organization can be explained within the framework of the Anderson-Bower generate-recognize theory in at least two different ways. According to the first explanation, if the number of list-marker elements associated with the node activated by an item in the recognition test does not exceed some criterion, a "retrieval check'" may be performed on this item (cf. Mandler, 1972, 1979, 1980) to determine if it is "recallable" by virtue of there being tagged pathways emanating from the node it activates. Since the presence of these tagged pathways depends on the semantic organization of the to-be-remembered items, a semantic-organization effect will occur in recognition whenever retrieval checks are made. Also, since these "retrieval checks" must necessarily have been made for all recalled items in a recall test but not for all recognized items in a recognition test, semantic-organization effects should be larger for recall than for recognition. However, as proponents of generate-recognize theory, we are biased against this explanation of semantic-organization effects in recognition because we believe it undermines the very foundation of that theory by eliminating the recall-specific search process.

We prefer a second, alternative explanation of semantic-organization effects in recognition, an explanation that is already built into Anderson's (1972) assumptions about node tagging. According to these assumptions, a node is tagged with list-marker elements only while it resides in the STS. A node is entered into the STS if it is (1) directly activated by a currently presented to-be-remembered item or (2) retrieved, by virtue of its being related to the currently presented to-be-remembered item, and recognized as representing a previously presented item from the current to-be-remembered list. Since the nodes corresponding to items that are unrelated to all other list items are tagged only under Condition 1, whereas the nodes corresponding to items that are related to other list items may be tagged under both Conditions 1 and 2, the nodes corresponding to semantically related list items are the most likely to be tagged with list-marker elements, thereby enhancing recognition performance on the semantically related list items.

If this explanation of organizational effects in recognition is correct, for the sake of parsimony one may begin to question whether one needs to appeal to

pathway tagging to accommodate the semantic-organization effect in recall. One could argue that the semantic-organization effect in recall occurs both during the recognition stage of recall, in the manner just described, and during the generation stage of recall, by virtue of a "random" search's being conducted along untagged pathways connecting memory nodes embedded in a semantically organized network. Since the organizational effect in recall is, according to this argument, produced by the structure of the network through which the search proceeds as well as by the greater number of list-marker elements that are associated with nodes corresponding to semantically related list words, this argument can account for the organizational effect's being larger in recall than in recognition, without making any appeal whatsoever to the pathway-tagging process. (Glanzer and Bowles, 1976, and Landauer, 1975, have developed explanations of semantic-organization effects in recall that do not appeal to pathway-tagging-like mechanisms. Glanzer and Bowles' explanation is similar to ours, whereas Landauer's is not.) Of course, without additional assumptions, our explanation leaves unexplained our finding that recall depends more upon the mere occurrence of semantically related words in the list than upon how far apart these related items appeared in the list, whereas the opposite was true in recognition. However, rather than enter into premature speculation concerning the nature of these additional post hoc assumptions, we choose merely to note that all of the other explanations of semantic-organization effects that we have entertained here also require additional post hoc assumptions to accommodate this finding. (For a detailed discussion of how the spacing between two exemplars of the same semantic category can affect their separate and conjoint recall, see Batchelder and Riefer, 1980.)

In short, we conclude that semantic-organization effects in recall and recognition can be accounted for within the framework of the Anderson-Bower theory without appealing to a pathway-tagging mechanism. This conclusion makes more comprehensible our finding that, compared with subjects expecting recognition, those expecting recall are not more likely to form episodically encoded interitem associations among the to-be-remembered words. If, as we have argued, the generation stage of recall is "guided" by an already existing structural organization in the memory network in which the nodes that have been tagged with list-marker elements are embedded, there would be no good reason for subjects expecting recall to lay down pathway tags to further guide this already "guided" search.

Is Pathway Tagging Necessary to Account for Other Recall-Recognition Differences?

Given that it is not necessary to appeal to a pathway-tagging mechanism in order to account for organizational effects' being larger for recall than for recognition and that there is no compelling evidence that people expecting recall do more pathway tagging than those expecting recognition, the question is raised whether there is any need to appeal to a pathway-tagging operation to account for why other variables have differential effects on recall and recognition performance. We believe the answer to this question is a provisional "no." The data that seem to require a pathway-tagging assumption the most are those results that show that recall and recognition are differentially affected by the type of encoding operations subjects perform on the to-be-remembered words. For example, as noted in the introduction to Experiment 1, Eagle and Leiter (1964) and Griffith (1975) found that intentional-learning (or categorization) instructions lead to better recall but poorer recognition than do incidental-learning (or imagery) instructions. Such data have been interpreted by Anderson (1972) as indicating that subjects given the intentional (or categorization) instructions were more likely to produce pathway tags and less likely to produce node tags than were those given the incidental (or imagery) instructions. Crowder (1976) has accounted for the finding that increases in maintenance-rehearsal duration facilitate recognition but not recall by assuming that maintenance rehearsal on an item increases the number of tags associated with that item's node but does not increase the number of list-marker tags associated with the pathways emanating from that item's node.

Unfortunately, the pathway-tagging accounts for these results are not as straightforward as they first seem. As has too often been the case when generate-recognize models have been applied to data in which a variable differentially affects recall and recognition, the general logic of these pathway-tagging accounts seems to be of the following form: (1) Experiments that yield an interaction between the effects of some variable and recall and recognition performance implicate the operation of different mechanisms in recall and recognition. (2) The generate-recognize model postulates the operation of different mechanisms in recall and recognition. (3) Therefore, the results of such experiments support the generate-recognize model. Such reasoning falls short whenever a variable affects recognition but has the opposite or no effect on recall. Since recall involves the same recognition process as recognition, in such cases one needs a detailed analysis of how the manipulated variable affects the generation process in recall so as to offset the effects that occur in the recognition process in recall, thereby yielding an effect in recall different from that in recognition. (When a variable affects recall but not recognition, one can simply argue that the variable affects only the generation stage of recall.) Since such a detailed analysis has not been made in the pathway-tagging explanations of the instructional set and maintenance-rehearsal results, these explanations must be regarded as working in principle only. (Tulving, 1976, has made this same general point.)

Given that the pathway-tagging account of the instructional set results is an in-principle explanation only, the question becomes one of whether it is possible to develop an in-principle explanation of these results that does not appeal to pathway tagging at all. Without going into detail here, we will outline a general approach for developing such explanations. This approach is a species of the generic argument that was developed to accommodate encoding specificity effects (e.g., Tulving & Thomson, 1973) and context effects in recognition (e.g., Light & Carter-Sobell, 1970) within the framework of the Anderson-Bower theory. Specifically, it could be argued that which particular nodes corresponding to the to-be-remembered word will be tagged will depend on the type of instructional set subjects are given. It must be further assumed that different nodes corresponding to the to-be-remembered word are accessed by the generation stage in recall and by the presentation of the to-be-remembered word in the recognition test. One then need merely assume that the particular nodes, corresponding to the to-be-remembered word, that were tagged under intentional learning (or categorization) instructions are those that are likely to be generated in recall but are unlikely to be accessed in the recognition test, whereas the opposite is true of the nodes that were tagged under the incidental (or imagery) instructions. Since pathway tags are not being appealed to, the likelihood that tagged nodes will be generated during recall must depend on the nature of the information they represent (e.g., semantic, phonological, graphemic, etc.) rather than on the number of pathway tags that connect them with other list words' nodes. Unfortunately, exactly what the nature of this information is must at present remain unspecified. Nevertheless, this brand of explanation can in principle account for Type of Encoding by Recall-Recognition interaction without appealing to pathway tags.

Obviously, both the present explanation and the pathway-tagging explanation are at this point too ill-specified and too ex post facto to be regarded as anything other than provisional explanations. Before either of these explanations can be accepted, it will be necessary to spell each of them out in detail and submit them to rigorous experimental tests. Nevertheless, the point for the time being remains that one need not necessarily appeal to a pathway-tagging mechanism in order to account for the differential effects that variables have upon free recall and recognition performance. (For a pathway-tagging-free explanation of the differential effects that word frequency has on recall and recognition, see Balota and Neely, 1980.) This does not mean, however, that we can totally abandon the concept of interitem associations (pathway tagging). Such a concept may turn out to be necessary to account for certain phenomena within the framework of generate-recognize theory. Our point is that it is not necessary that generate-recognize theories appeal to pathway tagging to account for the currently available data on recall-recognition differences.

CONCLUSION

The main empirical thrust of the present research is embodied in three findings: (1) People expecting a recall test do better in both recall and recognition than do those expecting a recognition test and do so under conditions in which differences in test expectancy are not confounded with differences in PI and retrieval practice. Thus, we can be reasonably sure that this test-expectancy effect is due to test-expectancy-induced encoding differences. (2) Recognition, as well as recall, is facilitated by increases in the semantic organization of the to-be-remembered words, but this semantic-organization effect is larger in recall than in recognition. (3) Test-expectancy and semantic-organization effects have additive effects in both recall and recognition. Our own theoretical biases predispose us to interpret these three findings within the framework of the Anderson-Bower generate-recognize theory of recall and recognition. When so interpreted, these and other findings converge on two conclusions: (1) Subjects expecting recall do more "cumulative node tagging" than those expecting recognition, and (2) there is at present no compelling empirical evidence that episodically encoded interitem associations are formed during list presentation by those expecting recall in order to benefit their recall performance. Thus, a direction for future research to take would be to try to develop a recall-recognition comparison that would provide compelling, direct evidence for the existence of a pathway-tagging mechanism that is qualitatively distinct from the node-tagging mechanism.

Of course, it may be the case that some of the more recent "unifactor" theories of recall and recognition (e.g., Lockhart, Craik, & Jacoby's, 1976, levels-of-processing theory and Tulving's, 1976, episodic-ecphory theory) can provide more elegant interpretations of test-expectancy and semantic-organization effects than can a modified version of the Anderson-Bower theory. We look forward to proponents of these other theories proffering such interpretations in enough detail that the adequacy of their interpretations of test-expectancy and semantic-organization effects in recall and recognition can be compared with the adequacy of the interpretations we have given here.

REFERENCE NOTES

1. Miller, M. E., Maisto, S. A., Fleming, J. P., & Rosinsky, R. W. *Storage processes for recall and recognition: The effect of instructions.* Unpublished manuscript, University of Wisconsin—Milwaukee, 1978.

2. Neely, J. H., Balota, D. A., & Schmidt, S. R. ***Test-expectancy effects in recall and recognition: A methodological, empirical, and theoretical analysis.*** Manuscript in preparation, 1980.

3. Toglia, M. P., Barrett, T. R., & Crothers, E. J. *Process differences in recall and recognition memory.* Paper presented at the annual meeting of the Psychonomic Society, Denver, Colo., November 1975.

4. Neely, J. H., Schmidt, S. R., & Roediger, H. L., III. *Output-*

interference and priming effects within categories in episodic recognition. Paper presented at the annual meeting of the Psychonomic Society, St. Louis, Mo., November 1980.

5. Leonard, J. M., & Whitten, W. B., II. *Information stored while studying for recall or for recognition.* Paper presented at the annual meeting of the Psychonomic Society, San Antonio, Tex., November 1978.

REFERENCES

AMBLER, B., & MAPLES, W. Role of rehearsal in encoding and organization for free recall. *Journal of Experimental Psychology: Human Learning and Memory,* 1977, **3**, 295-304.

ANDERSON, J. R. FRAN: A simulation model of free recall. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 5). New York: Academic Press, 1972.

ANDERSON, J. R., & BOWER, G. H. Recognition and retrieval processes in free recall. *Psychological Review,* 1972, **79**, 97-123.

ANDERSON, J. R., & BOWER, G. H. *Human associative memory.* Washington, D.C: Winston, 1973.

ANDERSON, J. R., & BOWER, G. H. A propositional theory of recognition memory. *Memory & Cognition,* 1974, **2**, 406-412.

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.

BATCHELDER, W. H., & RIEFER, D. M. Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review,* 1980, **87**, 375-397.

BATTIG, W. F., & MONTAGUE, W. E. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph,* 1969, **80**(3, Pt. 2).

BOWER, G. H., CLARK, M., LESGOLD, A. M., & WINZENZ, D. Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 323-343.

BROWN, J. (Ed.). *Recall and recognition.* New York: Wiley, 1976.

BRUCE, D., & FAGAN, R. L. More on the recognition and free recall of organized lists. *Journal of Experimental Psychology,* 1970, **85**, 153-154.

CAREY, S. T., & LOCKHART, R. S. Encoding differences in recognition and recall. *Memory & Cognition,* 1973, **1**, 297-300.

COHEN, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1977.

CONNOR, J. M. Effects of organization and expectancy on recall and recognition. *Memory & Cognition,* 1977, **5**, 315-318.

CROWDER, R. G. *Principles of learning and memory.* Hillsdale, N.J: Erlbaum, 1976.

D'AGOSTINO, P. R. The blocked-random effect in recall and recognition. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 815-820.

EAGLE, M., & LEITER, E. Recall and recognition in intentional and incidental learning. *Journal of Experimental Psychology,* 1964, **68**, 58-63.

GLANZER, M., & BOWLES, N. Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, **2**, 21-31.

GLENBERG, A., SMITH, S. M., & GREEN, C. Type I rehearsal: Maintenance and more. *Journal of Verbal Learning and Verbal Behavior,* 1977, **16**, 339-352.

GRIFFITH, D. Comparison of control processes for recognition and recall. *Journal of Experimental Psychology: Human Learning and Memory,* 1975, **1**, 223-228.

HALL, J. W., GROSSMAN, L. R., & ELWOOD, K. D. Differences in encoding for free recall vs. recognition. *Memory & Cognition,* 1976, **4**, 507-513.

HALL, J. W., MISKIEWICZ, R., & MURRAY, C. G. Effects of test expectancy (recall vs. recognition) on children's recall and recognition. *Bulletin of the Psychonomic Society,* 1977, **10**, 425-428.

JACOBY, L. L. Effects of organization on recognition memory. *Journal of Experimental Psychology,* 1972, **92**, 325-331.

JACOBY, L. L. Test appropriate strategies in retention of categorized lists. *Journal of Verbal Learning and Verbal Behavior,* 1973, **12**, 675-682.

KAHNEMAN, D. *Attention and effort.* Englewood Cliffs, N.J: Prentice-Hall, 1973.

KINSBOURNE, M., & GEORGE, J. The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior,* 1974, **13**, 63-69.

KINTSCH, W. Recognition and free recall of organized lists. *Journal of Experimental Psychology,* 1968, **78**, 481-487.

KINTSCH, W. Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory.* New York: Academic Press, 1970.

KUČERA, H., & FRANCIS, W. N. *Computational analysis of present day American English.* Providence, R.I: Brown University Press, 1967.

LANDAUER, T. K. Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology,* 1975, **7**, 495-531.

LIGHT, L. L., & CARTER-SOBELL, L. Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior,* 1970, **9**, 1-11.

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.

MAISTO, S. A., DEWAARD, R. J., & MILLER, M. E. Encoding processes for recall and recognition: The effect of instructions and auxiliary task performance. *Bulletin of the Psychonomic Society,* 1977, **9**, 127-130.

MAKI, R. H., & SCHULER, J. Effects of rehearsal duration and level of processing on memory for words. *Journal of Verbal Learning and Verbal Behavior,* 1980, **19**, 36-45.

MANDLER, G. Organization and recognition. In E. Tulving & W. Donaldson (Eds.), *Organization of memory.* New York: Academic Press, 1972.

MANDLER, G. Organization and repetition: Organizational principles with special reference to rote learning. In L. G. Nilsson (Ed.), *Perspectives in memory research.* Hillsdale, N.J: Erlbaum, 1979.

MANDLER, G. Recognizing: The judgment of previous occurrence. *Psychological Review,* 1980, **87**, 252-271.

MANDLER, G., PEARLSTONE, Z., & KOOPMANS, H. S. Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 410-423.

NAUS, M. J., ORNSTEIN, P. A., & KRESHTOOL, K. Developmental differences in recall and recognition: The relationship between rehearsal and memory as test expectation changes. *Journal of Experimental Child Psychology,* 1977, **23**, 252-265.

NISBETT, R. E., & WILSON, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 1977, **84**, 231-259.

PAIVIO, A. *Imagery and verbal processes.* New York: Holt, Rinehart, & Winston, 1971.

PALERMO, D. S., & JENKINS, J. J. *Word association norms: Grade school through college.* Minneapolis: University of Minnesota Press, 1964.

POLTROCK, S. E., & MACLEOD, C. M. Primacy and recency in the continuous distractor paradigm. *Journal of Experimental Psychology: Human Learning and Memory,* 1977, **3**, 560-571.

REDER, L. M., ANDERSON, J. R., & BJORK, R. A. A semantic interpretation of encoding specificity. *Journal of Experimental Psychology,* 1974, **102**, 648-656.

RUNDUS, D. Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology,* 1971, **89**, 63-77.

SLAMECKA, N. J. Intralist cueing of recognition. *Journal of Verbal Learning and Verbal Behavior,* 1975, **14**, 630-637.

TODRES, A. K., & WATKINS, M. J. A part-set cuing effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory,* 1981, 7, 91-99.

TULVING, E. Ecphoric processes in recall and recognition. In J. Brown (Ed.), *Recall and recognition.* New York: Wiley, 1976.

TULVING, E., & THOMSON, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review,* 1973, **80**, 352-373.

TULVING, E., & WATKINS, O. C. Recognition failure of words with a single meaning. *Memory & Cognition,* 1977, **5**, 513-522.

TVERSKY, B. Encoding processes in recognition and recall. *Cognitive Psychology,* 1973, **5**, 275-287.

TVERSKY, B. Eye fixations in prediction of recognition and recall. *Memory & Cognition,* 1974, **2**, 275-278.

TZENG, O. J. L., LEE, A. T., & WETZEL, C. D. Temporal coding in verbal information processing. *Journal of Experimental Psychology: Human Learning and Memory,* 1979, **5**, 52-64.

WNEK, I., & READ, J. D. Recall and recognition encoding differences for low- and high-imagery words. *Perceptual and Motor Skills, 1980,* **50**, 391-394.

WOODWARD, A. E., BJORK, R. A., & JONGEWARD, R. H. Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior,* 1973, **12**, 608-617.

NOTES

1. It should be noted that two experiments that have used blocked-categorized lists and have induced test expectancy using a UP procedure involving three (Jacoby, 1973) or five (Carey & Lockhart, 1973) practice tests have failed to find a recognition superiority by subjects expecting recall over those expecting recognition. However, in the Jacoby experiment, those expecting recall also did no better in recall than those expecting recognition. In the Carey and Lockhart (1973) experiment, there was no main effect of test expectancy in recall, whereas in recognition, those expecting recall did worse than those expecting recognition. However, the interpretation of these results is complicated by two factors. The first complication is that statistically significant Intracategorical Serial Position by Test Expectancy interactions were obtained in both recall and recognition. The nature of these interactions was that recall and recognition performance was equivalent across the intracategorical serial positions for subjects expecting recognition but decreased across the intracategorical serial positions for those expecting recall, such that those expecting recognition did better than those expecting recall in both recall and recognition but only for items presented as the last two or three members of each category. A second complication was that recall performance decreased across the five recall practice lists, whereas recognition performance increased across the five recognition practice lists. (Carey and Lockhart explicitly ignored these effects because they were so small.) This is a complication because the effects of test expectancy were evaluated by comparing performance on the expected test given on the fifth list with performance on the unexpected test given on the sixth list. Because of these complications and because this is the only study we know of using word lists that has obtained superior memory performance for subjects expecting recognition over those expecting recall, we have chosen to ignore this one discrepant finding in our discussion of the nature of test-expectancy effects obtained in recall and recognition.

2. We consider here only "traditional" list-learning experiments in which subjects are merely told to memorize a list of words that is presented with only one of the to-be-remembered word's being exposed to the subject at any one time. Semantic-organization effects in recognition are well documented when subjects themselves manipulate semantic organization by sorting "unrelated" words into their own subject-defined categories (Mandler, 1972; Mandler, Pearlstone, & Koopmans, 1969) and in experiments in which semantic organization is induced by presenting the to-be-remembered items in a format corresponding to a tree-like conceptual hierarchy (Bower, Clark, Lesgold, & Winzenz, 1969). However, since these results may not extrapolate to the traditional list-learning paradigm, we will not give them further consideration here.

3. It should be noted that the test-expectancy effect obtained in recognition memory for words is opposite to the one obtained for pictures. For example, Tversky (1973, 1974) found that subjects expecting recall do worse than those expecting recognition in a recognition test for picture memory. However, since those expecting recall expected to recall the verbal labels naming the objects represented by the pictures, whereas those expecting recognition expected to be tested on the pictures themselves, Tversky's results are probably more relevant to the idea that pictorial information and verbal information are encoded in distinct and independent formats (Paivio, 1971) than they are to test-expectancy-induced differences in encoding per se. Thus, Tversky's findings will not be given further consideration here. (See Neely et al., Note 2, for a more detailed discussion of Tversky's results.)

APPENDIX

To best understand how we computed our standardized z-score measure of category dependency in memory, consider a recall (or recognition) protocol for the 10 categories representing the U, R-S, or R-M condition. In this protocol, a 1 represents the recall (recognition) of a critical target (CT) or a related-pairmate target (RPMT) in a particular semantic category and a 0 represents the nonrecall (nonrecognition) of a CT or RPMT in a particular semantic category. An example of such a protocol is as follows:

	Category									
	1	2	3	4	5	6	7	8	9	10
CT	0	0	1	1	0	1	0	0	0	0
RPMT	0	0	0	1	0	1	0	1	1	0

In this example protocol, the subject recalled (recognized) three CTs (c = 3) and four RPMTs (r = 4) and in two cases (p = 2) the recalled (recognized) CTs and RPMTs were members of the same category (i.e., Categories 4 and 6). Each subject contributed three such 10-category protocols, one each for the U, R-S, and R-M conditions. (In the U condition, membership in the same "category" was based on an arbitrary designation established before the scoring began.)

What we sought to compute for each such protocol was the z score for the p obtained in that protocol. To compute such a z score, one needs to compute (1) the mean p, E(p), that would be obtained in that protocol under the assumption that the subject randomly recalled (recognized) the c CTs and the r RPMTs and (2) the standard deviation, SD(p), of the probability distribution of all possible values of p that could be obtained for the given values of c and r. If, in any of the three conditions U, R-S, and R-M, c or r was equal to 0 or c or r was equal to 10, all of that subject's data were excluded from further analysis using this z-score measure. The reason for such exclusions is that, in the former case, p must always be 0 and, in the latter case, it must always be that value of c or r that is not 10, with the result that SD(p) would be 0 and z(p) would be undefined.

To describe our computations, we need to designate for each condition three other values: the total number (n) of items recalled (recognized), the greater (g) of c and r, and the lesser (l) of c and r. In our example, n = 7, g = 4, and l = 3. In general, the possible p values that can be obtained for a given protocol is given by the integer series i . . . j, where i is 0 when $n \leq 10$ and i is n − 10 when $n > 10$, and where j is always equal to l. Thus,

in our example, the possible values of p are 0, 1, 2, and 3. To compute the probability of obtaining exactly x pairs, P(p = x), one needs (1) to compute the total number, N(p = x), of different ways one could obtain exactly x pairs for the given c and r values, (2) to compute the total number, N(c,r), of different ways the c CTs and r RPMTs could be arranged within the 10 categories, and (3) to compute N(p = x) ÷ N(c,r), the quotient of which is P(p = x). The following formulas are germane to computing our z score:

$$N(p = x) = \binom{10}{l} \times \binom{10 - l}{g - x} \times \binom{l}{x}. \quad (1)$$

$$N(c,r) = \binom{10}{c} \times \binom{10}{r}. \quad (2)$$

$$E(p) = \sum_{x=i}^{j} P(p = x) \times x, \quad (3)$$

where i and j are defined as they are above.

$$SD(p) = \left\{ \sum_{x=i}^{j} P(p = x) \times [x - E(p)]^2 \right\}^{1/2}, \quad (4)$$

where i and j are defined as they are above.

In Equation 1, the first term is the number of ways the *l* items can be distributed across the 10 categories. The second term is the number of ways that g – x of the g items that are not "paired" with an *l* item can be distributed across the 10-*l* categories not represented by the *l* items. The third term is the number of ways that x of the g items that are "paired" with an *l* item can be distributed across the *l* categories represented by an *l* item. In Equation 2, the first and second terms are the number of ways the c and r items, respectively, can be distributed across the 10 categories. In our example, P(p = 0, 1, 2, 3) = .167, .500, .300, and .033; E(p) = 1.2; SD(p) = .7483; z(p = 0, 1, 2, 3) = −1.60, −.27, +1.07, +2.41.

One final comment: one potential problem with this z-score measure is that it does not correct for guessing (false alarms). However, since the false alarm rates were relatively low and did not vary much as a function of the different organizational conditions in the present experiment (see note in Table 2), we do not see this failure to correct for guessing as being very problematic in the present application.

(Received for publication July 14, 1980;
revision accepted December 12, 1980.)