Journal of Experimental Psychology:
Human Learning and Memory
1980, Vol. 6, No. 5,576-587

Test-Expectancy and Word-Frequency Effects in Recall and Recognition

David A. Balota
University of South Carolina

James H. Neely
Purdue University

To determine if people who expect a recall (RCL) test encode a list of to-be-remembered (TBR) words differently than those who expect a recognition (RGN) test, people were first induced to expect a RCL or a RGN test and then were asked to remember a critical list consisting of both high-frequency (HF) and low-frequency (LF) words. Following presentation of the critical list, different groups received either an expected RCL test, an unexpected RCL test, an expected RGN test, or an unexpected RGN test. There were two main results: (a) People who expected RCL did better in both RCL and RGN than did people who expected RGN, but to a much greater degree for HF than LF words. (b) The standard word-frequency effect was obtained; namely, HF words were better recalled but more poorly recognized than LF words. These data were interpreted within the framework of Anderson and Bower's generate-recognize theory as indicating that, compared to people expecting a RGN test, people expecting a RCL test more variably encode the semantic interpretations of the TBR word. The implications that these data have for Glanzer and Bowles' theory of the word-frequency effect and for classroom examinations were also discussed.

One of the more common questions students ask about a forthcoming exam is whether it will be an essay or a multiple-choice exam. In asking such a question, students supposedly seek to determine whether the to-be-remembered (TBR) information must be recalled from memory without explicit cues (essay) or must be recognized in the midst of incorrect information (multiple choice). Presumably, they then try to tailor their study strategies to maximize their performance on the designated exam. Such test-specific study strategies are relevant to both educational and theoretical issues. For the educator, the issue is one of determining the extent to which test-specific study strategies lead to increased retention and better understanding of the information. For the theoretician, the issue is one of determining the manner in which different encoding processes are differentially emphasized as a function of the type of memory test that is expected. Although the present research focuses primarily on the theoretical issues, our results can perhaps shed light on the educational issues as well.

The writing of this article was supported by a Faculty XL grant from Purdue Research Foundation to the second author. Because both authors contributed equally, the order of authorship was determined by lot. We thank Henry L. Roediger III for his helpful comments on various drafts of the article. Portions of this research were reported at the 1980 meeting of the Southeastern Psychological Association in Washington, D.C.

Requests for reprints should be addressed to James H. Neely, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

Anderson and Bower's Generate-Recognize Theory of Recall and Recognition

Two Kinds of Encoding: Node Tagging and Pathway Tagging

Anderson and Bower's version of generate-recognize theory (e.g., Anderson, 1972; Anderson & Bower, 1972, 1974) provides one framework within which to interpret test-expectancy-induced differences in en-

coding.[1] This theory assumes that free-recall performance depends on two sequentially ordered processes. A search-guided generation process occurs first and provides output to a recognition process presumably identical to the one mediating performance in recognition tests of memory. These generation and recognition processes access information stored in a memory network consisting of (a) conceptual nodes corresponding to different meanings and (b) associative pathways connecting semantically and/or associatively related nodes. Information about list membership is stored as a set of contextual elements encoded together as a list-marker node that represents the exteroceptive and interoceptive stimulation occurring during list presentation. During encoding, information about list membership is incorporated into the memory network via two kinds of tagging. When a TBR word is presented, a node corresponding to a particular meaning of that word is associated with list-marker elements via *node tagging*. At the same time, nodes associatively and/or semantically related to that node are retrieved and submitted to a recognition test. If any of these related nodes are recognized as having been represented by a prior word in the current list, the pathways connecting these nodes to the node corresponding to the current TBR word are also associated with list-marker elements via *pathway tagging*.

According to the Anderson-Bower theory, a word presented in a recognition (RGN) test activates a node in memory corresponding to one of its meanings. RGN performance is directly proportional to the number of retrieved contextual elements that are associated with both the activated node and the list-marker node. In recall (RCL), there is an extra process of generating items (viz., searching for and finding nodes) that are submitted to this recognition process. To guide this search, the person uses the associative pathways that were tagged during input with list-marker elements.

Since node tagging and pathway tagging are differentially important in mediating RCL and RGN, it should be possible to affect differentially RCL and RGN by inducing differential emphases on node tagging or pathway tagging. Some support for this is provided by the finding that intentional learning instructions can produce better RCL, but poorer RGN, performance than incidental learning instructions (Eagle & Leiter, 1964). To account for this, Anderson (1972, p. 370) argued that, compared to the incidental learners, the intentional learners produced more associative-pathway tags (thereby facilitating their generation process in RCL) and fewer node tags (thereby hurting their RGN performance).

An Account of Test-Expectancy-Induced Differences in Encoding

Given the above assumptions, what does the Anderson-Bower theory predict about the effects of test-expectancy-induced differences in encoding on RCL and RGN performance? In line with Anderson's (1972) account of Eagle and Leiter's (1964) instructional-set results, one might surmise that, compared to those expecting RGN, those expecting RCL should be more likely to lay down pathway tags and less likely to lay down node tags. To validate this analysis, one must more directly test the idea that people who expect RCL do more pathway tagging than those who expect RGN. The rationale for such a test can be based on the Anderson-Bower theory's account of semantic organization effects in RCL and RGN. Indeed, one empirical cornerstone of generate–recognize theories is the finding that manipulations of the semantic organization of the TBR list have differential effects on RCL and RGN performance. For example, Kintsch (1968) found that RCL performance was better for lists containing words with strong semantic relationships with one another than for

[1] Although Kintsch's (1974, chap. 4) theoretical characterization of the encoding processes involved when people try to learn a list of words is described in a theoretical language seemingly different from that utilized by Anderson and Bower, as far as we can tell the two theories make the same predictions about the effects investigated here. Thus, to shorten our discussion, we will describe only the Anderson and Bower theory.

lists containing words with weak semantic relationships with one another; however, no such difference occurred in RGN performance. Such results nicely conform to the Anderson-Bower generate–recognize theory. That is, the more semantically related the TBR words are, the more likely it is that the associative pathways connecting their nodes will be tagged with pathway tags during encoding. These tagged associative pathways can then facilitate the generation of nodes in RCL. Once generated, these nodes are then submitted to a RGN test which, like the experimenter-generated RGN test, is unaffected by the semantic relationships among the TBR words.

Given this analysis, one can, by examining the joint effects of test expectancy and semantic organization on RCL performance, test the idea that people who expect RCL do more pathway tagging than those who expect RGN. Since the additional pathway tagging performed by those expecting RCL should be more likely to occur for the associative pathways connecting the nodes of TBR words bearing strong semantic relationships to other list items, the Anderson-Bower theory predicts that the superiority in RCL by those who expect RCL, relative to those expecting RGN, should increase as the semantic organization among the TBR words increases. Neely and Balota (Note 1, Experiment 1) tested for this predicted Test Expectancy × Semantic Organization interaction and found that those who expected RCL did better in RCL than those who expected RGN, but to the same degree for both semantically related and unrelated words. This result was also obtained in a second experiment involving a RGN test.

Given that Neely and Balota's (Note 1) results contravene the most straightforward application of the Anderson-Bower theory to test-expectancy effects, how can the theory accommodate the finding that people who expect RCL do better in both RCL and RGN than do people who expect RGN? Neely and Balota (Note 1) argued that the most parsimonious interpretation of their results is that, compared to people who expect RGN, people who expect RCL better perform an encoding operation that facilitates both RCL and RGN performance. If one chooses not to embellish the current version of the Anderson-Bower theory with additional assumptions, this encoding operation is node tagging. Thus, Neely and Balota interpreted their data as indicating that people who expect RCL do more node tagging than do people who expect RGN.[2]

---

[2] Three points need to be interjected here. First, Connor (1977) has obtained data which, taken at face value, indicate that test expectancy and semantic organization effects are interactive, rather than additive, in both RCL and RGN. However, several aspects of her data and her experimental design make her data difficult to interpret (see Neely & Balota, Note 1). Furthermore, there are other data, besides those reported by Neely and Balota (Note 1), that converge on the conclusion that test expectancy and semantic organization effects are additive (see Neely & Balota, Note 1). Second, the notion that people expecting RCL do more node tagging (rather than more pathway tagging) than do people who expect RGN requires further explication. The relative amounts of processing capacity that people expecting RCL will commit to pathway tagging and node tagging should depend on the relative difficulties they encounter during RCL in performing the generation and recognition processes. An assumption that the recognition process is relatively more difficult than the generation process would help to explain why those expecting RCL allocate more of their limited processing capacity to node tagging than to pathway tagging. However, such reasoning fails to explain why those expecting RCL actually end up doing more node tagging than those expecting RGN, since those expecting RGN should ideally be devoting all of their resources to node tagging and none to pathway tagging, because pathways tags are presumably not used in RGN. To account for why those expecting RCL end up doing more node tagging than those expecting RGN, it must be further assumed that compared to people who expect the "easier" RGN test, those who expect the "more difficult" RCL test allocate a larger reservoir of processing capacity to the encoding of the TBR words (see Kahneman, 1973). The third point is that the test-expectancy effect obtained in RGN memory for words is opposite to the one obtained in RGN memory for pictures. For example, Tversky (1973, 1974) found that, compared to those who expect RGN, those who expect RCL do worse in a RGN test for picture memory. An explanation of the discrepancy that exists between the test-expectancy effects that are obtained in RGN memory for words versus pictures can be found in Neely, Balota, and Schmidt (Note 2).

The Present Research

The major purpose of the present research is to examine the nature of the additional node tagging people presumably do when they expect a RCL test. Within the framework of the Anderson-Bower theory, "additional node tagging" can have at least two different meanings. It could mean that, compared to people who expect RGN, people who expect RCL either (a) associate more list-marker elements with one particular node corresponding to the TBR word or (b) associate list-marker elements with more different nodes corresponding to the different meanings of the TBR word. Or, to put it another way, compared to people who expect RGN, people who expect RCL more variably encode either (a) the contextual elements or (b) the semantic interpretations of the TBR word (cf. Hintzman, 1974). To distinguish between these two senses of additional node tagging, one must control how many different memory nodes corresponding to a TBR word are available for tagging. In the present experiment this will be controlled by the frequency of occurrence in the language of the TBR words, since there is evidence that high-frequency (HF) words generally activate more different memory nodes than do low-frequency (LF) words (e.g., Glanzer & Bowles, 1976; Reder, Anderson, & Bjork, 1974). To make the derivations of the predictions clear, we will assume the limiting case in which a given LF word activates only one node in memory.

Consider first what is predicted if, for each LF and HF word, the people expecting RCL and those expecting RGN select for tagging only one of the memory nodes corresponding to that word, with those expecting RCL tagging it with more list-marker elements than those expecting RGN. If this were the case, the effects of test expectancy and word frequency should be additive in both RCL and RGN. That is, people expecting RCL should do better in both RCL and RGN than those expecting RGN and do better to the same degree for both the HF and LF words.

A different pattern of results is predicted if, in comparison to people expecting RGN, people who expect RCL associate list-marker elements to more different nodes corresponding to the different meanings of the TBR word. For a LF word, the person expecting RCL would have only one node, corresponding to the single meaning of that word, available for tagging and therefore would be unable to tag more different nodes than the person expecting RGN, thereby eliminating any test-expectancy effect on that LF word. However, for a HF word, the person expecting RCL would have several different nodes available for tagging and would tag more of these different nodes than would the person expecting RGN. If this were the case, the effects of test expectancy and word frequency would interact in both RCL and RGN. More specifically, the people expecting RCL should do better in both RCL and RGN than those expecting RGN and to a greater degree for HF words than for LF words.

To discriminate between the two senses of additional node tagging described earlier, in the present experiment we first induced people to expect an RCL or an RGN test and then asked them to remember a list consisting of both HF and LF words. To obtain information on the role that practice on RCL and RGN tests can play in modulating test-expectancy effects, six practice lists were given under two different conditions. In the unbalanced-practice (UP) condition, people who were to expect a RCL test and people who were to expect a RGN test were given practice lists tested only by RCL or RGN tests, respectively. In the balanced-practice (BP) condition, three practice lists were tested by RCL and three by RGN, with each practice list being 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 RCL or RGN test expectancy for the critical list, the presentation of which was followed by different groups' receiving either an expected RCL test, an unexpected RCL test, an expected RGN test, or an unexpected RGN test. As just described, if the effects

of the word-frequency and test-expectancy variables are additive in both RCL and RGN, the results will favor the view that, compared to those expecting RGN, those expecting RCL associate more list-marker elements with one particular node corresponding to a particular meaning of the TBR word. If, on the other hand, the effects of the word-frequency and test-expectancy variables interact in both RCL and RGN (with those expecting RCL doing better in both RCL and RGN and to a greater degree for HF words than for LF words), the results will favor the view that, compared to those expecting RGN, those expecting RCL associate list-marker elements to more different nodes corresponding to the different meanings of the TBR word.

Method

Design. Three between-subjects factors (test expectancy, RCL vs. RGN; balancing of practice tests, BP vs. UP; and type of test received, RCL vs. RGN) and one within-subjects factor (word frequency, HF vs. LF) were crossed to produce a $2 \times 2 \times 2 \times 2$ mixed-factor design. To determine if the order of presentation of the RCL and RGN tests in the BP groups influences performance on the critical test, two different practice-test orders were nested under the BP factor. In the RGN-last ordering, the order of practice tests was RCL, RGN, RGN, RCL, RCL, RGN; in the RCL-last ordering, the order of practice tests was RGN, RCL, RCL, RGN, RGN, RCL.

Materials. Six different 20-word lists were constructed to serve as practice lists. The TBR words (targets) for these practice lists and the lures in the RGN 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 RGN 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 a random ordering of 50 HF and 50 LF concrete nouns that could serve as the object of a sense verb such as *see*, *hear*, *touch*, or *smell*. The HF and LF words had frequency counts in Kučera and Francis (1967) greater than 34 per million and less than 4 per million, respectively. Four critical lists were constructed. The 50 HF and 50 LF targets for List 1 were randomly selected without replacement from the 100 HF and 100 LF words that constituted the critical-list word pool. The 50 HF and 50 LF targets for List 2 were the words that remained in this pool. The targets for List 1 served as the lures in the RGN test for List 2, and the targets for List 2 served as the lures in the RGN test for List 1. Lists 3 and 4 were derived from Lists 1 and 2, respectively, by exchanging the input positions of the HF and LF targets. The same two-page RGN test was used for all four critical lists. The confidence rating scale used for the practice RGN tests appeared at the top of each page, and each page contained an equal number of HF and LF targets and lures, which were randomly interspersed in four 25-word columns. Thus, across the four critical lists, each input position in the TBR critical list was equally often occupied by a HF and a LF target; each HF and LF word equally often served as a target and a lure in the RGN tests; and each output position in the RGN test was equally often occupied by a target and a lure.

Procedure. All subjects received six practice lists. In the BP groups, three lists were tested by RCL and three by RGN, with two different test orderings (RGN last or RCL last). BP subjects were given both RCL and RGN 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 test they would actually receive on the critical list.

In the UP groups, the practice lists were tested only by RCL or only by RGN 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 test on the critical list received these same instructions before the critical test was administered; UP subjects who received the unexpected test on the critical list did not receive instructions appropriate to the unexpected type of test until just before the critical test.

Test instructions conveyed only information about the mechanics of the tests. People who received RGN instructions were told to rate how sure they were that a particular word actually occurred on the most recently presented list. People who received RCL 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 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 seventy-three male and female introductory psychology students participated in this 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 (Footnote a). (In the BP groups, nearly equal numbers of people received the RGN-last and RCL-last orderings of the practice tests.)

Results

The mean percent RCL and mean percent correct RGN scores for the critical-list test are given in Table 1. Each cell is based on at least 1,600 observations (32 people × 50 items). In RGN, targets and lures that received confidence ratings of 4 or 5 were treated as hits and false alarms, respectively, and separate false-alarm rates were computed for HF and LF lures. The RGN scores are based on the high-threshold correction, that is, % correct = (% hits − % false alarms) ÷ (100% − % false alarms). However, the pattern of data shown in Table 1 is not specific to this measure of RGN performance; the same pattern was also obtained for hits alone. Also, the only substantial difference in false-alarm rates (false-alarm rates were higher for HF lures than for LF lures) was in the opposite direction of the corresponding difference in hit rates (hit rates were lower for HF targets than for LF targets). Thus, there is little danger that a high hit rate in any condition was due to people's adopting a perversely lenient criterion for their high confidence ratings in that condition.

Three points need to be made about the results presented in Table 1. First, the patterns of data obtained within the BP and UP groups were quite similar, and the corresponding BP and UP means never differed by more than 4%. Second, and most important in terms of present purposes, people who expected RCL did better in both RCL and RGN than those who expected RGN, with this test-expectancy effect being more or less confined to HF targets. More specifically, the memory superiority of those expecting RCL over those expecting RGN was 7% (in RCL) and 10% (in RGN) for HF targets but was only 3% (in RCL) and 0% (in RGN) for LF targets. This Test Expectancy × Word Frequency interaction suggests that people who expect RCL associate list-marker elements to more different nodes corresponding to the different meanings of the TBR word than do people who expect RGN. Third, the standard word-frequency effect was obtained; RCL scores were 5% higher for HF targets than for LF targets, whereas RGN scores

Table 1

Mean Percent Correct RCL and Mean Percent Correct RGN for HF and LF Targets as a Function of Test Expectancy and Balancing of Practice

Type of test and test expectancy	Type of target and balancing of practice					
	HF			LF		
	BP	UP	*M*	BP	UP	*M*
RCL						
Expects RCL (39, 32)[a]	27	23	25	18	17	18
Expects RGN (34, 32)	20	16	18	16	14	15
M expectancy difference			7			3
RGN						
Expects RCL (36, 32)	57	53	55	73	69	71
Expects RGN (36, 32)	44	46	45	70	73	71
M expectancy difference			10			0

Note. RCL = free recall; RGN = recognition; HF = high frequency; LF = low frequency; BP = balanced practice; UP = unbalanced practice. Percent correct RGN is based on the high-threshold correction procedure. See text for details.

[a] Numbers in parentheses indicate the number of people tested in the BP and UP groups, respectively.

were 21% lower for HF targets than for LF targets.

These three conclusions were supported by analyses of variance. Preliminary 2 (RCL vs. RGN test expectancy) × 2 (RGN-last vs. RCL-last ordering of practice tests) × 2 (HF vs. LF targets) analyses of variance were performed on the RCL and on the RGN scores from the BP groups to determine if the ordering of practice tests influenced performance. (Unless otherwise specified, all significant effects have two-tailed p values less than .05.) The only significant effect in which ordering of practice tests participated was its main effect on RGN performance. RGN scores with RGN-last were 8% higher than with RCL-last, $F(1, 68) = 4.93$, $MS_e = 522.66\%^2$. Since ordering of practice tests did not participate in any significant interactions, this variable was ignored in all subsequent analyses.

For the RCL scores, a 2 (BP vs. UP) × 2 (RCL vs. RGN test expectancy) × 2 (HF vs. LF) analysis of variance indicated that (a) balancing of practice tests did not participate in any significant effects; (b) the main effects of test expectancy, $F(1, 133) = 8.37$, $MS_e = 187.95\%^2$, and word frequency, $F(1, 133) = 41.01$, $MS_e = 48.72\%^2$, were both highly significant; and (c) the Test Expectancy × Word Frequency interaction was significant, $F(1, 133) = 5.06$, $MS_e = 48.72\%^2$. A post hoc t test, the error term for which was computed from the MS_e for the Test Expectancy × Word Frequency interaction, indicated that the difference in RCL scores between those who expected RCL and those who expected RGN was larger for HF targets (7%) than for LF targets (3%), $t(133) = 3.18$. The latter difference was, however, also significant, $t(133) = 2.43$.

For the RGN scores, a similar analysis of variance indicated that (a) balancing of practice tests did not participate in any significant effects; (b) the main effect of test expectancy, $F(1, 132) = 2.68$, $MS_e = 647.32\%^2$, was only marginally significant ($p = .10$), whereas the main effect of word frequency was highly significant, $F(1, 132) = 255.18$, $MS_e = 120.83\%^2$; and (c) the Test Expectancy × Word Frequency interaction was also highly significant, $F(1, 132) = 15.92$, $MS_e = 120.83\%^2$. A post hoc t test, the error term for which was computed from the MS_e for the Test Expectancy × Word Frequency interaction, indicated that the superiority by those who expected RCL over those who expected RGN was larger for HF targets (10%) than for LF targets (0%), $t(132) = 5.66$.

General Discussion

Implications for Test-Expectancy Effects

The most interesting finding was that the superiority in memory performance of those who expected RCL over those who expected RGN was quite pronounced for HF words but nearly nonexistent for LF words. Similar findings have been obtained in other laboratories. For example, Miller, Maisto, Fleming, and Rosinsky (Note 3, Experiment 4) examined the joint effects of test expectancy (manipulated through instructions only), word frequency (manipulated between lists), and concreteness–abstractness (manipulated within lists) on RCL and RGN performance. Their RCL data for concrete words were similar to the results reported here. That is, the superiority in RCL by those who expected RCL over those who expected RGN was 10.9% for HF words but only 1.5% for LF words. (Since Miller et al., Note 3, reported their RGN data collapsed across the test-expectancy variable, we do not know whether the same pattern of data was obtained in RGN.) Toglia, Barrett, and Crothers (Note 4, Experiment 2) also jointly manipulated test expectancy (through the administration of practice lists involving only RCL or only RGN tests) and word frequency (between lists) and found that the superiority in performance by those who expected RCL over those who expected RGN was 4% larger for HF than for LF concrete words in both RCL and RGN. Given the appearance of this Test Expectancy × Word Frequency interaction across variations in the procedures used to induce test expectancy and the between- versus within-list nature of the word frequency manipulation, it would seem that this interaction is robust enough that it must be ac-

commodated by extant theories of RCL and RGN.[3]

As outlined in the introduction, Anderson and Bower's (1972, 1974) theory can accommodate this Test Expectancy × Word Frequency interaction. If one accepts both their assumptions about node tagging and the assumption that HF words generally activate more memory nodes than LF words, this interaction indicates that people who expect RCL associate list-marker elements to more different nodes corresponding to the TBR word than do people who expect RGN. However, this interaction does *not* provide information as to whether those expecting RCL associate list-marker elements either to a greater number of semantically distinct nodes corresponding to the TBR word (e.g., Martin, 1975; Reder et al., 1974) or to a greater number of nodes corresponding to different shadings of one semantic interpretation of the TBR word (see Anderson, 1976, pp. 390–406). The important point is that the greater variability in encoding by those who expect RCL over those who expect RGN is occurring in the *semantic* domain, associated with the TBR word, rather than occurring in the domain of the contextual list-marker elements (cf. Hintzman, 1974).

The finding that the superiority in RGN performance by those expecting RCL over those expecting RGN was equivalent for the BP and UP conditions is also important. In the UP condition, those induced to expect RGN were exposed to many lures during practice, whereas those induced to expect RCL were exposed to none. If the presentation of these lures produced proactive interference (PI) with memory for critical-list items, the superiority in memory performance by those who expected RCL over those who expected RGN in the UP condition could be due to differences in PI. However, people who expected RCL also did better than those expecting RGN in the BP condition, in which differences in test expectancy were not confounded with differences in PI, suggesting that the test-expectancy effect obtained in this condition was due to test-expectancy-induced differences in encoding rather than to differences in PI. This is important because prior to the present results and those of Neely and Balota (Note 1) the results had indicated that people expecting RCL do better in RGN than people expecting RGN only when differences in test expectancy are confounded with PI differences (Hall, Grossman, & Elwood, 1976, Experiment 1; Maisto, DeWaard, & Miller, 1977, silent-rehearsal condition; Poltrock & MacLeod, 1977, Experiment 1; Toglia et al., Note 4, Experiments 1 and 2) and not when these differences are not confounded (Hall et al., 1976, Experiment 3; Hall, Miskiewicz, & Murray, 1977; Naus, Ornstein, & Kreshtool, 1977; Miller et al., Note 3, Experiments 1, 2, and 4). However, the latter failures to find a test-expectancy effect in RGN might have been due to the fact that those expecting RCL received either no RCL practice lists or only one practice list each on RCL and RGN tests. Thus, the previous and present data suggest that when differences in test expectancy are not confounded with PI differences, people expecting RCL do better in RGN than people expecting RGN only if all subjects have developed and differentiated their test-specific encoding strategies through practice. Based on the available data, one must conclude that one practice list each on the RCL and RGN tests is insufficient practice for detecting a test-expectancy effect in RGN, whereas three practice lists each on the RCL and RGN tests are sufficient.

Implications for Word-Frequency Effects

As has been typically found (see Gregg, 1976, for a review), the present results indicate that HF words are better recalled than LF words, whereas the opposite is true in RGN. Even though the present experiment employed a mixed list containing both HF and LF words, superior RCL of HF words over LF words was still obtained (also see Smith, Glenberg, & Bjork, 1978, Experiment 4). Thus these data contradict Gregg's (1976, pp. 196–197) con-

[3] Although the Miller et al. and Toglia et al. Test Expectancy × Word Frequency interactions were not statistically significant, it should be pointed out that their experiments contained fewer than 25% of the number of observations made in the present study.

clusion that the superiority in RCL of HF words over LF words is obtained only with unmixed lists. Although it remains unclear what conditions are necessary for obtaining superior RCL of HF words over LF words in a mixed-list design (e.g., see Gregg, Montgomery, & Castaño, 1980, for mixed-list results contrary to the present results), the important point is that data now exist which indicate that Gregg (1976) was premature in concluding that the RCL superiority of HF words over LF words is confined to unmixed-list designs.

The fact that word frequency has a differential effect on RCL and RGN has been one of the empirical cornerstones of generate–recognize theories. This effect can be accommodated by these theories by arguing that the nodes corresponding to HF words are much more likely to be generated than are nodes corresponding to LF words. Thus, even though it may be more difficult to recognize a generated node corresponding to a HF word than to recognize a generated node corresponding to a LF word, there are so many more HF than LF nodes generated that the net result is superior RCL of HF words. Of course, one must also explain why LF words are more easily recognized than HF words. Such an explanation has recently been presented by Glanzer and Bowles (1976).

Glanzer and Bowles (1976), like Anderson and Bower, assume that nodes in memory are tagged with list-marker elements. In what Glanzer and Bowles call *direct tagging*, a subset of possible meanings available for a TBR word is sampled during its presentation, and all members of this subset are tagged with list-marker elements. This direct tagging process is akin to Anderson and Bower's node tagging. In *derivative tagging*, a node may be tagged with list-marker elements whenever a node associatively or semantically related to it is being directly tagged. This derivative tagging process is akin to Anderson and Bower's pathway tagging in that it is affected by the semantic relationship among the TBR words. However, it differs from pathway tagging in that it involves the association of list-marker elements to a node and not to pathways between nodes.

According to Glanzer and Bowles, in RGN a subset of the nodes corresponding to the to-be-recognized word is activated. The person then determines the proportion of nodes in the activated subset that were tagged with list-marker elements during list presentation, with no distinction being made between list-marker tags laid down by direct or derivative tagging. The larger the *proportion* of tagged nodes in the activated subset, the greater is the probability that the person will say that the word was in the list. RCL of a word, on the other hand, is based on the retrieval of a *single tagged node* corresponding to that word. In accounting for the word-frequency effect, Glanzer and Bowles make three assumptions: (a) HF words have more total nodes in memory available for activation and tagging than do LF words; (b) HF words are more likely to be semantically related to other list words and thus the nodes corresponding to HF words are more likely to be derivatively tagged; (c) the number of a word's nodes that are sampled and *directly* tagged during list presentation is a constant that is the same for HF and LF words.

What implications do the present data have for these assumptions? Since the superiority in RCL and RGN of those expecting RCL over those expecting RGN was much larger for HF words than for LF words, to account for the test-expectancy effect one must appeal to an encoding process that differentially affects performance on HF and LF words. In Glanzer and Bowles' theory, that process would be *derivative* tagging (see assumption b). However, two pieces of evidence argue against the suggestion that people expecting RCL do more derivative tagging than those who expect RGN. First, if people expecting RCL do more derivative tagging than those expecting RGN, the test-expectancy effect should be larger for semantically related TBR words than for unrelated TBR words. This prediction is based on the assumption that derivative tagging of a node can occur only when a semantically related node is being directly tagged. Thus, the idea that people who expect RCL do more derivative tagging than do people who expect RGN is contra-indicated by Neely and Balota's

(Note 1) results. They found that the superiority in memory performance by those expecting RCL over those expecting RGN was not affected by the semantic relatedness of the TBR words. The second piece of evidence against the notion that people expecting RCL do more derivative tagging than those expecting RGN comes from the present false-alarm data. Glanzer and Bowles have argued that the false-alarm rate to HF lures is higher than that to LF lures because (a) false alarms are based solely on derivative tags and (b) nodes corresponding to HF lures are more likely to be derivatively tagged (see assumption b above). We too found higher false-alarm rates for HF lures (15%) than for LF lures (10%), $F(1, 132) = 24.80$, $MS_e = 49.87\%^2$. However, the difference in false-alarm rates for those expecting RCL versus those expecting RGN was not larger for HF lures than for LF lures, as would be the case had those expecting RCL done more derivative tagging. In fact, there was less than 1% difference in false-alarm rates between those expecting RCL and those expecting RGN for both HF and LF lures. Thus, the Test Expectancy × Word Frequency interaction we obtained cannot be attributed to differences in derivative tagging as a function of test expectancy.

Since the data indicate that those expecting RCL are not doing more derivative tagging than those expecting RGN, within the framework of Glanzer and Bowles' theory one is led to conclude that they are doing more direct tagging. But if so, how can Glanzer and Bowles' theory accommodate the Test Expectancy × Word Frequency interaction we obtained? In order to accommodate this interaction and at the same time attribute it to an encoding effect, assumption c above must be modified. (Of course, one could retain assumption c and add new assumptions about other encoding mechanisms.) Apparently, assumption c holds only when people expect RGN (as was the case in Glanzer and Bowles' experiment). When people expect RCL, one must assume that they directly tag more nodes corresponding to HF words than to LF words. But even if one modifies the Glanzer-Bowles theory by making such an assumption, it can (with the appropriate choice of parameter values) still predict the word-frequency effect in RCL and RGN. Thus, the present analysis should be viewed as an extension of the Glanzer-Bowles theory rather than as a refutation of it.

Practical Implications

Let us now briefly consider the relevance these data have for classroom instruction and examinations. The most seductively obvious conclusion to be drawn from the present data is that one should always recommend that students study for an essay exam. However, two issues are raised by this extrapolation to the classroom. The first is whether the recommendation that the student study for an essay exam is any more than a recommendation that the student study harder (more). The present data indicate that test-expectancy effects are more complicated than this simple study-harder hypothesis would lead one to believe. If people expecting RCL merely study harder, one would expect either no Test Expectancy × Word Frequency interaction or one different from the one obtained here. More specifically, if the test-expectancy manipulation was merely a motivational one, people expecting RCL might study harder only on the words most likely to be forgotten in RCL, that is, the LF words. Thus, one version of an unembellished study-harder hypothesis incorrectly predicts that the superior memory performance of those expecting RCL should be larger for LF words than for HF words. Also, to the degree that people realize that LF words are easier to recognize than HF words, the present data rule out the contrary version of the study-harder hypothesis in which people study harder on the easy items that they believe they have some chance of remembering. If this were the case and if people expecting RGN realize that LF words are easier to recognize than HF words, those expecting RGN should have done better than those expecting RCL in recognizing LF words. However, note that these arguments merely discard the most simplistic versions of the study-harder hypothesis. Indeed, the interpretation that we have

made of our data within the framework of the Anderson-Bower theory can be construed as a more molecular account of a version of the study-harder hypothesis in which the people expecting RCL, in comparison to those expecting RGN, study harder in terms of episodically encoding more information about each TBR word's meaning.

A second, and more important, question concerns the generalizability of the present results to other situations and materials. According to our interpretation, a RCL test expectancy should facilitate memory performance only when the TBR material permits those expecting RCL to exercise a greater variability in their semantic encodings than those expecting RGN. In the present experiment we restricted this variability by using LF words as the TBR items; we found nearly equivalent performance on these items by those expecting RCL and those expecting RGN. Given this result, it is interesting that in those few cases in which a RCL test expectancy has failed to facilitate RCL performance, the potential for variability in semantic encoding has been minimized. For example, a RCL expectancy does not facilitate RCL of HF concrete TBR words when they share membership in a semantic category and are presented blocked-by-category in the list (Carey & Lockhart, 1973; Jacoby, 1973), nor does it facilitate the paired-associate RCL of HF response terms in paired-associate learning with HF stimulus terms (Lovelace, Note 5, Experiment 8). It can be argued that under such conditions even the people expecting RCL restrict their semantic encoding to those semantic features of the TBR word that are shared in common either with other TBR words from the same semantic category or with the stimulus term. The semantic encoding under these conditions by those expecting RCL might be functionally equivalent to the semantically impoverished encoding performed by those expecting RGN. Further support for this argument comes from the finding that the word-frequency effect in free RCL is eliminated when variability in the semantic encoding of HF TBR words is restricted either by the presence of an input cue that the person expects to serve as a retrieval cue at test (Reder et al., 1974, Experiment 1) or by incidental learning instructions (Dunlap & Dunlap, 1979). These latter two findings may be analogous to the reduction in the word-frequency effect that we obtained in RCL when variability in the semantic encoding of HF TBR words was restricted by the expectancy of a RGN test. Of course, acceptance of these post hoc arguments must await the collection of data that show that the magnitude of the test-expectancy effect obtained with HF concrete words is directly modulated by the degree to which the task or TBR list structure allows for variability in semantic encoding.

It is unfortunate that the reasoning in the preceding paragraph leads to the conclusion that test-expectancy effects might not be obtained for TBR materials of most interest to the educator, that is, coherent, fact-oriented prose passages. Such passages may too greatly restrict semantic encoding variability. Indeed, studies that have manipulated essay versus multiple choice test expectancies and used prose passages as the TBR materials have failed to obtain a statistically significant effect of test expectancy in RCL performance (e.g., Hakstian, 1971; Kulhavy, Dyer, & Silver, 1975). However, because the data on test-expectancy effects in prose RCL are sparse, we believe that any conclusion concerning the effect that test expectancy has on the retention of information learned in the classroom would be premature. We hope that additional research on test-expectancy effects can provide educators with a solid empirical basis for their development of classroom examination procedures that will maximize students' retention and retrieval of classroom information in problem-solving situations outside the classroom.

Reference Notes

1. Neely, J. H., & Balota, D. A. *Test-expectancy and semantic-organization effects in recall and recognition.* Manuscript submitted for publication, 1980.
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. (Draft available from James H. Neely, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907).

3. 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.
4. Toglia, M. P., Barrett, T. R., & Crothers, E. J. *Process differences in recall and recognition memory*. Paper presented at the meetings of the Psychonomic Society, Denver, Colorado, November 1975.
5. Lovelace, E. A. *Effects of anticipated form of testing on learning* (Final report, Project 2-C-019 on Grant OEG-3-72-0033). Office of Education in the U.S. Department of Health, Education, and Welfare, Washington, D.C., August 1973.

References

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. *Language, memory, and thought*. Hillsdale, N.J.: Erlbaum, 1976.

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. A propositional theory of recognition memory. *Memory & Cognition*, 1974, *2*, 406–412.

Carey, S. T., & Lockhart, R. S. Encoding differences in recognition and recall. *Memory & Cognition*, 1973, *1*, 297–300.

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

Dunlap, G. L., & Dunlap, L. L. Manipulating the word frequency effect in free recall. *Memory & Cognition*, 1979, *7*, 420–425.

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.

Gregg, V. Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition*. New York: Wiley, 1976.

Gregg, V. H., Montgomery, D. C., & Castaño, D. Recall of common and uncommon words from pure and mixed lists. *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 240–245.

Hakstian, A. R. The effects of type of examination anticipated on test preparation and performance. *Journal of Educational Research*, 1971, *64*, 319–324.

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.

Hintzman, D. L. Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium*. Hillsdale, N.J.: Erlbaum, 1974.

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.

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

Kintsch, W. *The representation of meaning in memory*. Hillsdale, N.J.: Erlbaum, 1974.

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

Kulhavy, R. W., Dyer, J. W., & Silver, L. The effects of notetaking and test expectancy on the learning of text material. *Journal of Educational Research*, 1975, *68*, 363–365.

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.

Martin, E. Generation–recognition theory and the encoding specificity principle. *Psychological Review*, 1975, *82*, 150–153.

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.

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.

Smith, S. M., Glenberg, A., & Bjork, R. A. Environmental context and human memory. *Memory & Cognition*, 1978, *6*, 342–353.

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.

Received November 11, 1979
Revision received April 14, 1980 ■