Metamemory for words and enacted instructions: Predicting which items will be recalled

RONALD L. COHEN

Glendon College, York University, Toronto, Ontario, Canada

Subjects can predict, during word list acquisition, which items they will subsequently recall. When the list items are subject-performed tasks (SPTs), recall prediction is totally inaccurate. Delaying the rating of recall probability (actually, trace strength rating) until list acquisition was complete did not affect the accuracy of the ratings. Word recall was predictable; SPT recall was not. When encoding was systematically manipulated, the effects of the manipulations on recall were accompanied by parallel effects on the strength ratings for SPTs, as well as for words. Several explanations for the data are discussed, the most probable being that of a mismatch between those characteristics of the SPTs that underlie strength ratings and those that determine recall.

One of the more interesting findings in metamemory research is that subjects possess the ability to predict, prior to test, which words they will subsequently remember. There are two basic procedures for studying this phenomenon. One of these, the feeling-of-knowing (FOK) procedure, requires subjects to make FOK judgments for non-recallable items, subsequently validating these judgments against performance in a recognition test (Hart, 1965, 1967). In summarizing FOK findings, Schacter (1983) described the FOK effect as being modest in size but reliable; subjects have some idea, but not too much, of which nonrecallable words they will later be able to recognize.

The second procedure, the feeling-that-I-will-know (FTIWK) procedure, requires subjects to predict, during list acquisition, which words they will later recall in a free recall test (Groninger, 1979; Lovelace, 1984a; Underwood, 1966) or in a paired associates test (Arbuckle & Cuddy, 1969; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984a, 1984b). This procedure also has been used in conjunction with the cued recall of sentences (Lovelace, 1984b) and in the recall of idea units from a text passage (Maki & Swett, 1987). The effects obtained in the FTIWK studies are not only reliable, but also rather substantial.

In addition to testing FTIWK for the free recall of words, Cohen (1983) tested FTIWK for minitasks enacted by the subjects (subject-performed tasks, or SPTs). Although he confirmed a positive FTIWK effect in the case of word recall, Cohen found no such effect for SPT recall. The subjects were quite unable to predict which SPTs they would recall on the subsequent free recall test. The main question raised by these findings is, of course, why subjects can predict the recall probabilities of individual words, but not of individual SPTs. The object

of the present study was to test some possible answers to this question.

The impact made by a word at acquisition probably depends on how the subject recodes it. Thus, if a word stimulus is elaborated or enriched by the subject, for example by the use of imagery or by relating the word to prior knowledge, then this could provide a basis for predicting probable recall. As Bäckman and Nilsson (1985) observed, individual SPTs are rich in and of themselves, and consequently are not subject to recoding. If this is indeed the case, then the subject may have no valid basis for judging how well any given SPT has been acquired. In short, successful prediction in word recall may reflect the immediate apprehension of recoding. Since SPTs neither require nor receive recoding, such a basis for prediction is lacking. This explanation refers specifically to what occurs during study.

One corollary of this explanation is that the advantage that words enjoy over SPTs in recall predictability may be confined to judgments made during acquisition. If the predictive judgments were to be made at a later stage, prediction of word and SPT recall might be equally accurate (or inaccurate). This was tested in Experiment 1, in which subjects studied a list of words or a list of SPTs. The list was presented a second time, and predictive judgments were made during this latter presentation: the subjects were asked to rate each item according to whether or not they judged it to be strong enough in memory to be recalled. The predictive power of the ratings was then validated against actual recall, either by having the same subjects recall the list, or by having a second subject sample study and recall the list. The use of this strength rating procedure not only enabled us to test the recoding apprehension hypothesis, but also facilitated the use of the acquisition manipulations introduced in Experiment 2. (The term *strength* is used in this discussion mainly because it provides a convenient label for the subjective rating scales used in the study. No particular theoretical significance is ascribed to the terms *strength* or *trace*

This research was supported by Grant A7023 from the Natural Sciences and Engineering Research Council of Canada. Send reprint requests to Ronald L. Cohen, Department of Psychology, Glendon College, 2275 Bayview Avenue, Toronto, Ontario M4N 3M6, Canada.

strength, although, on an intuitive level, it seems more likely that in rating strength subjects would rate the availability of an event's representation in memory, rather than its accessibility.)

A more general conclusion also could be drawn from Cohen's (1983) data, namely that adults have somehow managed to acquire a working knowledge of how their memory systems function in memory tasks involving unconnected words, but not in memory tasks involving enacted events. To test this general view of word and SPT memory, the conditions of encoding were systematically manipulated in Experiment 2, with the object of providing subjects with memory traces of varying strengths. The subjects' knowledge of their word and SPT memory subsystems could then be further examined by measuring the relationships between the strength ratings of word and SPT memory traces and recall, as these were mediated by the encoding manipulations.

EXPERIMENT 1

Three procedures were employed in this experiment. One of these was a replication of Cohen's (1983) procedure (within-subjects, simultaneous acquisition and rating, or WSSIM, procedure), in which each subject received one presentation of a list of items (words or SPTs). Immediately following each item's presentation, the subject predicted whether or not the item would be recalled at test, using a 4-point rating scale ranging from 1 (*this word/task is too weak in my memory for me to recall it*) to 4 (*this word/task is strong enough in my memory so I will be able to recall it*). When the total list had been presented, the subject was given 2 min for written recall of the items.

In the second procedure (within-subjects, successive acquisition and rating, or WSSUC, procedure), predictive ratings were made only after the total list had been encoded. The list was first presented as a study list, together with the information that a rating trial and a recall test would follow. Subsequent to the study presentation, the experimental items (words or SPT instructions) were re-presented, and the subject was asked to rate each one on a 4-point scale ranging from 1 (*this word/task is too weak in my memory for me to have recalled it*) to 4 (*this word/task is strong enough in my memory so that I could have recalled it*). Finally, the subject was given 2 min for written recall of the list.

In the WSSUC procedure, a second presentation of the list, the rating trial, intervened between the study trial and recall test. This second presentation could have functioned as an acquisition trial, in which case this procedure may not have wholly fulfilled its purpose of delaying the prediction ratings until after acquisition was complete. For this reason, a third procedure was included in the study (between-subjects, or BS, procedure), which employed four subject samples. Each sample was given a study trial, in which one list was presented together with the information that a memory test (unspecified) would follow.

Two of the samples received the word list, and the other two received the SPT list. For each list type, the study trial was followed by written free recall for one sample, and by the strength ratings in the other sample (see the above description of the WSSUC procedure).

Method

Subjects. A total of 160 undergraduate students served as subjects on a volunteer basis, each receiving \$3 (Canadian) as payment for participation. This sample was randomly split into eight groups, each of 20 students. Two groups were assigned to the WSSIM procedure, one group receiving the word list and the second group the SPT list; two groups were similarly assigned to the WSSUC procedure. The remaining four groups were assigned to the BS procedures, one to the word rating, a second to word recall, a third to SPT rating, and the fourth to SPT recall.

Materials. A list of words, 37 items long, was selected from the Paivio, Yuille, and Madigan (1968) lists. The first 3 and last 4 words in each list were considered buffer items; the remaining 30 words constituted the experimental items. One known correlate of recall probability, namely degree of concreteness, was included as a variable in the list construction, with half of the experimental words in each list being concrete (rated higher than 6.0 on the Paivio et al., 1968, lists) and half being abstract (rated lower than 3.0). Otherwise, the selection of list items was nonsystematic.

An SPT list, 37 items long, was constructed from SPTs selected from Cohen's (1981) pool. Corresponding with the division of the word lists into abstract and concrete words, the experimental items in each SPT list were divided into 15 items that required the use of extraenvironmental objects and 15 items that did not require such objects. Although the evidence is not unequivocal, some studies have found object-required items to be less well recalled than object-not-required SPTs (Cohen, Peterson, & Mantini-Atkinson, 1987).

In all lists, the 30 experimental items were sequenced in the same fixed random order across the three procedures.

Procedure. For the SPT procedures, the subject sat at a table, half of which was screened from view. Each SPT was presented by the experimenter's reading aloud an instruction for a task, which the subject then enacted. When the SPT required the use of an extraenvironmental object, such as a toothpick for the SPT *break the toothpick*, the object was given to the subject together with the verbal instruction. Immediately following the completion of the task, the object (in this case, the pieces of toothpick) were retrieved by the experimenter. All such extraenvironmental objects were kept behind the screen, hidden from the subject, except when actually in use. For the word procedures, the words were read aloud to the subject by the experimenter.

In the two procedures that did not require predictive ratings during acquisition, the rate of presentation was 1 item/3 sec in the case of the word lists and approximately 1 item/6 sec in the case of the SPT lists. In the procedure that required ratings during acquisition, these rates decreased somewhat, to approximately 1 item/6 sec for words and 1 item/8 sec for SPTs.

When the prediction ratings required a separate trial, the 30 critical items were re-presented in a second fixed random order, preceded by two of the list's buffer items. Each item (word or SPT instruction) was read aloud by the experimenter, following which the subject rated the probability of its recall, this normally taking about 4 or 5 sec. At test, the subject was allowed 2 min for the written recall of the list items.

Results and Discussion

The performance data are given in Tables 1 and 2. The significance of all statistical tests was evaluated using the .05 probability level.

Table 1
Proportion of Items Recalled Under the Three Procedures in Experiment 1

	Words		Subject-Performed Tasks	
	Concrete	Abstract	EEO Required	EEO Not Required
WSSIM	.31	.20	.36	.40
WSSUC	.38	.28	.44	.47
BS	.26	.21	.35	.37

Note—WSSIM = within-subjects, simultaneous acquisition and rating procedure. WSSUC = within-subjects, successive acquisition and rating procedure. BS = between-subjects procedure. EEO = extraenvironmental object.

Recall (Table 1). The concrete words were recalled significantly better than the abstract words [$F(1,57) = 24.67$, $MSe = 2.27$]. In addition, there was a main effect of procedure on word recall [$F(2,57) = 11.15$, $MSe = 1.76$]. Using the least significant difference (LSD) test, this effect was traced to the superiority of the WSSUC procedure over the other two procedures. It would appear that the second presentation of the list, during the rating phase of the WSSUC procedure, had a facilitatory effect on word recall. The interaction between the two main variables was not significant [$F(2,57) = 1.64$, $MSe = 2.27$].

For the SPTs, the items requiring the use of extraenvironmental objects tended to be recalled more poorly than those that did not require such objects, although this tendency was not significant [$F(1,57) = 3.12$, $MSe = 1.67$, $p < .10$]. As in the case of word recall, procedure also had an effect on SPT recall [$F(2,57) = 7.66$, $MSe = 2.95$], which was again traceable to the superiority of the WSSUC procedure over the other two procedures. Again, the interaction between the two main variables was not significant ($F < 1.0$).

Comparisons were also made between SPT recall, collapsed across the extraenvironmental object required/not required variable, and word recall. In all three procedures, SPT recall proved to be significantly greater than recall of the concrete or the abstract words [all $t(38)$ values ≥ 2.24].

Strength ratings (Table 2). The pattern of the word ratings mirrored that of the recall data. The concrete words were rated significantly higher than the abstract words [$F(1,57) = 48.26$, $MSe = .23$]. There was also a main effect of procedure [$F(2,57) = 13.83$, $MSe = .47$], although in this case the WSSIM procedure was the odd man out. The subjects apparently had less confidence in their ability to recall words when the ratings were made in conjunction with acquisition than when the ratings were made during the list presentation given subsequent to ac-

quisition. The interaction between the two main variables was not significant [$F(2,57) = 1.70$, $MSe = .23$].

In the case of the SPTs, the extraenvironmental object required/not required variable did not significantly affect the ratings [$F(1,57) = 3.52$, $MSe = .06$, $p < .10$]. It should be noted, however, that the recall and ratings trends are in opposite directions. The object-not-required SPTs showed higher recall but lower ratings than the object-required SPTs. There was also an effect of procedure on the SPT ratings [$F(2,57) = 29.02$, $MSe = .42$]. As with the words, this effect could be attributed to the WSSIM procedure's yielding lower ratings than the other two procedures. In addition, the BS procedure yielded significantly lower ratings than the WSSUC procedure. Because these two conditions were similar with respect to acquisition and rating procedures, this difference may be due to a sampling anomaly. The interaction between the two main variables was once again nonsignificant ($F < 1.0$).

Comparisons between the SPT ratings, collapsed across the extraenvironmental object required/not required variable, and word ratings reflected some of the trends found in the recall data. The SPT ratings were greater than those of the abstract words, significantly so in the case of the WSSIM and WSSUC procedures [$t(38) = 4.15$ and $t(38) = 4.11$, respectively], but not in the case of the BS procedure [$t(38) = 1.81$, $p < .10$]. With respect to the concrete words, the results were inconsistent. The ratings for the SPTs were higher than those for the concrete words in the WSSUC procedure, although not significantly so [$t(38) = 1.66$]. This trend more or less disappeared in the BS procedure ($t < 1.0$), and reversed in the WSSIM procedure, in which the concrete words received significantly higher ratings [$t(38) = 2.20$].

The aspects of the above-described data that are directly relevant to the aim of this study are the parallel effects of the concrete/abstract variable on word recall and rat-

Table 2
Mean Strength Ratings Made Under the Three Procedures in Experiment 1

	Words		Subject-Performed Tasks	
	Concrete	Abstract	EEO Required	EEO Not Required
WSSIM	2.72	1.88	2.47	2.38
WSSUC	3.21	2.72	3.54	3.48
BS	3.12	2.78	3.19	3.09

Note—Maximum rating = 4.00. WSSIM = within-subjects, simultaneous acquisition and rating procedure. WSSUC = within-subjects, successive acquisition and rating procedure. BS = between-subjects procedure. EEO = extraenvironmental object.

ings, and the opposing tendencies exerted by the extra-environmental object required/not required variable on SPT recall and ratings. Thus the performance data provide some evidence for metamemorial capability in the word recall but not in the SPT recall.

To test the hypotheses described in the introduction, correlational analyses were performed. The subjects within each group were numbered according to their order of arrival in the laboratory. Each group was divided into two subgroups, one comprising the 10 odd-numbered subjects and the other the 10 even-numbered subjects. For each list the mean rating value and the mean recall value were calculated for each of the 30 experimental items for the two subgroups, and split-half reliabilities were calculated. In all cases, these reliabilities proved to be significant, ranging from .48 to .72 for the SPTs, and from .41 to .75 for the words, except in the simultaneous acquisition and rating condition, where the reliability for the word ratings reached 0.96.

Table 3 shows the ratings-recall correlations. The results are quite consistent across the three replications. The strength ratings were predictive of recall frequency in the case of the words, but not in the case of the SPTs.

As already described, the WSSUC procedure yielded significantly higher recall levels than the WSSIM and BS procedures, for both words and SPTs, probably because the separate rating trial in the WSSUC procedure provided an extra opportunity for acquisition. However, the similarity of the correlational data patterns suggests that the three procedures provided comparable measures of trace strength and recall.

As may also be seen in Table 3, partialling out the abstract/concrete and object required/not required variables for the recall-ratings correlations of the words and SPTs, respectively, had little effect on these measures, although the coefficient for the words in the WSSIM and WSSUC replications dropped just below the .05 significance level.

The results of the correlational analyses are in complete agreement with those reported by Cohen (1983), namely that subjects have some knowledge at acquisition of which words, but not of which SPTs, they will subsequently recall. In addition, these analyses have demonstrated that this pattern of results is not confined to ratings made during acquisition, but also extends to ratings made subse-

quent to acquisition. Consequently, the encoding impact explanation for Cohen's (1983) data, described in the introduction, would appear to be inadequate. The ability to predict the future recall of words, but not of SPTs, does not simply depend on the immediate apprehension of word recoding during acquisition.

A second explanation for these data, which is an implicit component of the encoding impact hypothesis, involves an extension of the notion that the SPTs do not systematically differ in their encoding impact, because they are all so inherently rich. The obtaining of reliable ratings, not only from the WSSIM procedure, but also from the other two procedures, suggests that there are systematic differences not only in the impact the various SPTs make at encoding, but also in the strengths of the resulting traces. What has still to be explained, then, is why these trace strength differences are not translated into differences in recall frequency.

As mentioned in the introduction, it makes intuitive sense to equate strength ratings with availability. If this assumption is valid, then one possible explanation is that the differences in SPT trace availability are offset by differences in trace accessibility at retrieval. Some evidence for this supposition is provided by the effect of the object required/not required variable in Experiment 1. There was a general tendency for subjects to rate SPTs that involved the use of an extraenvironmental object, such as a deck of cards or a toothpick, as being stronger in memory than those that did not. The involvement of such objects could well increase the distinctiveness (or strength) of SPT traces. In recall, the SPTs that involved extra-environmental objects generally tended to be recalled more poorly than SPTs that did not. A significant effect of the object required/not required variable on recall was reported previously by Cohen et al. (1987), who suggested that in the case of a no-object SPT such as *point to the door*, the *door* is a permanent feature of the experimental environment and can subsequently act as a retrieval cue. In the case of an object SPT such as *cut the cards*, the cards are removed from view as soon as the SPT is complete, and consequently provide no environmental support at recall. In this scenario, any superiority the traces of object SPTs might possess in distinctiveness, and therefore in rated availability, would be offset by their lower accessibility. It should be noted, however, that partialling out the object required/not required variable did not increase the magnitude of the recall-ratings correlations. Consequently, if the SPT results are to be explained in terms of a trade-off between availability and accessibility features, the basis of this trade-off must involve a more general mechanism than that just described.

Another explanation of the above data that deserves consideration is that subjects are generally knowledgeable about how their word memory subsystem, but not their action memory subsystem, functions. In speculating on some mechanisms that could serve as a basis for making FOK judgments, Nelson, Gerler, and Narens (1984) distinguished between trace access mechanisms, where judg-

Table 3
Correlation Coefficients (*r*) Between Group Mean Ratings and Group Mean Recall Frequencies, for the Sets of 30 Experimental Items Comprising Each List in Experiment 1

Condition	Words		Subject-Performed Tasks	
	<i>r</i>	<i>r_p</i>	<i>r</i>	<i>r_p</i>
WSSIM	.44†	.35*	-.19	-.06
WSSUC	.46†	.34*	-.38†	-.38†
BS	.56†	.49†	-.11	.02

Note—The *r_p* values are correlation coefficients between the ratings and recall frequencies with the abstract/concrete and the object required/not required variables partialled out for the words and SPTs, respectively. WSSIM = within-subjects, simultaneous acquisition and rating procedure. WSSUC = within-subjects, successive acquisition and rating procedure. BS = between-subjects procedure. **p* < .10. †*p* < .05.

ments are made according to the state of the actual trace, and inferential mechanisms, where extratrace information is used. Thus the subjects in Experiment 1 may have known, for example, that concrete words are generally better recalled than abstract words and used this knowledge to make their rating judgments. In fact, one finding in Experiment 1 suggests that this was indeed the case, namely that the abstract/concrete variable produced a much greater effect on the ratings than on recall. Point-biserial correlations between the abstract/concrete variable and the ratings were .88, .60, and .61 for the WSSIM, WSSUC, and BS procedures, respectively. When recall frequencies were substituted for the ratings, the corresponding correlation coefficients were only .36, .34, and .31. That the subjects in Experiment 1 put more faith in the abstract/concrete variable when rating trace strength than was justified by the actual recall frequencies may well be an indication that extratrace information was used in rating trace strength. A similar conclusion was drawn by Groninger (1979) with regard to word frequency, as well as to word concreteness, information.

Roughly equivalent to the trace/extratrace distinction are Koriat and Lieblich's (1974) use of the distinction between differential detection and class detection in the context of the tip-of-the-tongue (TOT) phenomenon and, more pertinent to our study, Lovelace's (1984b) distinction between idiosyncratic and normative components of predictive ratings in the FTIWK procedure. *Normative information* refers to general knowledge of the memorial properties of items shared by the subjects in a sample (e.g., the knowledge that concrete words tend to be recalled better than abstract words). *Idiosyncratic knowledge* refers to the specific knowledge that any specific subject has about the state of a specific item's representation in memory. This knowledge is not shared by the other members of the sample.

Some light can be shed on the possible use of idiosyncratic and normative information in the within-subjects procedures by analyzing the rating data across subjects instead of across items. A measure that takes into account individual differences in the use of the rating scales is the predictive accuracy quotient, or PAQ (see Lovelace, 1984b). For any given subject, this score is defined as the difference between the mean predictive rating for items recalled and the mean predictive rating for items not recalled, divided by the square root of the pooled variance of the subject's ratings on the recalled and non-recalled items. The group mean PAQ scores in the case of the word lists were 0.41 (WSSUC) and 0.35 (WSSIM). Both of these scores are significantly greater than zero ($p < .01$), which means that the subjects' strength ratings could effectively discriminate between words subsequently recalled and those not recalled. For the SPT lists, the corresponding PAQ scores were -0.05 and -0.03 , which indicate a complete failure of the subjects' ratings to predict subsequent recall.

In addition to calculating PAQ scores for the individual subjects, Lovelace (1984a, 1984b) calculated PAQ scores in pairs of subjects randomly yoked together. In this

procedure, the predictions of one subject under the yoke (S_A) are validated against the recall of the second subject (S_B). The PAQ scores are then defined as the difference between the mean predictive ratings of S_A for items recalled by S_B and the mean ratings of S_A for items not recalled by S_B , divided by the square root of the pooled variance for S_A 's ratings. Lovelace (1984b) argued that if the success of recall predictions depends on "basic properties intrinsic to the materials and common to all individuals" (p. 763), then the PAQ scores calculated by the yoked pairs procedure should not differ too much from those calculated within individuals. In the present instance, if yoked PAQ scores for strength ratings proved to be significantly greater than zero, this would indicate that at least some of the predictive power of the strength ratings depends on normative (extratrace) information. The yoking procedure, in fact, yielded mean group PAQ scores for the word lists of -0.02 (WSSUC) and 0.00 (WSSIM). This manifest decrease in PAQ scores that accompanied the yoking procedure replicates Lovelace's (1984a, 1984b) findings. In the case of the SPT conditions, neither group mean yoked PAQ score differed significantly from zero, (-0.09 and -0.14 , respectively).

These observations give rise to two obvious conclusions. First, the two within-subjects conditions provided similar data and very probably involve the same mechanisms. Second, these results strongly suggest that although the subjects appeared to use extratrace information in making their ratings, the predictive value of the word strength ratings for subsequent recall depended entirely on idiosyncratic, or trace, information.

The above conclusions refer only to recall-trace strength dependencies as these vary across list items. Experiment 2 was designed to explore the generality of these conclusions by testing whether encoding manipulations, known to have systematic effects on recall level, would show corresponding effects on strength ratings for SPTs as well as for words. This shift in emphasis from inter-item differences to the effect of manipulated variables entailed the use of analysis of variance as the main method of data treatment in Experiment 2.

EXPERIMENT 2

The ability of subjects to judge strengths of SPT memory traces, as compared with strengths of word memory traces, was again measured. To obtain event memory representations of varying strength, two encoding manipulations were employed for different groups of subjects. One manipulation involved frequency of presentation, and the other involved divided attention.

In the frequency-of-presentation procedure, a list of events (words or SPTs) was presented, some of which occurred once in the list and some twice. Not only has this variable been designated a strength variable (Jacoby, Bartz, & Evans, 1978; Wickelgren, 1981), but, more important, comparable mnemonic effects of repeating list items have been demonstrated for words and SPTs (Cohen, 1985). Furthermore, Zechmeister and Shaughnes-

sey (1980) demonstrated that, in the case of word lists, the advantage enjoyed by repeated items in recall is matched by higher prediction ratings during acquisition.

Following list presentation, subjects were re-presented with the list items and were required to judge whether or not each event was strong enough to be recalled. Finally, the subjects were asked to free recall the list. It was expected that the subjects would recall more twice-presented than once-presented items from both word and SPT lists. Furthermore, it was expected that subjects would rate the twice-presented word traces as being stronger than the once-presented word traces. The data of most interest, however, were the ratings for the SPTs. Would the subjects be capable of discriminating between strong and weak SPT traces, and rate twice-presented SPT traces as being reliably stronger than once-presented SPT traces?

For the divided-attention manipulation, half of the list (words or SPTs) was presented when the subject was engaged in counting backward by threes, and the other half of the list was presented in the absence of counting. The assumption that the performance of a concurrent task during acquisition would yield weak memory traces is supported by previous word memory studies (Anderson & Craik, 1974) and SPT memory studies (Bäckman, Nilsson, & Chalom, 1986), both of which showed that requiring a concurrent task during list presentation decreased level of recall. Again, it was expected that the optimal encoding condition (no subsidiary counting task) would yield higher recall levels for both words and SPTs, and also higher ratings in the case of the words. As before, the effect of the encoding manipulation on the SPT strength ratings was of prime interest.

Method

Subjects. Forty undergraduate students volunteered to participate as subjects in return for payment of \$3 (Canadian).

Design and Materials. For the frequency-of-presentation manipulation, 23 SPTs were selected from Cohen's (1981) pool. Of these, 7 items acted as buffer items (3 primacy and 4 recency), and the remaining 16 items constituted the experimental items. A list was constructed that consisted of the 3 primacy items, 8 experimental items that occurred once intermingled with 8 experimental items that occurred twice, and the 4 recency items. The occurrences of the twice-presented items were separated by a lag of 3 to 6 other items. A counterbalanced list was also constructed, which included the same 23 SPTs, but in which the 8 items that were presented once in the original list were now presented twice, and vice versa. Two 23-item word lists were constructed using the same procedures, employing two-syllable words having frequencies of 20 or more occurrences per million, selected from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982).

For the divided-attention manipulation, 23 SPTs were selected from Cohen's (1981) pool to form a list. Again, the first 3 and last 4 items were considered to be buffer items. Counterbalancing was achieved in this case by having half of the subjects count backward during the presentation of the first portion of the list (3 primacy and 8 experimental items) and the other half of the subjects count backward during the presentation of the second portion of the list (8 experimental and 4 recency items). A 23-item list was also constructed for use in the word memory condition, utilizing two-syllable words having frequencies of 20 or more occurrences per million.

Twenty subjects were tested with each of the two experimental manipulations, each subject receiving a word list and an SPT list. A further counterbalancing was included in the experimental design by testing half of the subjects with a word list first and an SPT list second, and testing the other half of the subjects using the reverse order.

Procedure. For the frequency-of-presentation manipulation, the list of items was presented, followed by an intervening task (two digit span lists). Subsequent to this, the 16 experimental items (words or SPT instructions) were read in a new fixed random order, and the subject was instructed to rate each item on a 4-point scale ranging from 1 (*this word/task is too weak in my memory for me to have recalled it*) to 4 (*this word/task is strong enough in my memory so that I could have recalled it*). Finally, the subject was given a fresh sheet of paper and asked to write down recalled list items.

For the divided-attention manipulation, a card bearing a three-digit number was held up to the subject's view. The subject was required to count backward by threes from this number, aloud, as long as the card was exhibited. For half of the subjects, the card was introduced right before the presentation of the 1st list item and removed right after the presentation of the 11th list item. For the other half of the subjects, the card was introduced during a short pause between the 11th and 12th list items and removed following the presentation of the final list item. The remainder of the divided-attention procedure followed that used in the frequency-of-presentation manipulation. Apart from these modifications, the procedure was basically the same as that used in the WSSUC condition in Experiment 1, in that the experimenter read aloud the words or SPT instructions. The approximate rates of presentation were 1 word/3 sec and 1 SPT/6 sec.

Results

The recall data for Experiment 2 are given in Table 4. In accordance with the earlier findings described in the introduction, two presentations yielded higher recall than one presentation, for both words and SPTs, and simultaneous counting yielded lower recall than did the absence of this competing task, again for both classes of items.

The two counterbalancings described in the Method were included in all of the analyses of variance (ANOVAs) reported in this section. In no case, however, did these variables interact significantly with the experimental variables.

For the frequency-of-presentation data (shown in Table 4), an ANOVA yielded a significant effect of number of item presentations [$F(1,16) = 20.24$, $MSe = 1.67$], a significant effect of class of item [$F(1,16) = 21.56$, $MSe = 1.69$], but no significant interaction between these two factors [$F(1,16) < 1.0$].

Table 4
Mean Proportion of Items Recalled Under the Once-Presented and Twice-Presented Conditions, and Under the Simultaneous Counting and Noncounting Conditions

Condition	Words	Subject-Performed Tasks
Frequency-of-Presentation Manipulation		
Once-Presented	.33	.51
Twice-Presented	.50	.66
Divided-Attention Manipulation		
Counting	.18	.45
Noncounting	.43	.55

For the divided-attention data (see Table 4), the simultaneous counting proved to have a significant main effect [$F(1,16) = 40.52$, $MSe = .96$], as did class of item [$F(1,16) = 27.80$, $MSe = 1.40$]. The interaction between the counting/noncounting variable and class of item showed a nonsignificant trend that exceeded the .10 level [$F(1,16) = 3.45$, $MSe = 2.12$].

Although neither analysis really warranted posttesting the encoding effect for each class of item separately, this was done to provide as stringent a test for the experimental predictions as possible. LSD tests applied to the data in Table 4 showed that both encoding manipulations significantly affected word recall. In the case of the SPTs, repetition showed a significant effect on recall level, with simultaneous counting showing a trend that exceeded the .10 level.

Table 5 shows the relationship between the ratings and the experimental variables. These data were obtained by calculating, for each subject, the mean ratings for the items in each of the eight cells of the matrix obtained by orthogonally combining the two classes of item (words vs. SPTs), the two levels of the encoding manipulation (one vs. two presentations, or counting during acquisition vs. no counting), and whether or not the items were subsequently recalled. These data follow the same pattern as the recall data shown in Table 4. The SPT ratings were higher than the word ratings in all eight comparisons, the favored encoding conditions yielded higher ratings in all eight comparisons, and the recalled words received substantially higher ratings than the nonrecalled words. Although the recalled SPTs also received higher ratings than the nonrecalled SPTs, the differences in this case were minimal.

Because of the experimental manipulations used in Experiment 2, a modified PAQ procedure was used to analyze the data. To take account of differences in the use of the rating scales, both between individuals and between the words and SPTs, separate ANOVAs were performed for the two types of items. The scores used in each repeated measures ANOVA were the mean individual ratings in each cell of the 2 (levels of encoding) $\times$ 2 (recalled or not recalled) matrix, divided by the square roots of the pooled variance of the relevant ratings. The results of the

analyses proved to be very similar for the two encoding procedures. The twice-presented items were rated significantly higher than the once-presented items [$F(1,16) = 8.16$, $MSe = .27$, for words, and $F(1,16) = 14.16$, $MSe = .68$, for SPTs]. The noncounting conditions produced significantly higher ratings than the counting conditions [$F(1,16) = 7.36$, $MSe = .28$, for words, and $F(1,16) = 7.61$, $MSe = 1.32$, for SPTs]. For both words and SPTs, then, the conditions favored in recall—two presentations and noncounting—produced significantly higher ratings.

The second main variable, namely recalled versus non-recalled items, proved to be significant in the case of the word lists [$F(1,16) = 5.66$, $MSe = .51$, and $F(1,16) = 84.63$, $MSe = .11$, for the multiple-presentation procedure and counting procedure, respectively]. The corresponding F values for the SPTs were less than 1.0 in both cases. The interactions between the two main variables were also less than 1.0 in all four analyses. Thus, the finding obtained in Experiment 1 that subjects' subsequent recall of the various list items could be predicted from their trace strength ratings for words but not for SPTs was replicated in Experiment 2 when the ratings were divorced from the influence of the encoding variables.

Application of the yoked-subjects procedure also yielded results similar to those obtained in Experiment 1, namely that the effect of the recalled/nonrecalled variable on the ratings was nonsignificant for the word lists (F values ≤ 1.15) as well as for the SPT lists (F values ≤ 2.01). Thus, this procedure has again provided data that are consistent with an idiosyncratic interpretation of the strength rating-recall relationship found with the word lists.

The other main variables, acquisition manipulations, retained their significance in the yoked procedure. This finding does not, of course, tell us anything about the idiosyncrasy of the relationship between the ratings and recall, as mediated by the experimental manipulations.

In summary, the frequency-of-presentation manipulation produced clear effects on recall for both words and SPTs. These effects were paralleled by effects on the strength ratings for both classes of item. Simultaneous counting clearly lowered word recall, and showed a similar trend in SPT recall. Again, the strength ratings for both classes of item were significantly affected by the encoding variable. It may be concluded, therefore, from the results of Experiment 2 that the encoding manipulations determined the strengths of the event representations in memory, which in turn affected recall levels. It may also be concluded that these systematic strength differences were detectable by the subjects in the case of the SPTs as well as in the case of the words. The poor metamemory for SPTs shown by the subjects in Experiment 1 does not appear to be a general characteristic of the human condition.

GENERAL DISCUSSION

The results of the two experiments appear to rule out four explanations for Cohen's (1983) findings. First, the notion that words vary in their encoding impacts or in their

Table 5
Mean Ratings for Recalled and Nonrecalled Items Under the Once-Presented and Twice-Presented Conditions, and Under the Simultaneous Counting and Noncounting Conditions

Condition	Words		Subject-Performed Tasks	
	Recalled	Nonrecalled	Recalled	Nonrecalled
Frequency-of-Presentation Manipulation				
Once-Presented	2.81	2.20	3.02	2.86
Twice-Presented	3.21	2.74	3.47	3.40
Divided-Attention Manipulation				
Counting	2.84	1.69	3.18	3.05
Noncounting	3.38	2.34	3.53	3.32

Note—The maximum rating (greatest strength in memory) = 4.0.

trace strengths, whereas SPTs do not, appears to be invalid. Interitem differences in ratings proved to be just as reliable for the SPTs as for the words. Second, the explanation of the data in terms of the apprehension of differences in encoding impact during acquisition is insufficient at best. Strength ratings for more established traces are just as good predictors of word recall as are ratings made during trace formation. Third, the possibility that subjects are generally more knowledgeable about the functioning of their word memory than of their action event memory is discounted by the results of Experiment 2. And, fourth, the yoked-subjects technique failed to provide any evidence that extratrace information contributed to the predictive power of the word ratings. Consequently, the findings of Experiment 1 cannot be explained in terms of the subjects' superior knowledge of the normative properties of words, as compared with SPTs.

Having discarded the above possibilities, we appear to be left with the conclusion that the strength ratings did not predict the recall frequencies of the SPTs because of a mismatch between the characteristics of the items (or item traces) that are utilized in making ratings and those that determine recall probability. One possible mismatch scenario involving the use of environmental cues has already been considered in Experiment 1, under the assumption that ratings reflect availability and recall reflects accessibility. A variation on this theme involves the assumption that whereas ratings reflect the strengths of individual item traces, recall reflects organization among the items. This hypothesis runs into a serious problem, however. If this assumption is valid for SPTs, it should hold equally for words, and we are then faced with the question of why the ratings were predictive in the case of word recall.

A third alternative is that ratings and recall are based on different characteristics of the SPTs, irrespective of environmental retrieval cues or organization. Although little is known about the properties of action events that determine their recall probabilities, these properties appear to be attached to the action components (Cohen et al., 1987), and more specifically to the goals of the actions (Cohen & Heath, 1988). Whatever the subjects used to discriminate among the SPTs when making strength judgments, it was unlikely to have been the goals of the actions. On this basis, then, the notion of a mismatch between ratings and recall characteristics of SPTs is quite credible. The positive correlations between the ratings and the recall of words signify some degree of matching between the characteristics of the words involved in these two memory tests. As already indicated, this match may have been mediated by variations in the recoding of the words.

When experimental manipulations are involved, such as those employed in Experiment 2, the picture changes somewhat. For example, the increase in recall probability that accompanies repeated presentations is clearly not

related to any specific characteristic of SPTs, such as the action goals, but to some more general aspect of multiple presentations, such as the multiplicity of information in the traces. This information may be utilized in the rating test, regardless of whether the list items are words or SPTs.

Finally, it might be worthwhile to draw a more general conclusion from our data, albeit a somewhat tentative one. Obviously, interitem differences in recall probability can depend on a number of factors. For example, there are several different experimental manipulations that can affect the makeup of the traces and, consequently, the recall level. Such variations in traces may be detected, and utilized, in making strength judgments, irrespective of the nature of the list items (Experiment 2). In the absence of an experimental manipulation, there may still be differences in the nature and degree of recoding, accorded to the various items. These recoding differences are both detectable by the subject in a strength rating test and predictive of recall probability. It is likely that verbal items are subject to such recodings, whereas SPTs are not (Experiment 1). When there are no interitem differences in recoding to guide strength ratings, the subjects base their ratings on other characteristics of the traces. Although the bases for these ratings are not obvious, it is clear that they are quite different from those aspects of the traces that determine recall.

REFERENCES

- ANDERSON, C. M. B., & CRAIK, F. I. M. (1974). The effect of a concurrent task on recall from primary memory. *Journal of Verbal Learning & Verbal Behavior*, 13, 107-113.
- ARBuckle, T. Y., & CUDDY, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, 81, 126-131.
- BÄCKMAN, L., & NILSSON, L.-G. (1985). Prerequisites for lack of age differences in memory performance. *Experimental Aging Research*, 11, 67-73.
- BÄCKMAN, L., NILSSON, L.-G., & CHALOM, D. (1986). New evidence on the nature of the encoding of action events. *Memory & Cognition*, 14, 339-346.
- COHEN, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, 22, 267-281.
- COHEN, R. L. (1983). The effect of encoding variables on the free recall of words and action events. *Memory & Cognition*, 11, 575-582.
- COHEN, R. L. (1985). On the generality of laws of memory. In L.-G. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 247-277). Hillsdale, NJ: Erlbaum.
- COHEN, R. L., & HEATH, M. (1988). Recall probabilities for enacted instructions. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 421-426). London: Wylie.
- COHEN, R. L., PETERSON, M., & MANTINI-ATKINSON, T. (1987). Inter-event differences in event memory: Why are some events more recallable than others? *Memory & Cognition*, 15, 109-118.
- FRIENDLY, M., FRANKLIN, P. E., HOFFMAN, D., & RUBIN, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14, 375-299.
- GRONINGER, L. D. (1979). Predicting recall: The "feeling-that-I-will-know" phenomenon. *American Journal of Psychology*, 92, 45-58.

- HART, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*, 208-216.
- HART, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning & Verbal Behavior*, *6*, 685-691.
- JACOBY, L. L., BARTZ, W. H., & EVANS, J. D. (1978). A functional approach to levels of processing. *Journal of Experimental Psychology: Human Learning & Memory*, *4*, 331-346.
- KING, J. F., ZECHMEISTER, E. B., & SHAUGHNESSY, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, *83*, 329-343.
- KORIAT, A., & LIEBLICH, I. (1974). What does a person in a "TOT" state know that a person in a "don't know" state doesn't know? *Memory & Cognition*, *2*, 647-655.
- LOVELACE, E. A. (1984a). Metamemory: Monitoring future recallability in free and cued recall. *Bulletin of the Psychonomic Society*, *22*, 497-500.
- LOVELACE, E. A. (1984b). Meta-memory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 756-766.
- MAKI, R. H., & SWETT, S. (1987). Meta-memory for narrative text. *Memory & Cognition*, *15*, 72-83.
- NELSON, T. O., GERLER, D., & NARENS, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, *113*, 282-300.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, *76*(Pt. 2).
- SCHACTER, D. L. (1983). Feeling of knowing in episodic memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *9*, 39-54.
- UNDERWOOD, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology*, *71*, 673-679.
- WICKELGREN, W. A. (1981). Human learning and memory. *Annual Review of Psychology*, *32*, 21-52.
- ZECHMEISTER, E. B., & SHAUGHNESSY, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, *15*, 41-44.

(Manuscript received March 5, 1987;
revision accepted for publication January 13, 1988.)