*Journal of Experimental Psychology:
Human Learning and Memory*
1976, Vol. 2, No. 4, 349-361

# Encoding Specificity: Relation Between Recall Superiority and Recognition Failure

Sandor Wiseman and Endel Tulving
*University of Toronto, Toronto, Canada*

The results of four experiments show that (a) recall superiority over recognition is reversed by the use of unrelated word pairs in the study list, and (b) the reversal of recall superiority leaves intact the phenomenon of recognition failure of recallable words. These results extend the generality of encoding specificity and suggest that although recall superiority is a sufficient condition for recognition failure of recallable words it is not a necessary condition.

In memory experiments in which people are required to remember the items of a list, the most effective retrieval cues are almost always nominal copies of to-be-remembered items. If such cues are ineffective, it is unlikely that any other cues will succeed. Under certain experimental conditions, however, copies of list items are *not* the most effective cues for retrieval—in the presence of other cues, people do recall learned list words that they fail to recognize.

This phenomenon of recognition failure of recallable words has been manifested in a number of experiments in which the same procedure, with only minor variations, has been used (e.g., Tulving, 1974; Tulving & Thomson, 1973; Watkins & Tulving, 1975; Wiseman & Tulving, 1975). In these experiments, learners studied a list of pairs of associatively related words. One word of each pair was designated as the to-be-remembered (TBR) or target word; the other word was studied with the expectation that it would serve as an aid for the recall of the target. Following the presentation of the study list, and usually after some interpolated activity, the learner was given two successive retention tests. The first of these was a free-choice recognition test in which the learner was provided with a test sheet containing copies of target words together with distractor words and was asked to identify the target items. The second was a typical paired associate test in which the list cues were given as aids for the recall of target words.

The phenomenon of recognition failure of recallable words accords with neither common sense nor extant theory of memory. Why, in a recall test, can a learner produce an item that he or she does not recognize? One interpretation of this phenomenon has been suggested by Tulving and Thomson (1973) in terms of the encoding specificity principle: A TBR item is encoded with respect to the context in which it is studied, producing a unique trace which incorporates information from both target and context. For the TBR item to be retrieved, the cue information at the time of retrieval must appropriately match the trace of the item-in-

This research was supported by National Research Council of Canada Grant A8632.

Requests for reprints should be sent to Endel Tulving, Department of Psychology, University of Toronto, Toronto M5S 1A1, Ontario, Canada.

context. Because of its unique features, the trace cannot always be retrieved through its copy cue, which occurs in a new context at recognition. The finding of recognition failure means that, for certain items, information in the list cue at recall successfully matches the information in the encoded trace, while the information in the copy cue at recognition does not. Thus, although the TBR item is identical with its nominal copy in the recognition test, the specifically encoded trace of that item may be sufficiently different from the copy cue to preclude recognition.

A number of recent experiments have been directed at determination of the generality of the conditions under which encoding specificity phenomena can be demonstrated. Some of these studies (e.g., Tulving, 1974; Watkins & Tulving 1975; Wiseman & Tulving, 1975) have suggested that the phenomena are rather general, while others (e.g., Postman, 1975; Reder, Anderson, & Bjork, 1974; Salzberg, Note 1) have led to conclusions that encoding specificity is narrowly limited. The main purpose of the present article is to analyze the reasons for such apparently contradictory conclusions and to suggest a resolution of the conflict. This analysis requires that a distinction be drawn between (a) recall superiority over recognition (or simply "recall superiority") and (b) recognition failure of recallable words (or simply "recognition failure"). We are concerned with these two phenomena and with the relation between them in terms of both logic and experiment. The resolution of the conflict takes the form of the argument that, first, the empirical fact indicating encoding specificity is recognition failure, *not* recall superiority, and, second, recognition failure occurs under a much wider set of circumstances than does recall superiority. A second and closely related purpose of the article is to describe four experiments in which we manipulated the nature of the relation between list cues and target words. This was done with a view to examining the effect of this relation on recall superiority and on recognition failure. In all previously published experiments, list cues and target items have been semantically related. Since it is known that the relation between the words of a pair is critical to the degree to which one member of the pair is successful in aiding recall of the second (e.g., Horowitz & Manelis, 1972), it seemed worthwhile to ask how critical this particular feature of the paradigm is for the production of encoding specificity phenomena.

## Relation Between Recall Superiority and Recognition Failure

As we have just mentioned, several writers have concluded that encoding specificity phenomena are limited. Reder et al. (1974) used target words of high and low frequency of occurrence in two experiments. They concluded, among other things, that their results with "low-frequency target words did not support the Encoding Specificity Principle" (Reder et al., 1974, p. 648). In another study, Salzberg (Note 1) described two experiments that were designed to investigate the "question of generality" of the "encoding specificity phenomenon." In these experiments, grammatical class and concreteness of list cues were varied. Salzberg concluded that his experiments had "clearly shown a boundary condition of Tulving and Thomson's (1973) encoding specificity effect" (Salzberg, Note 1, p. 31). More recently, Postman (1975), in an investigation of the generality of the encoding specificity principle, reported eight relevant comparisons of recall and recognition. He concluded that the "encoding specificity effect indexed by the superiority of recall over recognition is not a robust one" (Postman, 1975, p. 671).

In each of these cases, the "encoding specificity effect" to which our colleagues refer, and on which they base their conclusions, is recall superiority, not recognition failure. This is a shift in emphasis from that of the initial demonstrations of Tulving and Thomson (1973). In the Tulving and Thomson experiments, both recall superiority and recognition failure were observed, but the discussion of the theoretical importance of the findings and the application of the encoding specificity principle centered on recognition failure. Because of the difference in emphasis between those arguing for and those arguing against the generality of encoding specificity, it is important to examine the nature of the relation between the two phenomena.

In the paradigm used, each subject was

tested successively on both a recognition and a recall test. Thus, it was possible to analyze the subjects' performance in terms of fourfold contingency tables showing four mutually exclusive retrieval outcomes of target items tested for both recognition and recall: items correctly identified on (a) both the recognition and the recall tests, (b) the recognition test alone, (c) the recall test alone, or (d) neither the recognition nor the recall test. (Other entries of the tables represent the probabilities of overall recognition success and failure, as well as overall recall success and failure.)

Recognition/recall contingency tables allow both a comparison of overall levels of recognition and recall and a calculation of the proportion of recognition failure of recallable words. This latter measure, $P(\overline{\mathrm{Rn}}|\mathrm{Rc})$, is defined as the conditional probability that an item is not recognized given that it is recalled (Watkins & Tulving, 1975). Of necessity, whenever recall exceeds recognition, some recognition failure of recallable words must occur. Thus, $P(\mathrm{Rc}) > P(\mathrm{Rn})$ implies that $P(\overline{\mathrm{Rn}}|\mathrm{Rc}) > 0$: Recall superiority is a *sufficient* condition for recognition failure. It is not, however, a *necessary* condition; recognition failure can also occur when the overall level of recall is *lower* than that of recognition. As Experiments 3 and 4 of the present series will demonstrate, it is quite possible to have a configuration of data in the recognition/recall contingency table showing both superiority of recognition over recall and recognition failure. The nature of the relation between these two phenomena—recall superiority entails recognition failure of recallable items, and recognition superiority does not preclude it—means that it is important to distinguish between recall superiority and recognition failure when assessing their relevance to encoding specificity.

From the point of view of encoding specificity, as well as for the purpose of evaluating classical theories of recognition and recall, the phenomenon of interest is recognition failure. Comparison of overall recognition and recall hit rates are of theoretical interest only if recall is in fact higher than recognition, since this state of affairs entails recognition failure. The contrary observation of recognition exceeding recall can be accounted for by many theories. Moreover, in situations in which recognition is higher than recall, the comparison of only overall levels of recall and recognition provides no information about the extent to which recognition failure may have occurred. Postman's (1975) data, for instance, show both recall inferiority and recognition failure, although Postman pointed out that the magnitude of the latter effect was rather low.

In addition to permitting a calculation of the measure of recognition failure, there is another advantage that derives from administering two successive retention tests to each subject. The fact that an item can be recalled at the second retention test (time T2) indicates that information about that item must have been available at an earlier time, T1. The failure to recognize an item at time T1 in such a situation cannot be attributed to the lack of appropriate stored information, but rather must be a consequence of inadequate retrieval information in the nominal copy of the target item in the recognition test. This assurance about the locus of recognition failure is not directly available in between-subjects comparisons of recognition and recall.

Some researchers have claimed that a potential problem, in the experimental paradigm in which encoding specificity phenomena have been demonstrated, concerns the possible influence of the recognition test on subsequent recall performance (Postman, 1975; Santa & Lamwers, 1974). It is held that recognition testing may enhance recall performance. If it does, then the finding of recall superiority cannot be used to infer that the list cues at recall contained more effective retrieval information than did the copy cues at recognition. Recall superiority might be an artifact of the test sequence.

We agree that comparisons of overall recall and recognition scores are not useful in evaluating ideas of encoding specificity. Indeed, the main point of this article is that such comparisons may be misleading, and that, instead, one should use the measure of recognition failure, $P(\overline{\mathrm{Rn}}|\mathrm{Rc})$. However, since the matter of the effect of one retention test on performance in a subsequent one remains of interest, we considered it in the design of our experiments.

By incorporating the method used in earlier experiments and by replicating the earlier findings, the first experiment serves as a convenient starting point for the series. Subsequent experiments were focused on the effect of the relation between list cues and target items on recognition and recall of target words, and on the dependence of recognition failure on this relation.

## Experiment 1

One feature which characterizes virtually all of the recognition failure experiments is the fact that the recognition test precedes the recall test. It is, thus, possible to argue that the conditions under which recognition and recall are tested are not comparable. Recognition is tested following one exposure to each list item (the presentation of the study list), whereas recall is tested following two exposures to target items (one in the study list and the other in the recognition test). The additional opportunity afforded subjects for study of to-be-remembered items prior to recall might inflate recall scores, thereby producing recall superiority.

### *Method*

Experiment 1 was designed to test the hypothesis that the recognition test influences subsequent recall performance. The experiment was patterned after the Tulving and Thomson (1973) experiments, except that subjects were given an opportunity to generate and subsequently recognize only *half* of the critical list targets. They were then tested for recall of *all* target words. The experiment thus permitted a comparison of recall performance for generated targets, which appeared on the recognition test, with other targets that could not be generated and would thus not be present on the recognition test.

*Materials.* Each study list consisted of 24 to-be-remembered (TBR) words (targets), each presented in the context of a weakly associated input cue. These lists of cue–target pairs were the same as those used by Tulving and Thomson (1973). Corresponding to each target word there was an extra-list cue word that was known to be strongly associated with the target word in semantic memory and that could thus be used to induce the subject to generate the target word in a free-association task. Of the four study lists employed in this experiment, two were chosen as set-establishing lists and the remaining two as critical lists. In the presentation slides, input cues were always typed in lowercase letters directly above their capitalized targets.

*Subjects and procedure.* Twenty-four undergraduate students at the University of Toronto participated as volunteer subjects, in groups of from two to six individuals. Before the presentation of the first list, subjects were instructed that they would be shown slides of words which they were to remember. They were told that each slide would contain two words, one capitalized and the other not, and that their task was to remember the capitalized words.

Following these instructions, the first of two set-establishing lists was presented. Each slide was exposed for 4 sec, with an interstimulus interval of 1 sec. Following the presentation of the list, subjects were given a cued-recall test sheet containing all 24 list cues, haphazardly ordered, and were required to write down as many targets as they remembered, each one next to its appropriate cue word. Subjects were allowed 3 min in which to complete the cued recall test. A second set-establishing list was then presented and tested in the same manner. Subjects were then shown the critical list. One of two different critical lists was presented to half of the subjects; the second was presented to the other half.

Following the presentation of the critical list, subjects were given a free-association task in which they were required to generate 3 free associates to each of 28 stimulus words. The first 2 and the last 2 stimulus words served as buffers, and responses to them were ignored in the scoring and analysis of the data. Twelve of the remaining 24 stimulus words were strong extra-list associates of targets on one of the two critical lists, and the remaining 12 words on the free-association task were strong associates of the targets on the other critical list, which the subject had not seen. Thus, for a given subject, 12 free-association stimuli were expected to elicit copies of target words, while 12 were not.

Subjects were given 10 min in which to complete the free-association (target generation) task. They were then instructed to examine each of the words they had just written down and to circle any words which they recognized as target words from the critical (last seen) list. Subjects were allowed as much time as they felt necessary for this test.

When all subjects had completed the recognition test, they were given a cued recall test of the critical list. The 24 cues were typed on a sheet of paper, with a blank beside each one. Subjects were instructed to write down as many of the capitalized words from the last presented list as they could remember, each next to its appropriate list cue.

### *Results*

*Scoring.* Stringent scoring criteria were used throughout. (The use of more lenient criteria would not materially affect either the theoretical arguments or patterns of experimental results). The total number of critical list targets generated to their strong extra-list stimuli in the free-association task was tabulated, and the fate of these words was recorded in the 2 × 2 recognition/recall contingency tables constructed for each subject. A target was considered correctly recognized only if a copy of that word elicited by the strong extra-list associate was circled. Similarly, a word was considered correct on the cued recall test if it was a target written next to its appropriate list cue.

*Set-establishing lists.* The mean numbers of words recalled from the two set-establishing lists were 14.25 from the first list (59%), with a standard deviation of 4.98, and 15.96 from the second list (67%), with a standard deviation of 4.62.

*Critical list.* The mean number of critical list targets produced in the free-association task, and thus present on the recognition test, was 9.21 (77% of the 12 possible) in the generated list-half, and 1.50 (13%) in the nongenerated list-half. Thus, not all targets from the generated list-half were present on the recognition test, and not all targets from the nongenerated list-half were absent from it. However, the large difference in the numbers of targets from each of the two list-halves did suggest that the experimental manipulation was effective.

The first row of Table 1 presents the recognition and recall data for the critical list targets of Experiment 1. An examination of these data reveals that 67% of those generated targets which were recalled were not recognized. Also, recall superiority over recognition was substantial. Fifty-nine percent of all generated targets were recalled, whereas only 26% of the generated targets were recognized. These data closely replicate the results of experiments demonstrating both recall superiority and recognition failure (e.g., Tulving & Thomson, 1973; Watkins & Tulving, 1975; Wiseman & Tulving, 1975).

The other data of interest in this experiment concern the proportions of targets recalled from the two halves of the critical list: the generated half, of which 77% of the total of 12 targets were present on the recognition test sheet, and the nongenerated half, of which only 13% of the targets were on the recognition test. The proportions of targets recalled from these two list-halves were identical, namely .59.

These data provide no support for the hypothesis that cued recall scores are inflated by the appearance of a target on a prior recognition test. They differ from the data reported by Postman (1975). In his more analytical experiment, he found that target items tested for recognition tended to be more easily recalled than those not so tested, and that the interpolated recognition test reduced subsequent recall performance. The possibility exists, therefore, that in the present experiment these two tendencies balanced, resulting in no net effect. Nevertheless, the obtained recall superiority in this

TABLE 1

PROPORTIONS OF RESPONSE CLASSES IN FREE-CHOICE RECOGNITION AND CUED RECALL TESTS

| Experiment and condition | Recognition hits | False positives | Cued recall | Recognized |  | Not recognized |  | Recognition failure |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  | Recalled | Not recalled | Recalled | Not recalled |  |
| Experiment 1 | .26 | .05 | .59 | .19 | .07 | .40 | .34 | .67 |
| Experiment 2 | .18 | .04 | .34 | .08 | .10 | .26 | .56 | .76 |
| Experiment 3 |  |  |  |  |  |  |  |  |
| RN–RC | .52 | .06 | .38 | .26 | .26 | .12 | .36 | .31 |
| RC–RN (RC) | .57 | .05 | .42 | .37 | .20 | .05 | .38 | .12 |
| (RC) RN–RC | .57 | .05 | .46 | .40 | .17 | .06 | .37 | .12 |
| Experiment 4 |  |  |  |  |  |  |  |  |
| Related | .54 | .03 | .64 | .40 | .14 | .24 | .22 | .37 |
| Unrelated | .51 | .03 | .34 | .21 | .30 | .13 | .36 | .38 |
| RN 1:7 | .53 | .03 | .50 | .31 | .22 | .19 | .28 | .38 |
| RN 1:1 | .52 | .03 | .48 | .30 | .22 | .18 | .30 | .38 |
| RC 12 | .56 | .03 | .53 | .34 | .22 | .19 | .25 | .36 |
| RC 24 | .50 | .03 | .46 | .28 | .22 | .18 | .32 | .39 |

experiment cannot be readily attributed to the overall facilitation of cued recall by an earlier recognition test.

## Experiment 2

One property shared by many experiments in which the phenomenon of recognition failure has been manifested is the composition of the study lists employed. The identical set of word triplets (target, weak list cue, and strong extra-list cue) which had been constructed by Thomson and Tulving (1970) has also constituted the stimulus materials for many subsequent experiments (Tulving & Thomson, 1973; Watkins & Tulving, 1975; Wiseman & Tulving, 1975). The input pairs in these latter experiments as well as those in many other experiments that included at least some of the identical stimulus materials (Postman, 1975; Reder, Anderson, & Bjork, 1974; Tulving, 1974) consisted of target words presented together with *weakly associated* list cues. Horowitz and Manelis (1972) have shown that the relation between members of a word pair is critical to the degree to which one member of the pair (the cue) is successful in aiding recall of the second member (the target). The possibility is thus raised that recognition failure might be critically dependent on the preexperimentally associated cue–target input pairs.

The primary purpose of Experiment 2 was to study the relation between recall and recognition for study lists in which cue-target pairs were not normatively associated. A second purpose of the experiment was to examine again the effect of generation and recognition testing of target words on subsequent list-cued recall.

### *Method*

The design and procedure of Experiment 2 were exactly like those of Experiment 1 with the exception of the nature of the study-list word pairs, which were constructed as follows. The target member of each study pair of one critical list was interchanged with that of a randomly selected pair from the other critical list to yield two new critical lists in which the cue-target relation was now arbitrary. Similarly, the target words of the set-establishing lists were interchanged to yield two new set-establishing lists.

Twenty-four undergraduate students of both sexes from the University of Toronto participated as unpaid volunteers.

### *Results*

The mean numbers and percentages of target words recalled from the first and second set-establishing lists were 8.88 (37%) and 11.71 (49%), with standard deviations of 5.53 and 5.14, respectively. These mean scores are considerably below those in Experiment 1, and appear to reflect the difficulty of encoding a target with respect to a randomly selected cue word.

The data from the critical list of Experiment 2 were scored and analyzed exactly as those of Experiment 1, and are summarized in row 2 of Table 1. Both recognition (hit rate) and recall performance are lower than the corresponding values in Experiment 1. Differences in recognition scores were not quite statistically reliable, $t(46) = 1.99$, but recall scores did differ significantly, $t(46) = 3.82$.[1]

The difference in recall but not in recognition performance between related versus unrelated cue–target word pairs is not entirely surprising, given the fact that the retrieval information in the recall test resides in the list cue, whereas in the recognition test it resides in the copy cue. Hence, a change in the list cue is more likely to affect recall than recognition.

Table 1 shows that neither recognition failure nor recall superiority was eliminated by the use of randomly selected study pairs. Indeed, the level of recognition failure in Experiment 2 (.76) was slightly higher than that of Experiment 1 (.67). It may be that mere random pairing of familiar words does not necessarily render the words in a pair unrelated. We consider this possibility in Experiments 3 and 4.

The design of Experiment 2, like that of Experiment 1, permitted a comparison of recall performance between the generated and nongenerated halves of the critical list. The mean numbers of targets produced in the free-association task from each of these

[1] Throughout this series of experiments, a difference was considered significant if $p < .05$.

halves, respectively, were 8.08 and 1.08, consisting of 67% and 8% of the total number of targets in each of the two halves. The mean number of target words recalled from the generated list-half was 4.08 (34%), whereas the mean number of targets recalled from the nongenerated half was 5.00 (42%). The recall scores for these two list-halves were compared, and the difference was found to be not sufficiently reliable, $t(23) = 1.73$. This result is consistent with that of Experiment 1.

## Experiment 3

In spite of the fact that recall superiority and recognition failure were not eliminated by the use of randomly assigned cue–target pairs, the surprisingly low recognition scores of Experiment 2 justified an attempt to replicate these results. Experiment 3 was designed to further pursue the effects of the relation between study cues and target words. Its primary purpose was to see once more whether recall superiority and recognition failure could be eliminated by employing semantically unrelated cue–target study pairs.

### *Method*

*Materials.* We constructed these new study pairs by initially assigning cues to targets by a random procedure and then, in order to reduce cue-target relatedness, further changing all pairs that appeared to be meaningful. This was done for both set-establishing and critical list pairs.

*Subjects and procedure.* Experiment 3 also incorporated a number of procedural changes from Experiment 2 in order to increase the likelihood of eliminating recall superiority. The most important of these was that subjects were not required to generate target words. The results of other experiments (Watkins & Tulving, 1975, Experiment 4; Wiseman & Tulving, 1975) suggest that such generation lowers recognition performance.

Another change was that recognition and recall tests were given in both orders. One group of subjects (Group RN–RC) took the recognition test before the recall test; the other (Group RC–RN–RC) had one cued recall test before and another after the recognition test. This variation of test order was included to examine the possible effects on recognition of an immediately preceding cued recall test.

A further comparison which Experiment 3 was designed to permit was between recognition test performance on free and forced-choice tests. Moreover, the forced-choice test was constructed so as to allow a comparison of performance when lures either were semantic associates of targets or were semantically unrelated to targets. This latter feature of the forced-choice tests was included in order to test the replicability of the finding (Watkins & Tulving, 1975) that the semantic relation of targets to lures does not influence recognition performance.

TABLE 2
Summary of Procedure for Experiments 3 and 4

| Step | Procedure |
|---|---|
| (a) | Presentation of first set-establishing list |
| (b) | Test of first set-establishing list |
| (c) | Presentation of second set-establishing list |
| (d) | Test of second set-establishing list |
| 1. | Presentation of the critical list |
| 2. | Free-association task |
| | Experiment 3 only |
| 3. | Cued recall test of the critical list (Group RC–RN–RC only) |
| 4. | Free-choice recognition test of the critical list |
| 5. | Forced-choice recognition test of the critical list |
| 6. | Cued recall test of the critical list |
| | Experiment 4 only |
| 3. | Free-choice recognition test (ratio 1:7 or 1:1) of the critical list |
| 4. | Forced-choice recognition test (ratio 1:7 or 1:1) of the critical list |
| 5. | First cued recall test of the critical list (with or without lure cues) |
| 6. | Second cued recall test of the critical list |

The steps in the procedure of Experiment 3 are summarized in Table 2. Preliminarily, the subjects were presented with two set-establishing lists, each of which was followed immediately by a cued recall test. A third and critical list was then presented (Step 1). There were two unique critical lists, each presented to one-half of each group of subjects. Following the presentation of the critical list, subjects were given an interpolated free-association test (Step 2), in order to produce general interference with the retention of words from the critical list. The 28 free-association stimulus words were, unlike those used in Experiments 1 and 2, "neutral" words intended to elicit words other than critical list targets. Subjects were allowed 10 min to generate three free associates to each word. Then, the subjects in Group RC–RN–RC were given a cued recall test for the 24 target words from the critical list (Step 3). These subjects were given a free-choice recognition test (Step 4), a forced-choice recognition test (Step 5), and another cued recall test (Step 6). The subjects in Group RN–RC skipped Step 3, but they were given Steps 4, 5, and 6. Thus, the only difference between the two groups of subjects was whether or not they received an initial cued recall test before the recognition tests.

The free-choice recognition test was constructed as follows. Subjects were given, on a sheet of paper, 24 rows of three words each. Each row contained a target and two distractors. On half of the rows, the distractor items were words which were semantic associates of their targets. For example, the copy of the target word *smoke* was tripled with the distractors *match* and *pipe*. For the other rows, target words were accompanied by distractors which were unrelated to the target, although related to some other target in the test, for instance, *smoke, night, lamb*. The pairing of target words with related and unrelated distractors was counterbalanced across all experimental conditions. Subjects were asked to look at each word in each column and to circle any words that they thought had been capitalized words (targets) in the critical list.

Following the unpaced free-choice recognition test, subjects were instructed to turn to the next page of their response booklets on which was presented a random rearrangement of the rows of the free-choice recognition test, with the constraint that the rows with the related and unrelated distractors were still alternated. They were asked to circle exactly one word on each row, the one that they recognized as being, or thought most likely to be, the target from the critical list.

In the cued recall test that followed the forced-choice recognition test, the 24 list cues were typed on the test sheet in a random order different from that on the initial cued recall test taken by the RC–RN–RC group.

Subjects were 56 undergraduate volunteers of both sexes at the University of Toronto who were randomly divided into two groups of 28 subjects each. They were tested either individually or in small groups of 2 to 6.

*Results and Discussion*

The mean recall scores for the first two set-establishing lists were 7.76 (32%) and 9.71 (41%), with standard deviations of 5.01 and 5.37, respectively. These scores are somewhat lower than the corresponding scores for Experiment 2; the difference may reflect the manipulated reduction in semantic compatibility of cue–target pairs.

The recognition/recall contingencies and other data for Group RN–RC are presented in row 3 of Table 1; the corresponding proportions for Group RC–RN–RC are presented in rows 4 and 5 of the same table. In row 4, Condition RC–RN(RC), are the recognition/recall scores from the initial cued recall test and the recognition test; in row 5, Condition (RC)RN–RC, are scores from the recognition test and the subsequent cued recall test.

Although there were minor differences in recognition and in recall scores between the RN–RC and RC–RN–RC conditions, these differences were not significant. In each condition, recognition exceeded recall, and $P(\overline{\text{Rn}}|\text{Rc}) > 0$, although the magnitude of the recognition failure effect was greater in the RN–RC (.31) than in the RC–RN–RC (.12) condition. We will return to this latter finding shortly.

Comparison of the data of Group RN–RC with those of Experiment 2, in which the test order was the same, shows that (a) recognition in Experiment 3 (.52) was substantially higher than that in Experiment 2 (.18), and (b) there was a large decrease in recognition failure from .76 in Experiment 2 to .31 in Experiment 3. These observations suggest that variations in recognition failure are concomitant with variations in recognition but not recall performance.

The results of the forced-choice tests, presented in rows 1 through 3 of Table 3, are similar in pattern to those from the free-choice tests, although with forced-choice tests, recognition hit rates are higher and recognition failure rates are lower. Comparison of forced-choice recognition accuracy for targets selected from semantically related versus semantically unrelated distractor words confirms earlier findings (Watkins & Tulving, 1975, Experiments 5 and 6) that the semantic relatedness of targets to distractors does not affect recognition performance. The mean proportions of targets recognized in the forced-choice test by Group RN–RC were .69 and .68 for the related and unrelated distractors, respectively. For Group RC–RN–RC, these proportions were .76 and .74, respectively.

Although the reasons for the difference between Experiments 2 and 3 in recognition performance, and thus in recall superiority and degree of recognition failure, are far from clear, some of the procedural differences between the two experiments may be responsible for it. First, the construction of the cue–target pairs differed. Second, the nature of the activity in the retention interval, the free-association task, was different.

TABLE 3

PROPORTIONS OF RESPONSE CLASSES IN FORCED-CHOICE RECOGNITION AND CUED RECALL TESTS

| Experiment and condition | Recognition hits | Cued recall | Recognized | | Not recognized | | Recognition failure |
|---|---|---|---|---|---|---|---|
| | | | Recalled | Not recalled | Recalled | Not recalled | |
| Experiment 3 | | | | | | | |
| RN–RC | .68 | .38 | .32 | .36 | .06 | .26 | .15 |
| RC–RN (RC) | .75 | .42 | .40 | .35 | .02 | .23 | .06 |
| (RC) RN–RC | .75 | .46 | .44 | .31 | .02 | .23 | .05 |
| Experiment 4 | | | | | | | |
| Related | .71 | .64 | .50 | .21 | .14 | .15 | .22 |
| Unrelated | .71 | .34 | .27 | .44 | .07 | .22 | .20 |
| RN 1:7 | .63 | .50 | .36 | .27 | .14 | .23 | .28 |
| RN 1:1 | .76 | .48 | .40 | .36 | .08 | .16 | .16 |
| RC 12 | .74 | .53 | .43 | .31 | .10 | .16 | .19 |
| RC 24 | .68 | .46 | .35 | .33 | .11 | .21 | .23 |

Third, the recognition test was constructed by the experimenter rather than generated by the subject. Data from experiments reported by Watkins and Tulving (1975) and by Wiseman and Tulving (1975) have demonstrated that recognition performance is reliably higher in cases where subjects do not generate targets. Finally, the ratio of targets to lures in the recognition tests differed greatly. In the subject-generated test of Experiment 2, this ratio was on the order of 10:1, since most of the words produced by subjects were not copies of targets. In the experimenter-produced test of Experiment 3, on the other hand, the target-to-distractor ratio was 2:1. The density of targets on the recognition test may have affected the hit rate.

Although Experiment 3 again failed to yield reliable evidence that one retention test can influence performance on a subsequent test, it did demonstrate that test order can affect the magnitude of recognition failure of recallable words. What are the implications of this finding for encoding specificity? The fact that the index of recognition failure of recallable items, $P(\overline{\text{Rn}}|\text{Rc})$ is lower with the RC–RN than with the RN–RC test order most likely reflects differences in what is being retrieved in the recognition test in the two conditions. In the RN–RC condition, the index is based on words presented but *not yet* recalled, whereas in the RC–RN condition, it is based on words presented and *already* recalled. To the extent that the act of recall of an item changes the nature of its trace in memory, the recognition test for recallable words is being applied to two different populations of stored items. For this reason, attenuation of recognition failure in the RC–RN condition has no particular implications for the notion of encoding specificity.

The critical assumption that underlies our suggestion that $P(\overline{\text{Rn}}|\text{Rc})$ be used as the index of encoding specificity in the standard RN–RC paradigm is that the first test, *when unsuccessful,* does not change the information stored about a given target item's occurrence in the original study list. This assumption and its implications have been discussed at some length elsewhere (Tulving & Watkins, 1975). Should this assumption be shown to be untenable, we would have to revise our thinking about the theoretical significance of the finding of recognition failure of recallable words in these and other experiments. But there seems to be nothing in the present data, nor in any other data of which we are aware, that casts doubt on this assumption. One test can facilitate subsequent test performance (Postman, Jenkins, & Postman, 1948; Postman, 1975), but this may be due entirely to the greater probability of retrieval in the second test of those items successfully retrieved in the first test.

## EXPERIMENT 4

Experiment 4 was designed to examine the influence of some of the factors by which Experiments 2 and 3 differed, in order to clarify their effects on recall superiority and recognition failure.

The results of Experiments 1, 2, and 3 together suggest that the semantic relation between cues and targets in the study list can determine levels of retrieval performance, and thus recall superiority and recognition failure. Experiment 4 was designed to examine the relation between recognition and recall for related and unrelated cue–target study pairs within a single experiment with better controls. It also permitted assessment of the importance of the lure to target ratio in the recognition tests. This latter factor has been uncontrolled in previous experiments and has varied widely from subject-generated to experimenter-produced tests. Finally, it was designed to provide information about the ratio of cue words to distractors on the final cued recall test. In the original experiments of Tulving and Thomson (1973) as well as in other comparable experiments (Reder, Anderson, & Bjork, 1974; Tulving, 1974; Watkins & Tulving, 1975; Wiseman & Tulving, 1975; Salzberg, Note 1), subjects were always presented with all of the relevant cues on the cued recall test, and always told that these cues had appeared in the list. Since the subjects' knowledge that each of the cue words presented in the recall test had appeared in the study list might well have influenced levels of cued recall performance, the exploration of this variable seemed desirable.

### *Method*

Experiment 4 incorporated a 2 × 2 × 2 factorial design comprising eight independent groups of 12 subjects each. Study-list word pairs were either related or unrelated, defining two experimental conditions henceforth referred to as the "related" and "unrelated" conditions, respectively. Crossed with this variable were two levels of the ratio of targets to distractors on the recognition test, either 1:1 (24 targets and 24 unrelated distractors) or 1:7 (12 targets and 84 unrelated distractors), referred to as Conditions RN 1:1 and RN 1:7, respectively. The type of cued recall test was crossed with the other two variables. These tests contained either all 24 list cues and no distractors (Condition RC 24) or 12 list cues and 12 distractors (Condition RC 12). Twelve subjects were assigned to each of the eight experimental groups, at random.

*Materials.* Study lists containing related word pairs were the same two set-establishing lists and the same two critical lists as those used in Experiment 1. Two unrelated word lists were constructed as in Experiment 3. There were, thus, two set-establishing lists and two alternate critical lists in each of the related and unrelated conditions.

*Subjects and procedure.* Ninety-six undergraduate students of both sexes at the University of Toronto participated as paid subjects.

The procedure is summarized in Table 2. Prior to the presentation of the set-establishing lists, subjects were instructed in the same manner as subjects in Experiments 1 through 3. Subjects in both the related and unrelated conditions were then given two appropriate set-establishing lists presented in the same way as those in previous experiments, with a list-cued recall test following the presentation of each. After the test of the second such list, a third and critical list was presented (Step 1). All subjects were then given an identical interpolated free-association task (Step 2) in which they were required to produce and write down three free associates to each of 24 neutral stimulus words.

An experimenter-constructed, free-choice recognition test was then administered (Step 3). The recognition test sheet for Condition RN 1:7 was constructed by randomly arranging the 12 targets among the 84 distractors in such a way that each of the four columns contained three targets and 21 distractors. Similarly, the recognition test for Condition RN 1:1 was constructed by assigning six target words and six distractors to each of four columns. Half of the subjects in Condition RN 1:7 were tested with one half of the critical list targets, and the other half of the subjects was tested with the remaining targets.

When all subjects had completed the free-choice recognition test, a forced-choice recognition test was administered (Step 4) with the same layout of target and lure items as on the free-choice test. Subjects were now informed about the exact number of targets on their respective tests (either 12 or 24) and were instructed to circle exactly that number of words on the test sheet.

Finally, all subjects were given two successive list-cued recall tests. For half of the subjects, the first such test (Step 5) consisted of a cued recall test comprising all 24 randomly ordered list cues (Condition RC 24). The other half of the subjects received a first cued recall test with 12 cues and 12 unrelated distractors (Condition RC 12). The list cues in these latter recall tests corresponded to targets that had been tested in the immediately preceding recognition tests. All subjects were informed of the number of list cues on the recall test, and subjects in Condition RC 12 were instructed to first circle those words that they thought were list cues and then to recall the target word that had been paired with each at input.

The second and final cued recall test (Step 6) was, for all subjects, a cued recall test in which all 24 list cues were presented in a different random order from that on the first such test.

*Results and Discussion*

The mean numbers of words recalled from the first and second related set-establishing lists were 14.91 (62%) and 18.67 (78%), with standard deviations of 5.01 and 4.36, respectively. The corresponding means for the unrelated set-establishing lists were 6.85 (29%) and 10.27 (43%), with standard deviations of 4.54 and 4.81.

The recognition/recall contingency data for the critical list are shown on the last six rows of Table 1, classified by the three factors of (a) input relatedness, (b) recognition target to distractor ratio, and (c) recall cue to distractor ratio on the first cued recall test. Rows 6 and 7 show that cue-target relatedness affected neither the overall recognition performance nor recognition failure of recallable words. Subjects presented with related study pairs were able to recall more words than subjects who saw unrelated study pairs. This result parallels those of Experiments 1 and 2. Recall performance was slightly but significantly higher than recognition in the related input condition, $t(47) = 2.50$, and significantly lower than recognition in the unrelated input condition, $t(47) = 3.74$. Recall superiority, thus, was manifested only for *related* study pairs.

The results of Experiment 4 make it clear that recall superiority can be eliminated by the use of unrelated cue–target study list pairs. More important, however, is the finding that both overall recognition and recognition failure of recallable words are unaffected by this variable. These findings suggest, as did the data of Experiments 2 and 3, that recognition failure is related to recognition performance but not to recall.

The data from Groups RN 1:7 and RN 1:1 show that the performance of these two groups was practically indistinguishable. This means that the ratio of targets to distractors affected neither overall hit rate in recognition nor recognition failure of recallable words. Retrieval performance of subjects in the two recall conditions, those with distractors among the list cues (RC 12) and those without (RC 24), was less similar; recognition and recall performance was slightly but not significantly higher in Condition RC 12.

Three analyses of variance were carried out using recognition, recall, and recognition failure proportions as the dependent variables, respectively. The three independent variables, (a) cue–target relatedness, (b) recognition target to distractor ratio, and (c) recall cue to distractor ratio, were the factors in each of the analyses. The results of the analyses indicated that none of the main effects or interactions was significant for either recognition performance or recognition failure. These analyses render meaningful the earlier presentation and discussion of the results of the experiment.

In the evaluation of recall performance, only one effect was significant—input pair relatedness, $F(1, 77) = 32.9$, $MS_e = .003$. Recall performance with targets from related pairs (.63) was significantly greater than that from related pairs (.34). Only one of the interactions (Relatedness × Recall Ratio) approached significance.

As the reader may remember, prior to the first recall test, subjects in Condition RC 12 were required to identify those words on the first recall test sheet that they thought were list cues. Subjects in the related input condition recognized a mean proportion of .85 of the cues with a false alarm rate of .11, whereas subjects in the unrelated input condition recognized a mean proportion of only .73 of the cues with a false alarm rate of .21. The difference in hit rates was found to be statistically reliable, $t(46) = 2.31$.

Experiment 4 also provided another opportunity to evaluate the effect of an item's appearance on the recognition test on subsequent recall. In Experiments 1 and 2, subjects were given strong extra-list associates of only half of the target items on a subject-generated recognition test. Recall was found to be independent of generation and recognition testing of target words. In Experiment 4, the free-association task served merely as a neutral interpolated activity intended to reduce retention performance. For subjects in the RN 1:7 condition, the recognition test included only 12 of the 24 target words, although half of these subjects were tested for recall of all 24 targets. For the related input pairs, the probability of recall of targets that had appeared on the recognition test was .71, whereas the probability of recall of targets not on the recognition test was .60. This difference was not statistically reliable. For the unrelated input pairs, these two probabilities were an identical .35.

The data of Experiment 4 permit an assessment of the reliability of retrieval performance over successive identical tests. If there existed substantial variability in a subject's performance when tested on separate occasions under otherwise identical conditions, then recognition failure might be explained in terms of momentary fluctuations in the accessibility of stored information. Our data do not support such a hypothesis. Subjects in Condition RC 24 received two successive cued recall tests of all target words in the critical list. Analysis shows that only 1% of the target items were recalled on the first but not the second test, and only 2% of the target items were recalled on the second but not the first test.

Retrieval data based on the forced-choice recognition tests are presented in the last six rows of Table 3. Recognition proportions were higher than those of free-choice tests, and recognition failure proportions were lower. Again, however, recognition failure seems to be inversely related to overall recognition performance and independent of recall.

## General Discussion

The results of the series of experiments reported in this article demonstrate that, in the paradigm under study, recall superiority can be eliminated by the use of unrelated cue–target study pairs. Where related pairs were used (Experiments 1 and 4), recall was found to be higher than recognition; the use of unrelated pairs (Experiments 3 and 4) resulted in the finding that recognition exceeded recall. In Experiment 2, where cues and targets were paired randomly, recall exceeded recognition, but it now looks as if such random pairing did not result in truly unrelated study pairs.

The important lesson to be learned from the research presented here, however, is that the reversal of recall superiority did not eliminate recognition failure. Even when the traditional result of the superiority of overall-recognition over overall-recall was obtained (Experiments 3 and 4), considerable proportions of recallable words were not recognized. In Experiment 4, where a direct comparison of the effects of related with unrelated study pairs was made, the degree of recognition failure remained unaffected by the reversal of recall superiority. In the related condition, recognition failure was .36, whereas in the unrelated condition it was .38, even though in the former condition recall exceeded recognition, and in the latter condition recognition exceeded recall. The encoding specificity phenomenon of recognition failure, thus, is not dependent upon recall superiority: It seems to obtain regardless of whether recall exceeds recognition, or vice versa.

As we have already pointed out, the logic of the relation between recognition and recall in the paradigm here is such that recall superiority must necessarily entail recognition failure. But, as our experiments have shown, superiority of recognition over recall does not mean that all recallable words are recognized. Since recognition failure obtains whether or not recall performance exceeds recognition, the conclusion follows that the failure to find recall superiority does not alone demonstrate a limit to encoding specificity. Indeed, other studies similar to those reported here have also demonstrated recognition failure under conditions of recall inferiority (e.g., Postman, 1975; Salzberg. Note 1), although the authors of these studies, by considering the recognition/recall difference, have reached a different conclusion.

Our findings, showing that recognition failure is dependent upon the overall level of recognition but is not dependent on recall superiority, agree with observations we have made in another article (Tulving & Wiseman, 1975). In that article we summarized the results of all the published experiments that have conformed to the paradigm under study and found a highly systematic relation between overall level of recognition and recognition failure of recallable words. The data from the four experiments of the present series fit the function describing that relation very closely. Moreover, with respect to their conformity to the function, the critical data from experimental conditions in which recall exceeded recognition are indistinguishable from those in which recognition exceeded recall. The data from Group RC–RN–RC do not fit the function, probably for reasons we discussed earlier.

The finding that under forced-choice conditions recognition failure is reduced is consistent with the relation expressed by the function: Recognition failure of recallable words decreases when recognition increases. This finding, in turn, is related to the general question of the magnitude of recognition failure necessary to suggest encoding specificity. We have sometimes been asked how large a recognition failure effect must be, in order to be considered theoretically significant. We feel that *any* amount of recognition failure indicates encoding specificity for the simple reason that recognition failure is a continuous function of recognition (Tulving & Wiseman, 1975). The use of an arbitrary criterion of recognition failure, above which it would be due to encoding specificity and below which it would be due to "chance," would be less parsimonious than the continuity position.

To conclude, we have seen that the controversy concerning the generality of encoding specificity phenomena derives from the use of two different indices of subjects' performance in the RN–RC paradigm. One is based on the comparison of overall recognition and recall scores, the other on recognition failure of recallable words. The results of our experiments show that the former index can be readily varied without affecting the conditional probability that recallable words are not recognized. We have suggested that the recognition failure index is more appropriate for studying encoding specificity. It expresses directly the magnitude of the interesting phenomenon of items retrievable through one cue (list cue) but not through another (copy cue), and also avoids certain difficulties inherent in the use of the recall superiority index, such as the problem of the effect of one test on performance on the other.

## REFERENCE NOTE

1. Salzberg, P. M. *On the generality of encoding specificity: Grammatical class and concreteness of cues.* (Program on Cognitive Factors in Human Learning and Memory, Report No. 18.) Boulder: University of Colorado, 1974.

## REFERENCES

Horowitz, L. M., & Manelis, L. Toward a theory of redintegrative memory: Adjective-noun phrases. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 6). New York: Academic Press, 1972.

Postman, L. Tests of the generality of the principle of encoding specificity. *Memory & Cognition,* 1975, *3,* 663–672.

Postman, L., Jenkins, W. O., & Postman, D. L. An experimental comparison of active recall and recognition. *American Journal of Psychology,* 1948, *61,* 511–519.

Reder, L. M., Anderson, J. R., & Bjork, R. A. A semantic interpretation of encoding specificity. *Journal of Experimental Psychology,* 1974, *102,* 648–656.

Santa, J. L., & Lamwers, L. L. Encoding specificity: Fact or artifact? *Journal of Verbal Learning and Verbal Behavior,* 1974, *13,* 412–423.

Thomson, D. M., & Tulving, E. Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology,* 1970, *86,* 255–262.

Tulving, E. Recall and recognition of semantically encoded words. *Journal of Experimental Psychology,* 1974, *102,* 778–787.

Tulving, E., & Thomson, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review,* 1973, *80,* 352–373.

Tulving, E., & Watkins, M. J. Structure of memory traces. *Psychological Review,* 1975, *82,* 261–275.

Tulving, E., & Wiseman, S. Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society,* 1975, *6,* 79–82.

Watkins, M. J., & Tulving, E. Episodic memory: When recognition fails. *Journal of Experimental Psychology: General,* 1975, *104,* 5–29.

Wiseman, S., & Tulving, E. A test of confusion theory of encoding specificity. *Journal of Verbal Learning and Verbal Behavior,* 1975, *14,* 370–381.

(Received August 15, 1975; revision received December 15, 1975)