JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 14, 370–381 (1975)

# A Test of Confusion Theory of Encoding Specificity

SANDOR WISEMAN AND ENDEL TULVING

*University of Toronto*

Subjects studied and were tested for recognition and recall of target words on four successive lists of cue–target word pairs. List-cued recall was higher than recognition of target words in the absence of list cues in all lists, suggesting that recognition failure of recallable words is independent of subjects' familiarity with the task requirements. These results do not support explanations that attribute phenomena of encoding specificity to various sources of confusion in the method and procedure used in previous experiments.

Both common sense and classical theory of human memory hold that a person cannot recall an item of experience that he cannot recognize. This belief, which has for a long time been supported by observations that recognition performance in memory tasks is generally higher than recall performance, has been formalized in a two-stage theory of recall (Anderson & Bower, 1972; Bahrick, 1969, 1970; Kintsch, 1970).

Since the two-stage theory assumes that recall entails both of the two stages while recognition entails only the second, it explicitly denies the possibility that a learned item that could not be recognized could, nevertheless, be recalled. Recent demonstrations of recognition failure of recallable words (Tulving & Thomson, 1973; Watkins & Tulving, 1975) are, therefore, more than experimental curiosities; their existence constitutes a challenge to existing theories. The results of these demonstrations suggest that the encoding of a word-event and the resulting memory trace are rather specific and unique, in the sense that only a relatively restricted set of retrieval cues can provide access to the trace. The cues that on pre-experimental grounds can be thought to be closely related to and associated with the words to be remembered, such as strong semantic associates (Thomson & Tulving, 1970) or nominal copies of to-be-remembered words (Tulving & Thomson, 1973; Tulving, 1974) are not always effective.

The early reactions by other students of human memory to the experimental demonstrations of recognition failure of recallable words and their implications has taken essentially three forms. First, some investigators have accepted the initial findings and have proceeded to delineate their precise nature and the constraints on the conditions under which they can be obtained (Lauer, 1974; Light & Schurr, 1973; Murphy & Wallace, 1974; Olson, 1974; Postman, in press; Reder, Anderson, & Bjork, 1974; Salzberg, 1974; Salzberg & Pellegrino, 1974) and have offered suggestions for the modification of existing theory (Anderson & Bower, 1974; Kintsch, 1974; Reder, Anderson, & Bjork, 1974). The second type of reaction has taken the form of skepticism concerning suggested interpretation of the findings as such. This attitude is exemplified by Martin (1975) who has argued that recognition failure of recallable words is a misnomer inasmuch as the words in the Tulving and Thomson (1973) experiments which the subjects failed to recognize were "not the

This research was supported by the National Research Council of Canada Grant A8632. Requests for reprints should be sent to Endel Tulving at the Department of Psychology, University of Toronto, Toronto M5S 1A1, Ontario, Canada.

same words" as those which the subjects subsequently recalled. This criticism will necessitate an inquiry into the conditions under which two nominally identical words are, and the conditions under which they are not, treated identically by the cognitive system. The third type of reaction reduces itself to attempted dismissal of the broader implications of the findings of recognition failure or recallable words on the grounds either that the results represent an experimental artifact or that they manifest known effects of familiar variables. In either case, the results can be understood without resorting to any new ideas, and hence existing theory need not be modified. This third type of reaction is exemplified in an article by Santa and Lamwers (1974), who described some data and provided several arguments in support of the position that the encoding specificity principle should be discarded.

The present article is addressed to this third type of reaction. We will refer to the position that the phenomena of encoding specificity are artifactual or comprehensible within the confines of old theory as the "confusion theory" of encoding specificity, since many arguments that the proponents of the position have made against the principle of encoding specificity have in common the concept of confusion. Santa and Lamwers (1974), for instance, have attributed certain early findings of Thomson and Tulving (1970) to confused subjects. The experiment that we will describe in this article was designed to evaluate the effects of several sources of confusion pointed out by our critics. However, before describing our experiment, we will present the criticisms made by confusion theory in greater detail.

## Criticisms of Phenomena of Encoding Specificity

Prior to a discussion of the criticisms of encoding specificity, it seems desirable to briefly set forth the encoding specificity phenomena with which the confusion theory is concerned. The encoding specificity principle (Tulving & Thomson, 1973), strictly speaking, is not a theory of memory, and hence its tenability is not open to empirical test. It can be useful, however, for interpreting the outcomes of experiments in which the differential effectiveness of various retrieval cues is examined. It is particularly beneficial in situations in which experimental outcomes are not readily accommodated by existing theory. These outcomes include those involving the relative lack of effectiveness of strong associates of target words as extralist retrieval cues (e.g., Thomson & Tulving, 1970), context effects in recognition memory (Tulving & Thomson, 1971), and recognition failure of recallable words (Tulving & Thomson, 1973; Watkins & Tulving, 1975). These phenomena, whose interpretation can benefit from the application of the encoding specificity principle, can be referred to as encoding specificity phenomena.

Santa and Lamwers (1974) were concerned with two kinds of encoding specificity phenomena. The first was the demonstration of Thomson and Tulving (1970) that strong extralist associates of studied list words (strong cues) were ineffective as retrieval cues under certain conditions where list words were encoded with respect to specific cues in the study list. The second phenomenon discussed by Santa and Lamwers (1974) was that subjects cannot recognize nominal copies of studied list words, even though they can recall these words in the presence of the list cues that had accompanied the target words in the study list (Tulving & Thomson, 1973). Santa and Lamwers did not compare recognition with list-cued recall in their experiments, but they did have several critical comments to make about such comparisons.

Both of the above phenomena can be interpreted in terms of the encoding specificity principle: A retrieval cue is effective if its informational content matches and complements the information contained in the trace of

the to-be-remembered (TBR) event. Thus, it is possible to argue that strong extralist associates of TBR words fail as retrieval cues whenever the specific encoding of the TBR word with respect to its input cue under the experimental conditions (Thomson & Tulving, 1970) produces a trace having relatively little overlap with the encoded version of the extralist cue. The same interpretation can be offered to account for the recognition failure of recallable words (Tulving & Thomson, 1973). Santa and Lamwers criticized this interpretation by questioning the validity of the two types of findings mentioned in the previous paragraph. We will discuss their criticisms of each kind of experimental finding in turn.

The first finding questioned by Santa and Lamwers (1974) was that when target words were encoded with respect to weak list cues, recall of TBR words to strong cues was no higher than free recall. Santa and Lamwers did two experiments in which they showed that subjects who had been given explicit information about the nature of the relationship between the strong extralist cues and the targets did benefit from the presence of such cues. In their Experiment I, extralist cued recall with information (42%) was found to be approximately twice as high as both free recall (23%) and cued recall when no information was given (22%). This result closely resembles the finding reported by Tulving and Thomson (1973, Experiment I) who demonstrated that when subjects were given strong extralist cues but were not instructed in the use of those cues, recall was 15%, but that when subjects were instructed to recall targets by mentally producing strong associates to the same cues, recall was 30% or twice as high as uninstructed extralist cued recall. Santa and Lamwers interpret their data in terms of a generation–recognition model of recall, and suggest that poor extralist cued recall in the Thomson and Tulving (1970) experiment resulted from the use of a strategy which led to the implicit generation of response alternatives which did not include TBR words. According to Santa and Lamwers (1974), subjects used this inappropriate strategy because they were confused.

The second type of finding criticized by Santa and Lamwers (1974) was the finding of recall superiority over recognition. They raised three methodological points in suggesting that this finding was the product of the specific procedure used by Tulving and Thomson (1973).

The first point of their criticisms is addressed to the "poor instructions" (Santa & Lamwers, 1974, p. 418) used in the Tulving and Thomson (1973) experiments. Subjects in those experiments were led to expect that they would be tested for recall of TBR words in the presence of the original list cues. In fact, however, they were given an unexpected recognition test. One could argue that such a deviation from the established experimental procedure might have caused the subjects confusion and uncertainty about the experimental task. This, in turn, would presumably lead to poorer recognition performance than would have been obtained had subjects been aware of the experimental procedure. In particular, subjects might not have adopted a strategy which would have enabled them to perform well on the recognition test. This implies that recall superiority, and the recognition failure of recallable words it entails, can be reduced or eliminated if subjects are made aware of the actual experimental procedure and the impending recognition test.

The second point raised by Santa and Lamwers is directed specifically at the finding of recognition failure of recallable words. In the Tulving and Thomson (1973) procedure, subjects generated their own recognition tests by producing several free associates to a number of stimulus words selected such that they would be likely to elicit copies of target words. Santa and Lamwers argued that when subjects are required to recognize copies of target words from among their own generated free-association responses, they are "presented

with a difficult list discrimination problem" (Santa & Lamwers, 1974, p. 419). A subject faced with a copy of a target word which he has produced in the free association task would have a problem in deciding whether that item looks familiar because he has seen it in the study list or merely because he has just produced it himself. Thus, the subjects might be unable to perform as well with a self-made test as they might with a more traditional experimenter-made recognition test.

The third methodological difficulty resulting from the Tulving and Thomson (1973) procedure lies in the "high similarity choice sets for recognition" (Santa & Lamwers, 1974, p. 421). In the free association task, subjects generated distractors as well as targets for the ensuing recognition test. Since both distractors and targets were high semantic associates of the stimuli in the free association task, the semantic relationship between the distractors and targets was necessarily quite high. A consequent difficulty of discrimination may have been experienced by subjects. Since such difficulty was absent in the recall test with list cues, recall performance may indeed have been found to be higher than recognition. Again, the contrived nature of the experimental procedure leads to memory performance which is inconsistent with both common sense and classical theory.

The experiment reported in this article was designed to evaluate the methodological criticisms just described. We will argue that confusion theories of encoding specificity in general, and the position of Santa and Lamwers (1974) in particular, are untenable. We will show that, even though it may be true that strong extralist cues can be effective for retrieval under certain circumstances, the "encoding specificity phenomena" cannot be accounted for by the generation–recognition model defended by Santa and Lamwers (1974). The experiment we will describe did not employ strong associates of target words as retrieval cues; rather, it examined retrieval of target words in the presence of copy cues (nominal copies of target items) in order to provide a stronger test of confusion theory than is possible with strong extralist cues. The reason for this is evident in a consideration of the role of retrieval cues in the two-stage, generation–recognition theory of recall.

## Retrieval Cues and the Two-Stage Theory of Recall

According to a simple generation–recognition theory of recall (Anderson & Bower, 1972; Bahrick, 1969, 1970; Kintsch, 1970), which is schematically diagrammed in Figure 1, the role of any retrieval cue is limited to the

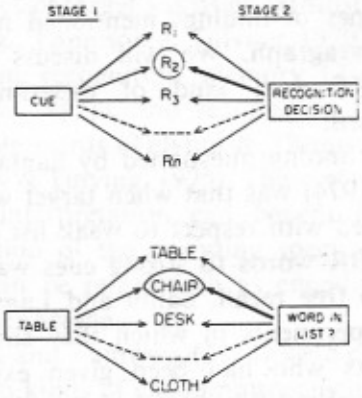


FIG. 1. A schematic diagram of the generation–recognition theory of recall.

first, or generation, stage of the two-stage process of recall. Within this stage, retrieval cues affect only the probability with which a TBR item will be generated and thereby included in the set of words upon which the second stage, or decision process, operates. In particular, such a view holds that strong extralist associates of TBR words are effective as retrieval cues only to the degree that the subject is successful in using them to generate TBR words. The important implication which follows from this view, therefore, is that by increasing the probability of target generation, one can increase the probability of successful recall.

Within the above framework, the finding of Thomson and Tulving (1970) that strong extralist cues did not lead to higher recall than

free recall can only be interpreted by making the assumption that the probability of target generation to strong extralist cues was less than the probability of target generation in free recall. The two-stage theory would also predict that if the probability of target generation to strong cues could be raised, for example by inducing subjects to generate more target words to these cues, then higher recall performance would result. Evidence consistent with this prediction is found in experiments described above (Tulving & Thomson, 1973; Santa & Lamwers, 1974) which demonstrated an improvement in recall performance when subjects were instructed in the use of strong cues. However, Tulving and Thomson went beyond that demonstration to show that recall to strong cues with instructions was still lower than recall to (weak) list cues. This comparison was not made by Santa and Lamwers (1974). The Tulving and Thomson result can still be accounted for by generation–recognition theory by assuming that target generation to weak cues was more likely than that to strong cues with instructions. In order to test this possibility, Tulving and Thomson (1973) made the target generation process explicit by asking subjects to write down the words that they produced. Moreover, they considered in their analyses only those target words which *had been generated* to strong cues. Since target generation to these strong cues was effectively 100% (by restricting analyses to generated targets), it became impossible to attribute the greater effectiveness of list cues for retrieval to greater likelihood of target generation. The differential effectiveness of (strong) extralist cues and (weak) list cues in Experiments II and III of Tulving and Thomson (1973) could not be explained in terms of differential generation of targets.

The point that we have been making is simply this: The simple generation–recognition theory clearly assumes that copies of target words should be at least as effective for retrieval as any extralist associates (e.g., strong cues). Therefore, the experiment which we report in this article compared retrieval of targets to list cues with retrieval to copy cues, rather than to strong extralist associates.

## A Test of Confusion Theory

The basic idea behind the experiment to be reported is quite simple: It will demonstrate that target retrieval to copy cues can be exceeded under conditions in which all of the methodological sources of difficulty and confusion that we have briefly reviewed earlier in this article are either minimized or precluded. If the phenomenon of recall superiority (retrieval with list cues) over recognition (retrieval with copy cues) occurs under conditions where subjects cannot or are very unlikely to be confused, then methodological idiosyncrasies or confusion can be eliminated as explanations of earlier results.

In the experiment, subjects successively (*a*) studied a list of cue-target pairs, as in previous experiments (e.g., Tulving & Thomson, 1973), (*b*) generated free association responses to strong extralist associates of target items, (*c*) took an experimenter-made recognition test, and (*d*) recalled target items in the presence of list cues.

In this experiment, we eliminated the possibility that target words would not be available to the subject in the recognition test. The reader may recall that the possibility was raised by Santa and Lamwers that subjects would generate only weak associates to strong extralist cues in the Thomson and Tulving (1970) experiments. In the present study, *all* target words were made explicitly available to the subject in the recognition test, obviating the need for their implicit generation. In this way, possible difficulties due to subjects' lack of information about the relation between extralist cues and targets was circumvented.

The problem of confusion created by subjects' ignorance of the experimental procedure was eliminated by giving subjects four successive critical lists, each followed by a

recognition test and a recall test as just described. Thus, even though initially subjects might have been uncertain as to the experimental method, they would become familiar with the task requirements in the course of being tested on successive critical lists and thus gain greater command of their "true" memory capabilities, unaffected by unexpected twists in the procedure. The use of four successive lists that would help the initially naive subjects to become sophisticated also minimized any possible confusion resulting from poor instructions.

The possible list discrimination problem created by subjects who generate copies of target items in a free association test was evaluated and circumvented by having each subject generate copies of only *half* of the critical list targets and by testing recognition with an *experimenter-made test* for *all* targets. This feature of the design allows the assessment and comparison of recognition failure of recallable words in situations where subjects do generate and in which they do not generate copies of TBR words.

Finally, the possibility that the recognition performance in the experiment of the type under discussion suffers because of semantic similarity of distractors to target words on the test was eliminated by using distractor items semantically unrelated to targets. There already exists a considerable body of evidence (Watkins & Tulving, 1975, Experiments V and VI) that the relatedness of targets to lures is not an important determinant of the phenomenon of recognition failure of recallable words. The results of the present experiment, therefore, were expected to replicate earlier findings with unrelated distractors.

## Method

### *Design and Materials*

Each subject was tested with five lists in succession. The first list was the same for all subjects and served as a set-establishing list. The other four were critical lists, designated Lists A, B, C, and D. The four lists are shown in Table 1. These four lists were assigned to the four critical list positions according to a Latin-square design. Thus, each of the four lists served equally frequently in each of the four test positions.

The word pairs in the critical lists were the same as those used by Tulving and Thomson (1973). In each pair, the to-be-remembered word was a weak (1% in free-association norms) associate of its accompanying list cue. Corresponding to one half of the target words in each critical list were strong (mean of 52% in free-association norms) associates of the target words that served as stimulus items in the free association task. These words are also shown in Table 1.

The observation of primary interest in the experiment concerns recognition and cued recall of target words, and their relation, across the four successive critical lists.

### *Subjects and Procedure*

A total of 24 summer session students at the University of Toronto participated as paid volunteers in this experiment. They were tested either individually or in groups of from two to six individuals.

Upon entering the experimental room, subjects were instructed that they would be shown a number of slides of words and that their task was to remember the words. In particular, they were told that each slide would contain two words, one typed in capital letters and the other typed in lowercase letters, and that they were to remember the capitalized words. Subjects were informed that each lowercase word appearing above the capitalized word was a cue word and that these cue words might help them to remember the capitalized target words.

Following these instructions, the set-establishing list was presented. The purpose of this list was to induce subjects to encode each target word with respect to its input cue. Each slide was exposed for 4 secs with an interstimulus interval of 1 sec—the time

TABLE 1

CRITICAL LISTS AND FREE ASSOCIATION TASK STIMULUS WORDS

| List A | | | List B | | |
|---|---|---|---|---|---|
| Input cue | Target word | Free association | Input cue | Target word | Free association |
| preach | RANT | | badge | BUTTON | |
| sword | DANCE | | base | HIT | |
| dim | HAZY | | book | TITLE | |
| sew | THIMBLE | | coarse | FLUFFY | |
| afraid | GUILT | | cushion | SHOCK | |
| army | TOUGH | | doodle | POODLE | |
| buy | SPEND | | family | SUCCESS | |
| dock | WHARF | | give | UP | |
| frigid | WEATHER | | lion | PAW | |
| pony | SOLDIER | | oil | FILM | |
| sea | WALL | | piano | LESSON | |
| tunnel | BURROW | | right | MIGHT | |
| grasp | BABY | infant | exist | BEING | human |
| pretty | BLUE | sky | train | BLACK | white |
| glue | CHAIR | table | covering | COAT | lining |
| blade | CUT | scissors | moth | FOOD | eat |
| fruit | FLOWER | blossom | tool | HAND | finger |
| cheese | GREEN | grass | hope | HIGH | low |
| command | MAN | woman | cottage | LOVE | hate |
| beat | PAIN | hurt | door | RED | color |
| cloth | SHEEP | lamb | roll | RUG | carpet |
| drink | SMOKE | tobacco | memory | SLOW | fast |
| home | SWEET | bitter | mountain | TREE | leaf |
| cave | WET | dry | brave | WEAK | strong |
| List C | | | List D | | |
| Input cue | Target word | Free association | Input cue | Target word | Free association |
| fraud | VICTIM | | balloon | STRING | |
| minor | MUSIC | | block | TACKLE | |
| wing | BROKEN | | city | VILLAGE | |
| flat | EVEN | | correct | COMPLETE | |
| ant | PICNIC | | apple | CIDER | |
| bare | ABSENT | | far | DISTANT | |
| collar | THROAT | | freeze | ROAST | |
| fancy | SIMPLE | | glow | SMILE | |
| made | BY | | monk | OLD | |
| rat | TRAP | | peak | TIP | |
| slime | DISGUST | | quota | RATIO | |
| salty | SOUP | | whole | TOTAL | |
| whistle | BALL | tennis | spider | BIRD | eagle |
| plant | BUG | insect | crust | CAKE | bake |
| ground | COLD | hot | barn | DIRTY | clean |
| sun | DAY | night | art | GIRL | boy |
| swift | GO | stop | glass | HARD | soft |
| stomach | LARGE | small | head | LIGHT | dark |
| bath | NEED | want | country | OPEN | closed |
| lady | QUEEN | king | cabbage | ROUND | square |
| deep | SLEEP | dream | stem | SHORT | long |
| butter | SMOOTH | rough | think | STUPID | smart |
| wish | WASH | soap | whiskey | WATER | lake |
| noise | WIND | blow | adult | WORK | labor |

necessary for the projector to recycle. Following the presentation of this list, subjects were given a sheet of paper on which was a list of the 24 input cues in a random order different from that of the input sequence. Subjects were instructed to write down as many of the target words as they remembered, each one next to its input cue. They were allowed 3 mins in which to complete the cued recall task.

When subjects had completed the cued recall test of the set-establishing list, they were instructed that they would be shown another list of words and that again their task was to remember the capitalized words and that the cue words might help them to do so. The first of four critical lists was then presented in exactly the same manner as the set-establishing list. Following the presentation of this list, subjects were given an unexpected free association task in which they were required to generate four free associates to each of 12 stimulus words. These stimulus words were the strong associates corresponding to 12 of the target words and were intended to elicit those 12 targets as free association responses.

Following the completion of the generation task, subjects were instructed to turn to the next page of their response booklets. On this page was an experimenter-prepared recognition test comprising 96 items: all 24 target words from the critical list and 72 unrelated distractor items. The recognition test sheet was laid out in 24 rows of four items each. Twelve of these rows each contained one target, six rows each contained two targets, and the remaining six rows contained no targets. The distractor items were semantically unrelated to the target words and differed for the recognition test of each of the four critical lists. Subjects were instructed to examine all recognition-test items and to circle those that were the same as target items in the list they had just studied. The recognition test was of the free-choice yes-no variety. Subjects were given as much time as they needed to complete the recognition test. They were then given a cued recall test with list cues, similar to that which followed the set-establishing list.

Following this sequence of steps with the first critical list, a second critical list was presented and tested exactly as the first one had been. The subjects were told that they would see and study a new list and that, again, their task was to remember the capitalized words and that the lowercase ones might help them to do so.

The third and fourth critical lists were then presented and tested exactly like the second one.

## Results

The mean number of words recalled from the set-establishing list was 12.88 (54%) with a standard deviation of 5.56.

The data relevant to recognition and recall of targets from the generated and nongenerated halves of the four successive critical lists are presented in Table 2 (Rows 1 through 8). Column 1 in Table 2 lists the experimental conditions. Columns 2 and 3 present the proportions of target items correctly identified in the recognition test (the hit rate) and the proportion of nontarget items incorrectly identified (the false positive rate). Column 4 presents the proportion of items correctly identified on the cued recall test. Columns 5 through 8 present the results of an analysis of the relation between recognition and cued recall with respect to individual subject-items. Since in this experiment each target item was tested in both a recognition and a recall test, it is possible to tabulate the data in terms of a fourfold contingency table which summarizes the proportion of items in each of four mutually exclusive combinations of test outcomes: (*a*) target items both recognized and recalled, (*b*) items recognized but not recalled, (*c*) items recalled but not recognized, and (*d*) items neither recognized nor recalled. The advantages of this mode of data presentation have been detailed by Watkins and Tulving (1975). The fourfold contingency data permit the

TABLE 2

PROPORTIONS OF RESPONSE CLASSES IN RECOGNITION AND CUED RECALL TESTS

| Experimental condition | Recognition | | Cued recall | Recognized | | Not recognized | | Recognition failure |
|---|---|---|---|---|---|---|---|---|
| | Hits | False positives | | Recalled | Not Recalled | Recalled | Not recalled | |
| List 1 | | | | | | | | |
| Generated | 0.46 | 0.01 | 0.61 | 0.34 | 0.13 | 0.28 | 0.26 | 0.45 |
| Nongenerated | 0.56 | | 0.64 | 0.47 | 0.09 | 0.17 | 0.27 | 0.26 |
| List 2 | | | | | | | | |
| Generated | 0.41 | 0.01 | 0.62 | 0.31 | 0.10 | 0.31 | 0.28 | 0.49 |
| Nongenerated | 0.49 | | 0.62 | 0.43 | 0.06 | 0.21 | 0.30 | 0.33 |
| List 3 | | | | | | | | |
| Generated | 0.34 | 0.01 | 0.59 | 0.25 | 0.09 | 0.34 | 0.32 | 0.57 |
| Nongenerated | 0.47 | | 0.64 | 0.39 | 0.07 | 0.25 | 0.28 | 0.39 |
| List 4 | | | | | | | | |
| Generated | 0.37 | 0.01 | 0.57 | 0.29 | 0.08 | 0.28 | 0.35 | 0.49 |
| Nongenerated | 0.47 | | 0.57 | 0.38 | 0.09 | 0.18 | 0.35 | 0.33 |
| Lists 1–4: Totals | | | | | | | | |
| Generated | 0.40 | | 0.60 | 0.30 | 0.10 | 0.30 | 0.30 | 0.50 |
| Nongenerated | 0.50 | | 0.62 | 0.42 | 0.08 | 0.20 | 0.30 | 0.33 |

calculation of a measure of recognition failure of recallable words. This measure is defined as the conditional probability that an item is not recognized given that it is recalled. Thus, for instance, in the first critical list 22% of *all* the items were recalled but not recognized, and 62% of *all* the items were recalled. The conditional probability that a recalled item was not recognized, therefore, is .22/.62, or .36. Column 9 of Table 2 presents these conditional probabilities for the four successive lists. The data in Table 2 thus permit both a direct comparison of overall levels of recall with those of recognition and also a more detailed analysis of the degree to which subjects cannot recognize words that they can recall.

Data in Table 2 show that (*a*) recognition was lower for generated list halves than for nongenerated halves, (*b*) recall in generated and nongenerated list halves did not differ greatly, (*c*) overall level of recognition, measured in terms of hit rate, decreased systematically over the four successive list positions, from .46 to .37 for generated halves and from .56 to .47 for nongenerated halves (the false positive rate for each of the four list positions was .01), (*d*) the cued recall hit rate remained constant over the first three lists and then declined slightly in the fourth, and (*e*) recall was higher than recognition in all four list positions, for both generated and nongenerated list halves. We will now consider these findings in greater detail.

One analysis of interest concerns the effect of generation of copies of target words in the free association task on the subsequent recognition and list-cued recall of the target words. The reader may remember that in the free association task, each subject had the opportunity to generate copies of only 12

target words. The mean numbers of target copies that were in fact generated were 9.04, 9.20, 9.45, and 8.66 for the four critical list positions, respectively, or 75, 76, 78, and 72% of the 12 possible.[1] The data relevant to the effect of target generation on recognition and recall accuracy are presented in Rows 9 and 10 of Table 2. These rows summarize the data from the generated and nongenerated halves of the critical lists, summed over all four lists. These data show that recall of the target words in the two conditions did not differ: 60% of the words in the generated half and 62% of the words in the nongenerated half were recalled. But, the data also show that the generation of targets did reduce recognition performance: hit rates were 40% for targets in the generated half and 50% for targets in the nongenerated half of the critical list. This difference was compared by means of a sign test and was found to be statistically reliable ($p < .001$). Recognition failure of recallable words was also higher for targets in the generated list halves (50%) than for those in the nongenerated halves (33%). The difference in recognition failure between the generated and nongenerated list halves was compared by a sign test, and was found to be statistically significant ($p < .001$). These analyses confirm earlier findings by Watkins and Tulving (1975).

The finding that recall was higher than recognition for both generated and nongenerated list halves of all four list positions is, of course, of primary interest. The difference in the proportion between recall and recognition for the generated halves was .15 in the first list, and .21, .25, and .20 in the second, third, and fourth lists, respectively. For the nongenerated list halves, these proportions were .08, .15, .17, and .09, respectively. An examination of the data in Table 2 makes it clear that these differences do not interact across list position with generation; the retrieval pattern for targets in generated list halves is closely paralleled by that for targets in nongenerated halves. Thus, the data from the generated and nongenerated halves of lists in each serial position were combined and the recognition and recall scores for each subject were compared for each of the four successive lists by means of *t*-tests. For each list position, recall was significantly higher than recognition ($p < .005$ for the first list; $p < .001$ for the second, third, and fourth lists). Thus, there is no evidence that the superiority of recall over recognition decreases as subjects gain familiarity with the nature of the experimental procedure. If anything, the data indicate that practised subjects show a larger effect of recall superiority than unpractised subjects, although the degree of practice is here confounded with the amount of prior learning material. At any rate, there is no support for the confusion theory in these data.

The proportions of unrecognized but recalled words show a rather similar picture. The proportion of recognition failure increased through the first three lists and then declined in the fourth. These proportions are, respectively, .45, .49, .57, and .47 for the generated halves, and .26, .33, .39, and .33 for the nongenerated havles. Again, there is no support for confusion theory in these data.

These analyses suggest two conclusions with respect to confusion theory. One, generation of copies of targets indeed appears to be a source of confusion for the subjects and it does reduce subjects' performance on the recognition test. Two, recall can be higher than recognition, and recognition failure of recallable words can occur, even when the source of confusion attributable to target generation is eliminated.

## Discussion

The purpose of this experiment was to evaluate certain methodological criticisms relating to experiments interpreted in terms of the encoding specificity principle. These criticisms had been proposed by Santa and

[1] There were a total of 25 targets from the nongenerated halves of the critical list produced in the experiment, or a mean of .26 targets per subject per critical list.

Lamwers (1974) to the effect that the results of earlier experiments showing lack of effect of extralist cues and recognition failure of recallable words (e.g., Thomson & Tulving, 1970; Tulving & Thomson, 1973) are attributable to the procedural idiosyncrasies in these experiments. In the experiment reported here, subjects' memory for target words in a list was tested first in a recognition test and then in a cued recall test using list cues. Four such lists were given to each subject successively. It was assumed that subjects were experimentally naive in the first test, and thus may have been confused as to the exact requirement of the task, and that by the fourth test they would have been less confused or not confused at all.

While there were some changes in recall and recognition performance across the four successive lists, there was no evidence that the critical finding of recognition failure of recallable words was attenuated as subjects became more familiar with the task requirements. Recall was higher than recognition and sizable proportions of recallable words were not recognized in each of the four list positions. The conclusion follows, therefore, that whatever findings were obtained here, and have been obtained in other experiments, with the first experimental list, cannot be readily attributed to naïveté of subjects.

A secondary finding of some importance was that both overall recognition and the proportion of unrecognized recallable words were lower for target words which the subject had generated in the free association task, interpolated between presentation of the study list and the recognition test, than for words not thus generated. This means that the generation procedure in earlier experiments may have been partly responsible for the obtained recall superiority (e.g., Tulving & Thomson, 1973). Nevertheless, it could not have been completely responsible, since even with nongenerated targets, recall was superior to recognition and recognition failure of recallable words was greater than zero.

It may be useful to remind the reader that in the experiment described here, recognition was tested with a typical experimenter-prepared recognition test, and that the distractor items were selected randomly from a larger pool of words. The fact that recall was superior to recognition and that recognition failure of recallable words was greater than zero in this experiment suggest that the same patterns of results of the original experiments (Tulving & Thomson, 1973) are not attributable to subject-generated recogntion tests and related distractors in those experiments.

These results thus provide little support for any theory that attempts to account for the kinds of data reported by Thomson and Tulving (1970) and by Tulving and Thomson (1973) in terms of confusion, either confusion of the subjects as to the nature of their task or the meaning of cues (Santa & Lamwers, 1974), or confusion in the dictionary sense of failure to distinguish between things. The explanation of various encoding specificity phenomena, such as context effects in recognition memory, ineffectiveness of extralist associates as retrieval cues, and recognition failure of recallable words, is still not in hand, and the search for an explanation constitutes an important research problem. The encoding specificity principle (Tulving & Thomson, 1973) provides a general conceptual framework in which solutions to problems posed by these phenomena can be sought. Attempts to ascribe phenomena to particular features of original experiments or to various sorts of confusion or artifact created in these experiments seem to constitute an unproductive reaction to the concept of encoding specificity.

## References

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

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

BAHRICK, H. P. Measurement of memory by prompted

recall. *Journal of Experimental Psychology*, 1969, **79**, 213–219.

BAHRICK, H. P. A two-phase model for prompted recall. *Psychological Review*, 1970, **77**, 215–222.

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

KINTSCH, W. *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1974.

LAUER, P. A. Encoding specificity in the cued and free recall of categorically and alphabetically organized words. *Bulletin of the Psychonomic Society*, 1974, **4**, 496–498.

LIGHT, L. L., & SCHURR, S. C. Context effects in recognition memory: Item order and unitization. *Journal of Experimental Psychology*, 1973, **100**, 135–140.

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

MURPHY, M. D., & WALLACE, W. P. Encoding specificity: Semantic change between storage and retrieval cues. *Journal of Experimental Psychology*, 1974, **103**, 768–774.

OLSON, A. M. The differential effect of syntactical pairings on cued recall and recognition. *Bulletin of the Psychonomic Society*, 1974, **3**, 232–233.

POSTMAN, L. Tests of the generality of the principle of encoding specificity. *Memory & Cognition*, in press.

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

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.

SALZBERG, P. M., & PELLEGRINO, J. W. The generation and recognition components of encoding specificity. *Bulletin of the Psychonomic Society*, 1974, **4**, 9–11.

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. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology*, 1971, **87**, 116–124.

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

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

(Received January 30, 1975)