Recognition Failure: Another Case of Retrieval Failure

JAN C. RABINOWITZ, GEORGE MANDLER, AND LAWRENCE W. BARSALOU

University of California, San Diego

A theoretical explanation of the phenomenon of recognition failure (failure to recognize items that can be recalled) is presented in terms of the dual access theory of recognition, which distinguishes between presentation codes and conceptual (retrieval) codes. In the paired-associate paradigm, retrieval involves access to the pair as a unit, which is holistically coded. As predicted by the theory, recognition failure is reduced when a more stringent recognition criterion is used, specific item pairs are more subject to recognition failure than others, and recognition failure is essentially eliminated when the proper access test (backward retrieval) is used and significantly reduced when variability in recognition performance is taken into account. These results were extended to an experimental paradigm that investigated performance at steady state.

In a series of experiments, Tulving and his associates (Tulving, 1968; Tulving & Thomson, 1973; Watkins & Tulving, 1975; Wiseman & Tulving, 1975; Wiseman & Tulving, 1976) have shown that the phenomenon of recognition failure, which is both counterintuitive and countertheoretical, can easily be demonstrated in the paired-associate paradigm. The phenomenon itself is defined as the failure to recognize certain items which subjects can, however, recall. The findings are counterintuitive at first glance because it appears that if one can recall something, one certainly ought to be able to recognize it. They are countertheoretical because they speak directly against two classes of theories; strength theories which directly predict that any item that can be recalled must have enough “strength” to be recognized, and generation/recognition theories which assume that recognition is a subprocess in the recall task. In the experiments to be presented here, we address neither of these theoretical questions, but rather accept the empirical demonstration. In particular, we will not enter into Tulving’s critique of generation/recognition theories, primarily because we believe, and have discussed previously, that this class of theories is not relevant to the paired-associate paradigm in which Tulving has demonstrated recognition failure (Mandler, 1976; Rabinowitz, Mandler & Patterson, 1977). Nor are we concerned with the methodological criticisms that have been directed against some of these findings (e.g., Santa & Lamwes, 1974, 1976).

The major purpose of the experiments to follow was to explore the mechanisms that are responsible for differential recognition and recall performance and the phenomenon of recognition failure. The origin of these investigations is a theoretical position that has not been addressed by Tulving and his associates in their critique of memorial theories. The position was originally presented by Mandler, Pearlstone, and Koopmans (1969); it was presented more elegantly by Atkinson and Juola (1974) and has been further explicated for free recall and for a variety of different memory paradigms (Mandler, 1972, 1976, 1977). Briefly stated, the position presents a dual code theory of
recognition. There are two major encodings of items at the time of presentation. The first class we call the presentation code. A variety of different authors have used such terms as occurrence information, familiarity, and perceptual and phonological cues for this class of encodings. We assume that this type of information is directly and automatically accessed upon presentation of a physical copy of the to-be-remembered item. At the time of testing, these copy cues generate automatic judgment of occurrence or familiarity that are presumably based on perceptual rather than conceptual information (e.g., Posner & Snyder, 1975). The second class of codes are conceptual codes that include what is usually considered under the rubric of retrieval information, including categorization, contextual relations, and a variety of class and category judgments about the item.

In a recall test, when no copy cue is present, only information relevant to the second, conceptual class of codes is available, and retrieval is determined by the type of encoding during presentation and by the cues available for retrieval. We should emphasize here that these conceptual codes will differ from situation to situation and possibly even from subject to subject, but that some classes of such codes can be identified. For example, during free recall, these are probably semantic and categorical, during paired-associate learning they are holistic and relational, and in serial learning they represent the serial structure of the presented event (cf. Mandler, 1970).

In the recognition test, the first step is automatic accessing of the presentation code. Instructions and situations presumably set some criterion for this presentation information and any item whose presentation code exceeds that criterion will be called old. In turn, if subjects are to make both “old” and “new” judgments, a second criterion at the lower end of the presentation code distribution will be invoked, and if the presentation code value falls below that criterion, that item will be called new. After the automatic accessing of the presentation code, there will be a large number of items which fulfill neither of the two criteria. It is at this point that we have invoked a retrieval phase during the recognition test. Some subset of items that fail the original presentation test will be subjected to a test of accessibility or retrievability. Again, we note that for different kinds of encodings these accessibility tests will be different. We have shown previously that a retrieval check which inquires whether a particular item is retrievable according to the encodings at the conceptual level is a satisfactory mechanism for explaining some of the variance of recognition data in the free recall paradigm (e.g., Mandler & Boeck, 1974; Rabinowitz et al., 1977). One other important aspect of this position must be stressed here and we will return to it again in connection with Tulving’s paired-associate paradigm. The retrieval mechanism which is used in order to check the accessibility of an item in the second stage of the recognition test may or may not be identical with that used as a retrieval mechanism during a recall test (cf. Mandler, 1976).

Given this theoretical context in which we wish to investigate the occurrence of recognition failure in the paired-associate paradigm, it is useful to review briefly Tulving’s explanation of the phenomenon and similarities and differences between that position and the one to be explored here. We start off with essential agreement when Tulving (1976) says that “recognition and recall differ only with respect to the exact nature of the retrieval information available to the rememberer” (p. 37). However, the next sentence in this passage discloses our disagreement when Tulving says that “in recognition, retrieval information is carried by a literal copy of the event or item to be remembered; in recall, the retrieval information is contained in cues other than copy cues.” First of all, we do not believe that any retrieval or other information is “contained” in a cue. More specifically, we part company on such statements as that “the retrieval information in the recall tests resides in the list cue” (Wiseman &
Tulving, 1976). All information, whether about presentation or a conceptual membership, is constructed at the time of presentation and retrieved at the time of recall or recognition. More important, however, is our position that a literal copy of the event provides one kind of retrieval information which is not the only information to be used during a recognition judgment. In fact, the presentation information provided at the time of the encounter with the copy cue (for example, after a long period of time) may be quite inadequate to make any nonrandom judgment about the previous occurrence of the item. On the other hand, the conceptual information that is available may be quite adequate to provide a limited and partially correct recognition judgment. For example, Rabinowitz et al. (1977) have shown that the false alarm rate to conceptually related distractors in the free recall situation may be as much as three to four times higher than that to conceptually unrelated items. In that kind of a test, the subject is not very efficient in terms of recognition (as, for example, evaluated by d'), but performance is very good if one is willing to score recognition as being able to tell what class of items was presented or not at the time of original acquisition. Thus, Wiseman and Tulving (1975) are only partially correct when they note that “the retrieval information contained in the copy cue is to a considerable extent independent of the retrieval information contained in the list cue.” Under specified conditions, the retrieval information available through access to the copy cue may be quite highly correlated with the retrieval information available from contextual factors.

The following series of experiments deals with recognition failure in the paired-associate paradigm within the context of the dual recognition theory. We assume that the presentation code available in the paired-associate task is essentially the same as that available for any other task and is a function of exposure frequency and perceptual encodings. However, the conceptual code must, as we have noted before, depend on the manner in which the original material is conceptually encoded and is therefore task dependent. We postulate that the modal encoding of associated pairs is a unitary code of the pair, a single holistic unit stored in memory (cf. Greeno, 1970; Mandler, 1970). The important consequence of this assumption is that the retrieval strategies available to the subject when the second stage of the recognition process is entered must focus on the retrieval of the complete pair. Specifically, if the to-be-remembered (TBR) half of the pair cannot be recognized on the basis of the presentation information we assume that the individual tries to find a presented pair of which the TBR item is a member. The most important access to such a pair is by the process of “backward” retrieval, that is, by finding the list cue (LC) which “goes with” the TBR item. If the appropriate LC is found by completion of the stored pair, then the judgment of “old” can be made about the TBR item. Our prediction therefore is that failures to recognize items that are successfully recalled (in the standard “forward” recall test) can be shown to depend on failures to recall the pair (or the LC item) in the “backward” test.

Within this context we shall first present an experiment, with new material and somewhat different procedures than used by Tulving, in order to test the prediction from our theoretical position that recognition failure should be significantly affected by the criteria used for recognition judgments. Having demonstrated such a phenomenon, we then turn to the question of why different materials provide different kinds of levels of recognition failure. We present a replication of Tulving’s experiments and, in the process, a demonstration that recognition failure is item dependent. The next experiment deals with a generalization of the item dependency and a specific prediction about the kind of retrieval

1 We are using Tulving’s terminology here to describe a pair as LC–TBR, which is equivalent to the traditional usage of A–B. LC and A refer to the first or left-hand member of the pair, TBR and B to the second or right-hand member.
processes that must occur during the recognition of items from the paired-associate list. We conclude that recognition failure in the paired-associate paradigm is due to failure of access to the encoding used during presentation, namely, the holistic pair. We then assess the effect of variability in recognition on the magnitude of the recognition failure phenomenon. Finally, we present an experiment using a within-subject design and an intensive investigation of the recognition failure phenomenon which addresses the magnitude of the phenomenon when no item selection is taking place.

EXPERIMENT 1

Previous demonstrations of the recognition failure phenomenon have typically used a recognition test which is best described as a high criterion test. Subjects are shown lists of items which include old and new items and are asked to pick out (circle) items that are old. This task does not require detailed examination of all items and presumably only triggers items with a high level of the presentation code. As a result, the number of items subject to conceptual retrieval is very large and the effect of retrieval strategies should be pronounced. If, in contrast, a lower criterion is set, then the number of recognized items will increase, the set of items subject to retrieval will decrease, and the number of items showing recognition failure because of failure to retrieve the appropriate pair should therefore also decrease. Experiment 1 was conducted in two parts: For half of the subjects the recognition test was an “OLD ONLY” test, while for the other half a low criterion “OLD/NEW” test for all items was used. In addition, we developed the test lists de novo, rather than using those previously used in the literature.

Method

Design and materials. Three lists of 24 word pairs each were prepared. For each word pair the to-be-remembered (TBR) word was a weak associate of the list cue (LC). Weak associates were those words that were given as free associates to the LC 1% of the time in the norms of Bilodeau and Howell (1965) or Riegel (1965).

Each subject was tested with the three lists in the same order. The first two lists served as set-establishing (practice) lists. Each list was followed by a 4-min arithmetic task which served as a recency buffer. Subjects were then given the cued recall test, in which the 24 LCs were presented in a new random order.

The third list was the critical list. Following its presentation and the buffer task a recognition test was introduced. The recognition test was divided into two pages. The first page consisted of 16 of the 24 TBR words randomly intermixed with 16 filler items. Half of the filler items were also weak associates of the LC words whose TBRs were tested. For example, for the pair GLASS-VASE, EMPTY, which is also a weak associate of GLASS, was used as a filler. The remaining half of the fillers were unrelated to any of the TBR or LC items.

The second page of the recognition test consisted of 16 of the 24 LCs and 16 fillers. Eight of the LCs were from pairs from which the TBR was tested on the preceding page, while the remaining eight were from the pairs from which the TBRs had not been tested. Half of the filler items were words which also elicited the corresponding TBR word 1% of the time in the free association norms. For example, for the pair GLASS-VASE, the word LAMP, for which VASE is also a weak associate, was used as a filler. The remaining half of the fillers were unrelated to any of the TBRs or LCs.

There were three versions of the recognition test such that across subjects, each TBR and LC was tested equally often. Within each test both members of eight of the pairs were tested. For the remaining 16 pairs the TBR was tested from eight of the pairs while the LC was tested from the remaining eight pairs. Fillers were also used equally often across tests. Thus for a given TBR (or LC) a related filler was included on half of the tests, while an unrelated filler
was included on the remaining half of the tests. There were two groups of subjects. Half of the subjects were given OLD/NEW instructions for the recognition test, while the remaining half were given OLD ONLY instructions. The differences between these two groups will be elaborated in the procedures section.

Subjects engaged in another arithmetic buffer task following the completion of the recognition test. They were then given a final recall test. For half of the pairs the LC was printed on the left-hand side of the recall sheet and recall of the corresponding TBR was requested. For the remaining 12 pairs the TBR was printed on the right-hand side of the recall sheet and recall of the corresponding LC was requested. There were two versions of the recall test such that recall of the TBR given the LC as a cue and recall of the LC given the TBR as a cue were requested equally often for each pair. The two forms of the recall test were crossed with the three versions of the recognition test to create six unique test sequences.

Subjects and procedures. Twenty-four students at the University of California, San Diego, participated in the experiment as part of an introductory psychology course requirement. Half of the subjects were assigned to the OLD ONLY recognition group and the remaining 12 were assigned to the OLD/NEW recognition group. The subjects were tested in groups of six individuals each.

Upon entering the laboratory the subjects were told that they would be presented with a series of slides, each containing two words which were slightly related to each other. Both words were typed on the slides in capital letters, with a hyphen between them. Subjects were told that their task was to remember the words on the right-hand side of each slide and that the words on the left-hand side were to be used as cues to help them remember the words on the right. They were told that during the recall test they would be given a list of the left-hand words and asked to recall the corresponding right-hand words.

Following the instructions, the first list, which was designated as a practice list, was presented. For half of the subjects in each group each slide was exposed for 4 sec; the exposure duration was 3 sec for the other half of the subjects. The interstimulus interval was approximately 0.5 sec, the time necessary for the projector to recycle. After the presentation of the list was completed the subjects were given the buffer task and then the recall test, which consisted of the 24 left-hand (LC) words. The order of the LCs on the recall test was different from the presentation order. The subjects were instructed to write down the corresponding right-hand (TBR) word for each of the LCs. They were required to respond to each LC, even if they had to guess. No time limit was imposed on the recall test.

The instructions were then reviewed and the second practice list was presented and tested in exactly the same manner as the first list. The subjects were told that this was the first of the two experimental lists.

The same procedure was followed for the third, critical list, up through the buffer task following the list presentation. The recognition test was then introduced. Subjects in the OLD/NEW group were given a recognition test consisting of two pages. The first sheet contained 16 TBRs randomly interspersed with 16 fillers, arranged in a single column. Next to each word were printed the words NEW and OLD and the numbers 1, 2, and 3. The subjects were told that this list contained words that were right-hand members of the pairs that had just been presented as well as new words that had not been presented during the experiment. They were told to go through the list and for each word to circle OLD if they thought that it was a right-hand member of one of the pairs that had just been presented or NEW if they thought that it had not been presented. They were also instructed to give a confidence rating for each judgment, with 1 corresponding to "just guessing" and 3 corresponding to "absolutely sure." After completing the first page of the recognition test (no time limit was imposed), the subjects were told
that the second page contained left-hand members (LCs) of the pairs that had just been presented as well as new words. They then judged each of these words as old or new and gave confidence ratings for each word.

The subjects in the OLD ONLY group were also given a recognition test consisting of two sheets. These sheets, however, contained only the test items, arranged in a single column. These subjects were also told that the first sheet contained only right-hand members (TBRs) of the pairs that had just been presented as well as new words that had never been presented. They were instructed to go through the list and circle those words that they thought were right-hand members of the pairs that they had just seen. Similar instructions were given for the recognition of the LC items. Upon completion of the second recognition test the subjects were asked to go back and give confidence ratings for those items that they had circled. The subjects had circled the words that they thought were old in pencil; they were issued pens for the confidence ratings to ensure that they did not circle any more words.

Following the recognition test the subjects again engaged in an arithmetic buffer task. They were then given the final recall test. For half of the pairs the left-hand member was printed on the left-hand side of the sheet with a blank on the right-hand side of the sheet. For the remaining pairs the right-hand member of the pair was printed on the right-hand side of the recall sheet with a blank on the left-hand side. These two types of items were randomly intermixed on the recall sheet. The subjects were instructed to recall the corresponding member of each pair and were required to respond to each word. Again, no time limit was imposed on this task. Upon its completion the experiment was terminated and the subjects were debriefed.

Results and Discussion

A strict criterion was used for scoring all of the recall tests. That is, a word was scored as correct only if it was recalled in response to the appropriate list cue.

For both recall and recognition, performance was superior for the subjects in the 4-sec exposure duration groups. Exposure duration did not, however, interact with either the relation between these two measures or the two recognition tasks. We have therefore combined the 3- and 4-sec exposure duration groups for all of our analyses.

The mean proportion of words recalled from the first and second practice lists was .694 and .722, respectively, for the OLD/NEW group and .698 and .746 for the OLD ONLY group. There was no significant improvement between the two lists, $F(1, 22) = .80$, $MS_e = .02$ (A .05 criterion for significance was used on this and all subsequent statistical tests.) More important, there was no difference between the two groups, $F(1, 22) = .04$, $MS_e = .06$, before the critical list and treatment differences were introduced.

The data of primary interest are the recognition of the TBRs and the recall of the TBRs when given the LCs as cues. Each subject's recognition score was based on 16 observations; the recall scores were based on 12 observations. Mean recall and recognition scores for the two groups are presented in Table 1, as are the false alarm rates. $A'$, a nonparametric measure of discrimination (Grier, 1971), was used to compare any overall difference in discriminability between the OLD ONLY and OLD/NEW groups. The two groups did not differ on this measure, $t(22) = 1.29$. A $2 \times 2$ analysis of variance with Groups (OLD ONLY vs OLD/NEW) as a between-subjects factor and Tests (recall vs recognition) as a within-subjects factor showed that recognition (.797) was significantly higher than recall (.663), $F(1, 22) = 8.20$, $MS_e = .03$. The interaction between Groups and Tests was also significant, $F(1, 22) = 7.15$. Thus, while there was only a small difference in the hit rate between the two groups, .828 and .766 for OLD/NEW and OLD ONLY, respectively, recall of the OLD ONLY group (.757) was superior to that of the OLD/NEW group.
<table>
<thead>
<tr>
<th></th>
<th>Recognition</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>HR</td>
<td>FAR</td>
<td>Recognized</td>
<td>Not</td>
<td>Recognized</td>
<td>Not</td>
</tr>
<tr>
<td>Expt 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLD/NEW</td>
<td>.569</td>
<td>.828</td>
<td>.151</td>
<td>.573</td>
<td>.010</td>
<td>.292</td>
<td>.125</td>
</tr>
<tr>
<td>OLD ONLY</td>
<td>.757</td>
<td>.766</td>
<td>.026</td>
<td>.674</td>
<td>.116</td>
<td>.095</td>
<td>.116</td>
</tr>
<tr>
<td>Expt 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Related</td>
<td>.551</td>
<td>.487</td>
<td>.055</td>
<td>.337</td>
<td>.214</td>
<td>.150</td>
<td>.299</td>
</tr>
<tr>
<td>Unrelated</td>
<td>.550</td>
<td>.487</td>
<td>.083</td>
<td>.340</td>
<td>.209</td>
<td>.147</td>
<td>.304</td>
</tr>
<tr>
<td>List A</td>
<td>.510</td>
<td>.484</td>
<td>.070</td>
<td>.318</td>
<td>.193</td>
<td>.167</td>
<td>.323</td>
</tr>
<tr>
<td>List B</td>
<td>.591</td>
<td>.489</td>
<td>.068</td>
<td>.360</td>
<td>.231</td>
<td>.129</td>
<td>.280</td>
</tr>
<tr>
<td>Combined</td>
<td>.550</td>
<td>.487</td>
<td>.069</td>
<td>.339</td>
<td>.212</td>
<td>.148</td>
<td>.302</td>
</tr>
<tr>
<td>Expt 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>List A</td>
<td>.526</td>
<td>.521</td>
<td>.062</td>
<td>.375</td>
<td>.151</td>
<td>.146</td>
<td>.328</td>
</tr>
<tr>
<td>List B</td>
<td>.733</td>
<td>.558</td>
<td>.062</td>
<td>.467</td>
<td>.267</td>
<td>.092</td>
<td>.175</td>
</tr>
<tr>
<td>Expts. 2 and 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>List A</td>
<td>.518</td>
<td>.503</td>
<td>.066</td>
<td>.346</td>
<td>.172</td>
<td>.156</td>
<td>.326</td>
</tr>
<tr>
<td>List B</td>
<td>.671</td>
<td>.528</td>
<td>.066</td>
<td>.420</td>
<td>.251</td>
<td>.108</td>
<td>.221</td>
</tr>
<tr>
<td>Combined</td>
<td>.599</td>
<td>.516</td>
<td>.066</td>
<td>.385</td>
<td>.214</td>
<td>.131</td>
<td>.270</td>
</tr>
</tbody>
</table>
The false alarm rate for the OLD/NEW group (.151) was also significantly greater, \( F(1, 22) = 9.46, M_{S_p} = .01 \), than the false alarm rate for the OLD ONLY group (.026).

These results were as predicted. The OLD/NEW recognition test produced the expected criterion shift as shown by the increased hit and false alarm rates. As a result, the set of items subject to retrieval during recognition should be greater for the OLD ONLY group. This effect presumably accounts for this group's increased retrieval in the subsequent recall test.

The relation between recognition and recall can be seen more clearly by breaking the data down into a four-way contingency table in which each subject-item is classified as either recalled or not recalled and recognized or not recognized. For each subject there were eight items that were tested for both TBR recognition and TBR recall given the LC as cue. (For half of these pairs the LC was also tested for recognition.) Table 1 presents the proportion of these items, pooled over subjects, that fell into each of these four categories. The degree to which recalled words cannot be recognized is summarized in the recognition failure measure, which is defined as the conditional probability that an item was not recognized given that it was recalled. The recognition failure measures can be computed directly from the data in Table 1. For the OLD/NEW group, whose members were required to examine carefully every item in the recognition test and adopted a relatively low criterion for recognition, recognition failure was practically nonexistent (.017).

Recognition failure did occur for the OLD ONLY group, reaching a level of .147. Thus, manipulation of the recognition criterion alone can significantly affect the magnitude of the recognition failure phenomenon.

The level of recognition failure in the OLD ONLY group, however, was rather low compared with that typically achieved by Tulving and his associates. The next experiment therefore uses their lists and an "old only" design. In fact, we shall continue to use the "old only" criterion throughout the rest of these studies in order to maximize recognition failures for a proper investigation of their source.

**Experiment 2**

Experiment 2 was designed to be an exact replication, in terms of both procedure and materials, of Tulving's encoding specificity experiments. Tulving and his associates have previously found that both generation of candidates in a free association task prior to recognition (Watkins & Tulving, 1975) and the use of these subject-generated lists for the recognition test (Wiseman & Tulving, 1975) tend to decrease the overall level of recognition performance and consequently increase the level of recognition failure (Tulving & Wiseman, 1975). Therefore, this experiment attempted to replicate Watkins and Tulving's (1975) Experiment 5, as that is one of the few reported experiments which uses experimenter-prepared recognition tests and does not require any free association task prior to the recognition test.

**Method**

**Design and materials.** Two practice lists and two critical lists, each with 24 word pairs, were prepared. The TBR word of each pair was a low-frequency associate of the LC. The two critical lists were the two lists used by Watkins and Tulving (1975). The two practice lists were prepared by selecting those item pairs from the four lists presented by Wiseman and Tulving (1975) that had not been used in either of the two critical lists.

Each subject was tested with the two practice lists and then one of the critical lists. Half of the subjects were tested with List A and the remaining half with List B (see Watkins & Tulving, 1975). The word pairs were typed on slides, with the LC typed in lowercase letters directly above the TBR, which was typed in uppercase letters. The slides were presented at a 3-sec rate with a 0.5-sec interstimulus interval. Immediately follow-
ing the presentation of each practice list the subjects were given a cued recall test, in which the 24 LCs were presented in a new random order.

The third list was the critical list. Immediately following its presentation the subjects were given a picture memory task, which lasted 12 min and served as a recency buffer. Upon its completion the recognition test for the TBR items was introduced. The recognition test consisted of two pages, each containing 36 words arranged in 12 rows of 3 columns each. Each row contained one TBR item and two fillers. The TBR items were arranged so that they occurred in each column an equal number of times. For each TBR item there were two “related” fillers. The fillers were related to the TBR item in as much as all three words are high-frequency associates of a fourth word (which was never presented in the experiment). For example, for the pair DRINK–SMOKE the two related fillers were CIGARETTE and CHEW. All three of these words are high-frequency free associates of TOBACCO. Half of the TBR items appeared in the same row of the recognition test as their related fillers. The remaining half of the TBR items appeared with “unrelated” fillers, which were actually the related fillers for other TBR items on the test. Thus, each filler in the recognition test was related to a TBR item on the test; the related–unrelated distinction refers to whether or not the fillers appeared in the same row as the TBR to which they were related. Related and unrelated rows on the recognition test alternated. Two versions of the recognition test were prepared so that each TBR item was in the same row as either the related or unrelated filler items equally often.

The recognition task was divided into four parts. In the first part, a free recognition test, subjects simply circled those words that they thought were TBR (target) items from the last list. They were then given an alternative forced choice test, in which they were informed that there was one TBR item per row and then asked to circle exactly one word in every row. Subjects then gave confidence ratings, on a three-point scale, for each word that they had circled. Finally, they were asked to recall the LC associated with each word that they had circled as an old TBR item.

Following the completion of the recognition test the subjects were again given a picture memory task, which lasted 4 min and served as a recency buffer. The final recall test, in which the 24 LCs were listed with a blank space next to each one, was then administered. Four versions of the test were prepared; on each form the 24 LCs were presented in a different random order. The four versions of the recall test were crossed with the two versions of the recognition test to create eight unique test sequences.

Subjects and procedure. Sixteen subjects, from the same population as those in Experiment 1, participated in the experiment. Half of the subjects were assigned to critical List A and half to critical List B. The subjects were tested in two groups of eight.

The subjects were told that they would be shown a series of slides, each with two slightly related words, and that their task was to remember the capitalized target word on each slide. They were instructed to use the word on the top, printed in lowercase letters, as a cue to help them remember the target word. They were informed that for each of the three lists they would be given a recall test in which they would be required to recall the target words, given their respective cue words. The first practice list was then presented, followed immediately by the cued recall test. The subjects were reminded to attempt to recall the TBR item that they thought had been presented with each LC. They were not required to produce a response for every LC, as they were in Experiment 1. Upon completing the recall the instructions were reviewed, any questions were answered, and then the second list was presented and tested for recall. The instructions were then again reviewed, with no indication of any future change in procedure, and then the third, critical list was presented.

Following the presentation of the critical list the visual memory buffer task was introduced.
Six slides of line drawings of real-world scenes were presented for 20 sec each, with 2 min between slides. The subjects were instructed to study each picture as it was presented and then draw it, as best they could, during the interval between slides.

The recognition test was then announced. The subjects were told that the test contained the capitalized target words from the last list as well as new items that had never been presented. They were asked to go through the list and circle those words that they thought were old. After completing this task the subjects exchanged their pens for pencils. They were then informed that there was one target word per row and asked to go through the test again and circle exactly one word per row. Finally, they were asked to recall the corresponding LC for each word that they had circled as an old target word.

The subjects were then given the picture memory task again. It lasted 4 min this time, with only two pictures being presented. The final cued recall test was then administered. As before, for each of the 24 LCs the subjects attempted to recall the corresponding TBR (target) item. No time limit was imposed on this or any of the other recall or recognition tests. Following completion of recall, the experiment was terminated and the subjects were debriefed.

Results and Discussion

As in Experiment 1, a strict criterion was used for scoring the recall tests. The recognition data are presented for the free recognition test only. The data, shown in Table 1, are presented separately for the related and unrelated filler conditions (collapsed over lists) and for the two lists (collapsed over filler conditions) and are summarized for the entire experiment (collapsed over both the list and filler factors).

Clearly, there are no differences between the related and unrelated filler conditions, thus confirming the findings of Watkins and Tulving (1975). While some might find this result surprising, we hasten to note that the distinction between related and unrelated fillers is relevant only to each row of the recognition test. Considering the test as a whole, all of the fillers are related. And indeed, every one of our subjects proceeded through the free recognition test by going down each column, not across the rows. Thus, in terms of the subjects’ perceptions, this is really an artificial factor and will therefore not be discussed further.

The phenomenon of primary interest is the degree to which recalled words cannot be recognized. With these lists, recognition failure appeared to be a rather robust phenomenon, achieving a magnitude of .385. Thus, over a third of the words which were recalled to the LC could not be recognized by themselves.

Although Wiseman and Tulving (1976) now disregard the relative overall levels of recall and recognition, these data do warrant brief consideration. While recall was slightly better than recognition (by 6%), this phenomenon appears to be confined to List B, in which recall was 10% better than recognition, as compared with a difference of only 2% on List A.

Comparing Experiments 1 and 2, we found a sizable amount of recognition failure in this experiment, but not in Experiment 1. It is doubtful that the differences were due either to the procedural variations or to a problem of subject selection. The most obvious, and most likely, source of the variation was the lists. If the obtained differences in recognition failure and the overall levels of recall and recognition were due to the specific item pairs used, then we should be able to demonstrate item-specific effects. We therefore examined the data on an individual item basis. Inspection of these data revealed large differences between items; some were easier to recall and evidenced recognition failure, others were easier to recognize, and many were recalled and recognized equally well.

If our original position is correct, then there may in fact be some pairs that, by the nature of the constituent items, are easy to recover in
forward recall, others in backward recall, and some for both. These differences might account for the differential occurrence of recognition failure between lists and items. If these item differences are stable we should be able to demonstrate an item correlation for recall and recognition across two experiments. Experiment 3 was therefore performed in search of differences in recallability and recognizability among items.

**EXPERIMENT 3**

**Method**

Experiment 3 is a replication of Experiment 2, with two simplifications in the recognition procedure. Both the alternative forced choice recognition test and the recall of the LCs for recognized items were omitted. In all other respects, Experiment 3 follows the exact same design and procedure as Experiment 2. Eighteen subjects, from the same source used in the previous experiments, participated in this experiment. Eight subjects were tested on List A and ten on List B.

**Results and Discussion**

The recall and recognition breakdown is shown in Table 1. Again we find a sizable degree of recognition failure (.335), with recall (.641) better than recognition (.542). As a prelude to our discussion of item-specific effects, we again note differences between Lists A and B. The overall levels of recall and recognition are practically identical for List A, while for List B recall was 17.5% higher than recognition. For the sake of completeness, we also present separately in Table 1 the data for Lists A and B combined over Experiments 2 and 3, as well as the data for the two experiments combined.

Our primary concern centers on specific differences between items in recognition failure. That is, are some items consistently easier to recall than to recognize and thus more likely to exhibit recognition failure than others? We therefore performed a correlation across the two experiments on the number of occurrences of recognition failure (recalled and not recognized) for each item. The correlation was significant, $r(47) = .516$.

We should note that the number of occurrences of recognition failure per item is not the exact analog of the recognition failure measures previously reported. Those measures are ratios of the number of recalled and not recognized subject-items to the total number of recalled subject-items. In this case, however, number of occurrences of recognition failure is the more appropriate measure, because our concern is the extent to which certain items are easier to recall than to recognize, independent of the overall level of recall for that item.

The other side of the recognition failure coin is recall failure, that is, the extent to which certain items are easier to recognize than to recall. The correlation for the number of times each item was recognized and not recalled, across the two experiments, was significant, $r(47) = .475$. Thus some items are easier to recall than to recognize and some are easier to recognize than to recall. These two distributions are not, however, nonoverlapping. That is, some subjects may recall a given item and not recognize it, while other subjects may recognize and not recall that same item. However, this is more the exception than the rule. Thus, there should also be a correlation in the difference between overall recall and overall recognition for each item. (This measure is, of course, equivalent to the difference between the number of occurrences of recognition failure and the number of occurrences of recall failure for a particular item.) The correlation performed on these difference scores was also significant, $r(47) = .557$.

These correlations suggest that there are specific differences between items, which can thus lead to the differences that we observed between Lists A and B. Specific item differences are also likely to be a large factor in our failure to find a high degree of recognition failure in Experiment 1.
Finally, and most important, the item correlation for recognition failure is completely consistent with our theoretical view of the processes underlying recognition failure. We suggested that for those items about which the subject is uncertain on the basis of presentation information an attempt is made to retrieve the other (LC) member of the pair (backward retrieval). Recognition failure occurs for those items for which backward retrieval fails (and thus the items are not called "old") while forward retrieval succeeds (the subject can recall the TBR item, given the LC). Thus, we interpret the recognition failure correlation as indicating consistent differences between the ease of forward and backward retrieval for different item pairs (retrieval asymmetry). Even though backward retrieval is expected to be generally poorer than forward recall, there remain specific differences between item pairs as to the relative ease of these two retrievals. Recognition failure is most likely to occur for those pairs for which forward retrieval is superior to backward retrieval.

**EXPERIMENT 4**

Experiment 4 is offered in specific support of our view that recognition failure is due to retrieval failure. We noted earlier that when the presentation information provided at the time of presentation of the copy cue (TBR) is inadequate to make an "old" recognition decision it is often necessary to access the encoding used during presentation, the holistic pair, in order to call that item "old." Accessing the holistic pair encoded during presentation implies the retrieval of the other member (the LC) of the pair. We refer to this type of retrieval as backward retrieval in contrast to the retrieval of the TBR item, given the LC, which is forward retrieval. If backward retrieval is attempted during recognition and fails, then the item will not be called "old," that is, it will not be recognized. Thus, if backward retrieval is unsuccessful, while forward retrieval (recall) for that item pair is successful, we will have an occurrence of recognition failure. This is, of course, a fairly likely situation, given that we instruct our subjects to study the items so that they can recall the TBR item, given the LC. That is, our instructions stress forward retrieval, not backward retrieval.

The most obvious way to test a hypothesis which holds that recognition failure is due to a failure in backward retrieval (followed by success in forward retrieval) is to include a direct test of backward retrieval. Therefore, we will present subjects with all of the TBR items and ask them to attempt to recall each item's corresponding LC. This test will occur after the recognition test and before the traditional forward recall test. Two major predictions follow directly from our hypothesis. First we shall compute a backward recognition failure measure, that is, the proportion of items recalled backwards and not recognized of all items recalled backwards. Given that the same type (direction) of retrieval is required for both recognition and backward recall, backward recognition failure is much less likely to occur. Thus, we expect backward recognition failure to be substantially less than forward recognition failure.

We also emphasize that given our instructions for studying the item pairs, backward recall will be worse than forward recall. The recognition failure measures, however, are conditional probabilities, conditionalized on the overall level of recall (forward or backward). Thus, our comparisons of forward and backward recognition failures are independent of differences between the overall levels of forward and backward recall. Notwithstanding these considerations, we must predict that backward recognition failure will be substantially less than forward failure.

Our second prediction concerns those items for which forward recognition failure occurs. If forward recognition failure is due to a failure of backward retrieval during recognition, then backward retrieval should also fail on the more direct, backward recall test. Thus, we predict a
failure in backward retrieval (recall) for those items that are forward recalled and not recognized.

One final prediction follows from the preceding one. In the previous two experiments we found item-specific differences for recognition failure. If recognition failure for a particular item is due to a greater likelihood of forward retrieval as compared to backward retrieval, then these recall differences among items in the present experiment should be correlated with recognition failures in the previous experiments.

Method

This experiment follows the methodology of Experiment 3, with a backward recall test inserted between the recognition test and the forward recall test.

Four versions of the backward recall test were prepared. On each version the 24 TBR items were listed in a different random order. A space was provided to the left of each item for the subject's response. The backward recall test was introduced as soon as the buffer task following the recognition test was completed. The subjects were told that these were the TBR (target) items from the last list and were instructed to try to recall the corresponding LC for each TBR item. The final forward recall test was administered as soon as the backward recall test was completed. In all other respects, the procedures and methodology of this experiment were identical to those of Experiment 3.

Sixteen undergraduates, from the same source used in the previous experiments, participated in this experiment. Half of them were assigned to List A, the remainder to List B. They participated in two groups of eight.

Results and Discussion

Each subject-item can be classified as recognized or not recognized, backward recalled or not backward recalled, and forward recalled or not forward recalled. The data can thus be completely and uniquely described in an eight-way contingency table. Table 2 presents such eight-way contingency tables for each of the two lists as well as the combined data, on which our discussion will center. Various summary measures derived from pooled data are presented in Table 3. All of the measures that we shall be discussing can be computed directly from the data in Table 2.

As expected, forward recall (.563) was superior to backward recall (.289). The hit rate in the recognition test was .570; the false alarm rate was .042. Of prime concern are the recognition failure measures. We predicted that forward recognition failure would be greater than backward recognition failure. This was indeed the case. Forward recognition failure (.306) was twice the magnitude of backward recognition failure (.153).

Our second prediction concerned those items for which forward recognition failure occurred. If recognition failure is due to a failure of backward retrieval during recognition then backward recall should also

<table>
<thead>
<tr>
<th>Recognized</th>
<th>Not recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FB</td>
</tr>
<tr>
<td>List A</td>
<td>.292</td>
</tr>
<tr>
<td>List B</td>
<td>.182</td>
</tr>
<tr>
<td>Combined</td>
<td>.237</td>
</tr>
</tbody>
</table>

*F, forward recalled; FB, not forward recalled; B, backward recalled; FB, not backward recalled.
fail. This was the case for 77.3% of those items. That is, over three-quarters of all the items that were forward recalled and not recognized were not recalled backwards. The probability of backward recall for these items that were forward recalled and not recognized is significantly less than the probability of backward recall for those items that were forward recalled and recognized, \( t(15) = 3.27 \).\(^2\)

Given that most forward recognition failures are due to failures in backward retrieval, we now wish to assess the magnitude of the phenomenon once failures in backward retrieval are accounted for. Of the items that were forward recalled and not recognized, 22.7% were also backward recalled. Forward recognition failure (which was .306), corrected for failures in backward retrieval, is then \(.306 \times .227 = .069\). Thus, the recognition failure phenomenon, due to factors other than a failure in backward retrieval, is really quite small. Another way of estimating the effect of backward recall on forward recognition failure is to determine the traditional recognition failure measure separately for those sets of items that were previously backward recalled and for those that were not. The resulting estimate (from Table 2) of recognition failure for backward-recalled items is .141, while it is .463 for items not backward recalled. Again, backward accessibility significantly modulates recognition failure.

Examining the two lists separately, we find once again that forward recall is slightly better than recognition on List B, while the reverse is true for List A. Examining this difference in more detail, we find that the overall level of forward recall is practically the same for List A and List B (.568 and .557, respectively). Recognition on List A (.630), however, is 12% higher than recognition on List B (.510). This is mirrored in a 12% difference in backward recall (.349 and .229 for Lists A and B, respectively). Or, as we would prefer to say, the differences in the recognition rates reflect the differences in backward accessibility (retrieval) of the pairs. These differences in backward accessibility are also reflected in the forward recognition failure measures. That is, forward recognition failure is greater on List B (.365) than on List A (.248) because backward retrieval is more difficult for the items on List B.

---

\(^2\) The data reported in the tables are pooled across subjects, but these statistical tests are \( t \) tests performed on individual subjects' data.
We claimed earlier, in connection with our prediction that forward recognition failure would be greater than backward recognition failure, that these measures are independent of differences in the overall levels of recall, because they are conditional probabilities. A demonstration of this point can be seen in the backward recall data. We noted that backward recall on List A was 12% higher than on List B. Recognition failure is, however, independent of this difference. Backward recognition failure is .149 on List A and .159 on List B.

In Experiments 2 and 3 we demonstrated that recognition failure is an item-specific phenomenon. We hypothesized that this was due to item-specific differences in the relative ease of forward and backward recall for different pairs. If this is true, we should be able to demonstrate a correlation between recognition failure in the previous two experiments and recall differences in the present experiment. In order to obtain a more reliable estimate of item differences in recognition failure we shall use the mean number of occurrences of recognition failure per item from Experiments 2 and 3. We predicted that recognition failure should occur when forward recall for a given item pair is easier than backward recall for that item pair. The difference score, of forward recall less backward recall, should thus be correlated with recognition failure. The correlation was significant, \( r(47) = .503 \).

As an alternative to difference scores, we can compute the correlations of recognition failure in the previous experiments directly with forward and backward recall in the present experiment. From these correlations we can compute the more relevant partial correlations. The partial correlation of recognition failure with backward recall (when forward recall is held constant) was significant, \( r(47) = -.423 \). Thus, as predicted, the probability of recognition failure for a given item increases as the probability of backward recall for that item pair decreases.

We also find that with backward recall held constant, the correlation of recognition failure with forward recall is .496. This correlation is, however, trivial. Given that recognition failure (which subsumes forward recall) still tends to occur for the same items (and it does; the correlation of recognition failure in the previous experiments with recognition failure in this experiment is .677), then, by definition, recognition failure of items will increase in Experiments 2 and 3 as forward recall of these items increases in Experiment 4.

Summarizing, Experiment 4 has provided direct support for our theory that recognition failure is just another case of retrieval failure. That is, recognition failure is most likely to occur when backward retrieval fails and forward retrieval succeeds. Recognition failure due to other factors occurs rather infrequently.

**Experiment 5**

There are two conditions under which some cases of backward recognition failure can and should occur. The first involves items whose presentation code falls below the low criterion. These old items are automatically called new, and no retrieval is attempted during recognition. However, these items may well be retrieved during recall. One test of this prediction is to take advantage of moment to moment variability in the presentation code. An item may fall below the criterion at one test but not at a subsequent test. Thus, if subjects are tested twice for recognition, then some items that are not recognized on the basis of low presentation codes on Test 1 should exceed that criterion on Test 2, be subjected to a retrieval test, and be successfully called old. Therefore, if the criterion for failure to recognize an item is a failure on both tests, then the value of recognition failure for backwardly recalled items should be less than the value obtained from a single recognition test. We tested this prediction in Experiment 5 and replicated the major findings of Experiment 4.

The second condition that might produce such recognition failure depends on variability in accessibility, that is, in the conceptual code.
Since the dual access hypothesis states that the code is used during both recognition and recall, variability in the code might produce pairs that fail to be retrieved during the recognition test, but are accessed during the backward recall test. These would then produce backward recognition failures. However, it is likely that this effect is very small since Wiseman and Tulving (1976) found very little, if any, variability in forward retrieval.

Concerning variability in general, Wiseman and Tulving (1976) note appropriately: "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" (p. 360). Curiously enough though, they tested the above hypothesis by administering two successive forward recall tests. There was very little variability between the two recalls and they therefore rejected the variability hypothesis. But our, and their, major concern is with recognition failure of recallable items. As we noted above, the more appropriate way to test for "variability in the accessibility of stored information" is to administer two successive recognition tests. Evidence of variability in recognition can then be obtained. The major issue to be addressed is: How much of recognition failure can be attributed to variability in the subjects' recognition performance?

We will again include the backward recall test, prior to the forward recall test, which will allow us to assess both forward and backward recognition failure and the relations between them. Given the considerations discussed earlier, we should expect backward recognition failure to occur very infrequently, especially after recognition variability is taken into account.

Method

Sixteen undergraduates, from the same source used in the previous experiments, participated as subjects in this experiment. Half of the subjects received List A as their critical list, the remainder received List B. The subjects participated in two groups of eight individuals each.

The procedures and methodology were identical to those of Experiment 4, up through the completion of the free recognition test. The second recognition test was then distributed. The subjects were told to go through this second test and again circle those words that they thought were target (TBR) items from the last list. They were instructed to reevaluate each item and not to worry about their responses on the previous recognition test. Confidence ratings were then obtained for those items marked "old" on the second recognition test.

The two recognition tests were the two forms used in the previous experiments. Thus, both tests contained all the TBR items and the identical fillers. The location of each item, as well as the surrounding items, differed between the two test forms. Each test form was administered first equally often.

Upon completion of the second recognition test, the 4-min buffer task was performed. Upon its completion the recall tests were administered. As in Experiment 4, the subjects were given the backward recall test followed by the forward recall test. The instructions and procedures used for the recall tests were the same as those used in the previous experiment.

Results and Discussion

Each subject-item can be classified as recognized or not recognized on each of the two recognition tests, backward recalled or not backward recalled, and forward recalled or not forward recalled. Thus, a 16-way contingency table is required to completely describe the data. Table 4 presents such 16-way contingency tables for each of the two lists as well as for the combined data, on which our discussion will center. Various summary measures, which can be calculated from the data in Table 4, are presented in Table 3. Where applicable, these measures are calculated separately for (a) the first recognition
Our major concern in this experiment was the degree of variability in recognition. This variability is, in fact, substantial. Of the items not recognized in the first test 16.8% were subsequently recognized in the second recognition test. On the other hand, 15.8% of the items recognized on the first test were not recognized on the second test. Summarizing the overall degree of recognition variability, we note that 24.1% of those items that were recognized were recognized on only one of the two tests.

As was found in Experiment 4, forward recognition failure (.321) is much greater than backward recognition failure (.139). These measures are based only on the first recognition test. In the introduction to this Experiment we derived the prediction from the dual access hypothesis that the amount of recognition failure should be reduced when recognition is based on two tests. If those items that were recognized on the second recognition test are also taken into account, both of the recognition failure measures are substantially reduced. Forward recognition failure drops to .253 and backwards recognition failure drops to .061. In fact, 10 of the 16 subjects showed no (0.000) backward recognition failure when both recognition tests were taken into account.

Again we find that forward recognition failure is due predominantly to a failure in backward retrieval. Over three-quarters of the items for which forward recognition failure occurred were not recalled backward. The probability of backward recall of those items for which forward recognition failure occurred was significantly less than the overall probability of backward recall; $t(12) = 4.39$ for both recognition tests combined.\(^3\) We can estimate the magnitude of forward recognition failure due to factors other than failure of backward retrieval by multiplying the forward recognition failure measure (the probability of nonrecognition given forward recall) by the proportion of those items that were previously backward recalled. Forward recognition failure then, corrected for backward access, is .078 when based only on the first recognition test; it drops to .034 when both recognition tests are combined.

The central findings of Experiment 4 have now been replicated, supporting our view that traditional (forward) recognition failure is due, in large part, to a failure in backward retrieval. Furthermore, a large degree of variability in recognition was found, and when this variability is taken into account recognition failure is substantially reduced. Thus, traditional estimates of recognition failure are inflated.

\(^3\) Three subjects were dropped from this analysis because they had zero recognition failure.
EXPERIMENT 6

All of our previous experiments have investigated the phenomenon of recognition failure under rather constrained conditions. The most obvious of these is, of course, the specific lists used. For Experiments 2 through 5, we used those lists that were used by Tulving and his associates in many of their experiments. Other investigators have also used these same lists in their studies of recognition failure and the relationship between recall and recognition in the paired-associate paradigm (e.g., Postman, 1975). Given our demonstration that recognition failure is item specific it becomes necessary to assess the magnitude of this phenomenon under more varied conditions. While other investigators (e.g., Reder, Anderson, & Bjork, 1974; Salzberg, 1976) have attempted to define the limiting conditions under which recognition failure occurs by systematically varying particular properties (grammatical class, concreteness, and relatedness) of the item pairs used, we prefer to assess the magnitude of the phenomenon under the most general of item conditions. We shall therefore allow such factors to vary freely and rid ourselves of the problems of item selection by using word pairs that are randomly generated.

Another limitation of most of the investigations of the recognition failure phenomenon is that subjects are relatively unpracticed at the tasks required of them and thus unlikely to have developed appropriate strategies. Mandler (1976) has stressed the importance of studying memory phenomena under “steady-state” conditions, that is, those in which the subjects have developed appropriate strategies for the task at hand. He notes that in everyday life people “typically have used their memory capacities over many years in a variety of different tasks, presumably with some economy of strategies and mechanisms.... In contrast, the laboratory experiment very often presents an individual with a novel kind of task which needs analysis and exploration before the relevant strategies can be found, strategies that typically have been used relatively infrequently in the past” (p. 5). However, most of our experiments and those of Tulving and his associates [with the exception of Wiseman and Tulving (1975)] present the subjects with two practice lists which are tested for recall and then the critical list which is tested for both recognition and recall. While it is doubtful that two practice lists are sufficient for the development of the strategies necessary for mastering paired-associate recall, the subjects certainly have not been able to develop the strategies necessary to deal effectively with a recognition test for TBR items. Yet it is precisely the failure to recognize items with which we are primarily concerned.

How much of the recognition failure phenomenon is due simply to the subjects’ initial lack of effective strategies for recognition? Will subjects develop appropriate and effective strategies for dealing with the recognition test (and recall test) after extended practice? How will this practice affect the magnitude of the recognition failure phenomenon? Experiment 6 addresses these issues using unrelated (randomly generated) item pairs and a within-subject design, in which subjects are tested for recognition, forward recall, and backward recall for each of 12 separate lists.

Method

Subjects. Six graduate students, from departments other than Psychology, at the University of California, San Diego, participated as paid subjects. Each subject participated in 12 sessions. Experimental sessions were separated by at least 1 day and never more than 1 week. The subjects were tested individually.

Materials and apparatus. An item pool of 550 high frequency (Kučera & Francis, 1967) nouns of six letters or less was prepared. Words that were homophones of other words were excluded from the item pool. Prior to the start of each experimental session, for each subject, 96 words were randomly drawn from the item pool. The first
24 words drawn were designated as the 24 LCs; the second 24 were designated the corresponding TBR items. Thus, new item pairs were randomly created for each subject's 12 sessions. The remaining 48 words were used as fillers on the recognition test.

All phases of the experiment were controlled by a PDP-12 computer. The computer program controlled the selection of the items for each test and randomized the sequences of items for each test. Items were presented to the subject on a remote CRT. All of the subjects' responses were oral. They were monitored by the experimenter who was located in an adjoining room.

**Design and procedures.** Each subject was presented with one list per session. After a short buffer task the recognition test for the TBR items was administered. After another short buffer task the item pairs were tested for both forward and backward recall. Each pair was tested both ways. Half of the pairs were tested for forward recall first and half were tested for backward recall first.

The item pairs were presented for 4 sec each, with a 1-sec ISI. The words of each pair were displayed on the scope side by side, with a hyphen between them. After the last pair had been presented a three-digit number appeared on the screen. This served as a signal to the subject that the last pair had been presented and initiated the buffer task. The subject then counted aloud backwards by 3s for 45 sec, until the words PLEASE WAIT FOR INSTRUCTIONS were displayed on the screen.

The recognition test was then initiated. Twenty-four words, arranged in eight rows of three columns, were displayed. The subject went through these words and called out those that he thought were TBR items from the presented list. The words remained on the screen until the subject indicated that he was finished. An array of 24 more words was then presented. Again the subject identified those words that he thought were old, and then the final set of 24 items was presented for recognition. The positions of the 24 targets and 48 fillers within the recognition test were completely random. Thus, there was no fixed number of TBR items in any given array, column, or row.

After the recognition test was completed another three-digit number was presented and the subject again counted backwards by 3s for 45 sec until the words PLEASE WAIT FOR INSTRUCTIONS were presented. The recall test was then initiated. Each pair was tested for both forward and backward recall. Half of the pairs were randomly selected to be tested for forward recall first and the remaining half were tested for backward recall first. After each pair had been tested once, each was tested again. Those pairs that had been tested for forward recall in the first half of the test were tested for backward recall and those that had been first tested for backward recall were now tested for forward recall. The order in which the items were tested was randomized for each half of the test. Forward and backward recall were also randomly intermixed.

The subjects knew whether each word was an LC or a TBR item by its position on the screen and thus whether forward or backward recall was required. On forward recall trials the LC was presented on the left-hand side of the screen with a hyphen in the center of the screen. On backward recall trials the TBR item was presented on the right-hand side of the hyphen. The subject was required to say the presented word out loud and then to try to recall the other member of the pair. Each word remained on the screen for 5 sec but the subjects were allowed another 5 sec in which to respond. At the end of this period the next item was presented. If the subject made no response within the allotted 10 sec that pair was scored as nonrecalled.

The identical procedures were followed in each of the 12 sessions and thus the subjects knew exactly what was expected of them. They knew that they would get a recognition test for the TBR items of each pair and that they would be required to recall both members of each pair when given the other member as a cue. At the end of each session the subjects
were given complete feedback on their performance. They were able to see exactly how well they did on each of the three tests as well as the fate of each item pair in each of the three tests. The subjects were well motivated and anxiously examined their performance at the end of each session. All three tasks were emphasized equally and each of the subjects strove to do as well as possible on each of the tests.

**Results and Discussion**

Somewhat surprisingly, the effects of practice were accomplished rather quickly. Performance on all three tasks improved over the first few sessions and then reached a plateau. Both recognition and backward recall reached a plateau by Session 4, while forward recall reached a plateau by Session 5. Various comparisons of interests, such as forward recognition failure and backward recognition failure, were examined for each session. No systematic trends in their relationships to one another could be detected. We shall therefore present the data combined over all 12 sessions.

The data are presented in Table 5 in two eight-way contingency tables. Each item was classified as either recognized or not recognized, forward recalled or not forward recalled, and backward recalled or not backward recalled. The proportions of items falling into each of these eight unique categories are the data presented in Table 5, separately for those items for which forward recall was tested first and backward recall second and for those items for which backward recall was tested first and forward recall second. Various summary measures are presented in Table 3.

Even with extended practice and knowledge of what was required of them, forward recall (.600) was better than backward recall (.500). Recognition was better than both types of recall; the hit rate was .724 and the false alarm rate was .030.

Again, forward recognition failure was almost twice that of backward recognition failure. Forward recognition failure, for those items for which forward recall was tested first, was .130. Backward recognition failure, for those items for which backward recall was tested first, was .074. But note that both of these measures are roughly half of the magnitude of those obtained using unpracticed subjects and related word pairs (Experiment 4).

We can also examine recognition failure based on the second recall. For those items for which forward recall was tested second, the forward recognition failure rate was .146, which is comparable to when forward recall was tested first (.130). As predicted though, more than two-thirds (.684) were not recalled backwards. Again, the probability of backward recall for these items was significantly less than the overall probability of backward recall, \( t(5) = 2.37 \). The effective forward recognition failure rate, corrected for failures in backward retrieval (which were \( 1 - .684 = .316 \)), is \( .146 \times .316 = .046 \).

Backward recognition failure, for those items for which backward recall was tested second, was .151. This measure is obviously inflated though, because many items are recalled backwards because they were previously recalled forwards. To see this more clearly, note that on the second recall test, \( .675 \) of the items were backward recalled, as opposed to \( .500 \) when backward recall was tested first. Thus, the forward recall test is responsible for a 35% increase in subsequent backward recall. It is therefore necessary to examine backward recognition failure
separately for those items previously recalled and those not previously forward recalled. Of those items that exhibit backward recognition failure, .628 were previously forward recalled and .372 were not. The effective backward recognition failure rate, uncontaminated by being previously forward recalled, is \(.151 \times .372 = .056\), which is comparable to the backward recognition failure rate when backward recall was tested first (.074).

Summarizing these data, we see that even with well-practiced subjects and unrelated item pairs, about 15% of those items that can be recalled fail to be recognized. Furthermore, two-thirds of these recognition failures can be attributed to a failure in backward retrieval. One may consider backward recognition failure to be the more interesting case of failing to recognize words that can be recalled because the same type of retrieval (backward) is required in both tasks, but the magnitude of this phenomenon is very small.

**EXPERIMENT 7**

The final experiment examines again the question of recognition variability, but now in the context of the steady-state performance produced by the subjects in Experiment 6. In particular, we wish to inquire how much the level of backward recognition failure can be reduced under these optimal conditions.

**Method**

Our indefatigable subjects from Experiment 6 were recalled for two more sessions. In this experiment, each subject was given two recognition tests, separated by 1 min of counting backwards, before they went on to the recall test. The two recognition tests contained exactly the same items, but their positions within the second test were randomized anew.

The subjects were not informed of the second recognition test until after they had finished counting backwards upon completing the first recognition test. As in Experiment 5, they were instructed to reevaluate each item and not to worry about how they had responded on the previous test.

When the subjects returned for their second session they were aware of the fact that they would again be given two recognition tests.

In all other respects, the procedures and methodology were the same as those of Experiment 6.

**Results and Discussion**

We shall only briefly summarize the relevant findings from this experiment. We will not present the complete data breakdown, but the relevant summary measures are presented in Table 3.

Combining the data from the two sessions, the hit rates for the first and second recognition tests were .799 and .785, respectively. Of those items not recognized on the first test, 17.2% were recognized on the second test. Of those items recognized on the first test, 6.1% failed to be recognized on the second test. As an alternative index of recognition variability, 10% of the items that were recognized were recognized on only one of the two tests. Thus, even with well-practiced subjects there remains a fair degree of recognition variability. It is, of course, less than that obtained with unpracticed subjects (Experiment 5).

The proportion of items forward recalled, when forward recall was tested first, was .750. Forward recognition failure, based only on the first recognition test, was .120. However, when we include as recognized those items that were also recognized on the second recognition test, forward recognition failure drops to .074.

When backward recall was tested first, .639 of the items were recalled. Backward recognition failure, based on the first recognition test, was .054. When those items that were recognized on the second recognition test are included, backward recognition failure is reduced to .043.

Forward recognition failure for those items for which the forward recall test came second was .130, based only on the first recognition test. This measure is reduced to .122 when
both recognition tests are combined. As expected, most (.714) of these items were not retrieved on the backward recall test. [Again, the probability of backward recall for these items that were forward recalled but not recognized was significantly less, \( t(5) = 2.20,^2 \) than the overall probability of backward recall.]. Thus, the effective recognition failure rate, corrected for backward access, is \( .122 \times .286 = .035 \).

Backward recognition failure, when backward recall was tested second, was .245; it was .094 if both recognition tests were combined. But again, as was found in Experiment 6, most of these items were previously forward recalled, and thus these measures are inflated. Backward recognition failure for those items that were not previously forward recalled was .034; it was .026 if both recognition tests were combined.

Clearly then, the findings of Experiment 6 have been replicated. Furthermore, as was found in Experiment 5, when recognition variability is taken into account the magnitude of recognition failure (both forward and backward) is reduced even further.

**General Discussion**

According to Wiseman and Tulving (1975) the primary use of the encoding specificity principle is in interpreting results from experiments in which the differential effectiveness of various retrieval cues is examined and found to be “not readily accommodated by existing theory.” These phenomena, whose interpretation can benefit from the application of the encoding specificity principle, are referred to as encoding specificity phenomena. The encoding specificity phenomenon on which this research has centered is that of recognition failure of recallable words. In recognition failure, an item is not recognized but it is recalled; thus, by definition, the nominal copy of the target word is a less effective retrieval cue than the word with which the target was paired at presentation. To account for this result the encoding specificity principle holds that an item is encoded into memory in relation to its cognitive environment (retrieval cues) and that access to the stored representation during testing is dependent on a reinstatement of the original input cues. Thus, with respect to recognition failure the TBR is encoded with the LC and therefore the LC is a more effective retrieval cue than the nominal copy of the TBR itself.

The astute reader will have noted that the encoding specificity principle does not explain recognition failure; it describes it. That is, the LC is a more effective retrieval cue than the TBR itself because the TBR was encoded with respect to the LC and can therefore only be accessed through the LC. Indeed, the encoding specificity principle is not supposed to offer an explanation of recognition failure. As Wiseman and Tulving (1975) note, an explanation of recognition failure (and other encoding specificity phenomena) “is still not in hand, and the search for an explanation constitutes an important research problem.” What the encoding specificity principle does do is “provide a general conceptual framework in which solutions to problems posed by these phenomena can be sought.”

Thus, Wiseman and Tulving seem to be calling for an underlying mechanism that allows recall to succeed while recognition fails. That is exactly what we have offered. However, this is not the path that most researchers have followed. Most research has been directed toward carefully delimiting the conditions under which encoding specificity phenomena are obtained. As an example of this approach, consider the recent work of Salzberg (1976), which is concerned with the generality of the phenomenon across different materials. He varied both the grammatical class and concreteness of the LCs while holding both the TBR (an abstract adjective) and the degree of relatedness between the LC and TBR constant. He found that encoding specificity phenomena were obtained only for the highest levels of cue concreteness and, together with other evidence, concluded that it is the relative concreteness of each member of
the pair which is the critical variable for encoding specificity effects to occur. Wiseman and Tulving have noted that Salzberg uses the superiority of recall over recognition as his criterion for an encoding specificity effect, rather than the recognition failure measure. In fact, he obtained a sizable degree of recognition failure in a number of conditions, and we shall return to this point later. Salzberg proposes that the pair is encoded into some sort of unit and that subsequent access to the encoded unit is required for recall or recognition of the target word. Up to this point his position is quite similar to ours. He then determines under what conditions one member of a pair should provide better access than another and concludes that the more concrete member of the pair is the more effective cue in providing access to the encoded unit. He suggests that the more concrete word is probably the more salient part of the image evoked by the pair and thus most likely to reinstate that image and provide access to the pair.

But there are difficulties with this explanation. First, since Salzberg is talking about the relative effectiveness of different cues to the stored representation of the unit, it is important that we shift our level of focus to recognition failure, rather than the relative difference between the overall levels of recall and recognition. According to Salzberg's hypothesis, recognition failure will occur when the LC, by virtue of being more concrete, provides access to the pair, while the abstract TBR does not. But recognition failure does not occur for all pairs in which the LC is concrete and the TBR abstract, and, furthermore, recognition failure also occurs when both LC and TBR are abstract, though not as frequently. Thus, as we interpret Salzberg's position, it states that for either recall or recognition to occur, a specific item (LC for recall, TBR for recognition) must provide access to the pair. Recognition failure occurs when the LC provides that access, but the TBR does not.

Rather than attempt to specify those particular stimulus attributes for which recognition failure will or will not occur, our approach is to examine those cases in which recognition failure does occur and to attempt to understand the underlying (retrieval) mechanisms. We are not content with the statement that the LC has provided access to the encoded unit while the TBR has not. Rather, our goal has been to determine more precisely what the term "provides access to" entails, in terms of underlying psychological processes, and how this access differs for recall and recognition.

We start with the assumption that the LC–TBR pair is encoded and stored as a holistic unit. This is, of course, at once a restatement and extension of what is meant by encoding specificity. Recall of one member of the pair, given the other member as a cue, requires accessing the entire unit, although we have left the exact nature of this retrieval mechanism undefined. While we speak of retrieving the TBR, given the LC as a cue, we are really referring to retrieving the entire unit. Successful retrieval of the unit produces the other member of the pair. The same retrieval mechanism is used for accessing the pair independently of which member of the pair is used as a cue. This does not mean though that retrieval is equally likely from both members of the pair.

Retrieval must be asymmetric in order for recognition failure to occur, and it is expected to be asymmetric given the nature of the instructions (study the pair so that you can recall the TBR given the LC as a cue). It also seems to vary as a function of various stimulus attributes such as concreteness (Salzberg, 1976), word frequency (Reder et al., 1974), and relatedness (Wiseman & Tulving, 1976). As far as we are concerned, the important point lies not in determining which cues will be better retrieval cues, but rather that retrieval asymmetry does occur for some pairs.

We have invoked the dual code theory of recognition to explain recognition failures. The presentation code is available for both members of the pair, as well as the pair itself; it is automatically accessed upon presentation
of the physical copy of an item. If the strength of this presentation code exceeds some predetermined criterion the item will be accepted as "old." If it falls below some lower predetermined criterion the item will be rejected. If the presentation code falls between these two criteria a second stage of processing will be undertaken in which the item is subjected to a test of accessibility. But what has been stored in the case of paired-associate tasks is the holistic unit, and the conceptual codes are used to retrieve that unit. If the unit of which the item is a part is retrieved the item will be accepted as "old." Thus, recognition can be based on either the presentation code or successful retrieval through the conceptual code. An item will not be recognized if it evokes a very weak presentation code or if the pair of which it is a part cannot be retrieved. Tatum and Ellis (1973) have also found that this type of accessibility influences recognition in paired-associate tasks.

Recognition failure, in which an item is recalled but not recognized, can occur for two classes of items: those that are not recognized on the basis of their presentation code and those for which retrieval was attempted and failed. It is the latter class of items that accounts for the majority of recognition failures. We have shown that over three-quarters of all recognition failures are due to a failure in backward retrieval, as determined by a direct test. For these pairs retrieval is asymmetric.

Items showing recognition failure on the basis of their presentation codes can be retrieved in a backward recall test (backward recognition failures), since the presentation code is presumably relatively independent of the conceptual code. Items that are not recognized but subsequently recalled in both the forward and backward recall tests occur rather infrequently, and their occurrence is even further reduced when recognition variability is taken into account.

We started off with a phenomenon that has perplexed psychologists: the phenomenon of recognition failure, which seemed at once both counterintuitive and countertheoretical. We have offered a resolution to this problem which makes it comprehensible and brings it within the bounds of a specifiable theory: the dual code recognition model. A form of this model has also recently been used by Tiberghien (1976) to account successfully for another encoding specificity phenomenon, the effects of associative context on recognition (Tulving & Thomson, 1971). Mandler (1976) has commented on the wide generality of this model, using different conceptual codes in different situations. The continued use of this model is likely to explain various memory phenomena.

REFERENCES


Mandler, G. A decennial tale--wherein the organization of memory becomes respectable and gains

In press.


(Received December 21, 1976)