

Recognition failure of categorized words

LARS-GÖRAN NILSSON
University of Umeå, Umeå, Sweden

and

MICHELLE DINNIWELL and ENDEL TULVING
University of Toronto, Toronto, Ontario, Canada

Previous research has demonstrated that the phenomenon of recognition failure of recallable words does not hold for categorized word pairs (e.g., *flower-ROSE*). We tested the hypothesis that such an exceptional finding is attributable to the use of homogeneous lists that create a situation in which the functional recall cues do not contain any information not already contained in the recognition cues. In three experiments in which categorized word pairs were used, but whose design ruled out the invariant informational overlap between recognition and recall cues, recognition failure was found to the extent expected by Tulving and Wiseman's (1975) function. These results add to the evidence that the relation between recognition and cued recall in recognition-failure experiments is largely invariant over many otherwise relevant variables that affect recognition and recall.

A large number of experiments reported in the literature have produced results that collectively seem to describe a general law of human memory, as follows: If a to-be-remembered item *T* is encoded in relation to some contextual information *C*, then recognition of *T* in the absence of *C* is largely independent of cued recall of *T* to *C* as the cue. The relevant experiments have been done under the rubric of "recognition failure," the term representing an abbreviated version of the more complete designation, "recognition failure of recallable words." In recognition-failure experiments, subjects study to-be-remembered words, each of which is presented together with a contextual word or phrase, and then they attempt to retrieve the to-be-remembered words in two successive tests.¹ The first test is one of recognition, in which copies of the to-be-remembered words appear without their study context; the second is a cued recall test, in which the intralist contexts of the to-be-remembered (or target) words are presented as retrieval cues. In this situation, subjects frequently recall words that they failed to identify as "old" words on the preceding recognition test, thus demonstrating the phenomenon of recognition failure of recallable words, or simply, recognition failure. A

review of this research was provided by Tulving (1983); other recent work includes that reported by Lieury (1979), Muter (1984), Neely and Payne (1983), Nilsson, Law, and Tulving (in press), Rabinowitz (1984), and Tajika (1978a, 1978b, 1979).

The extent of recognition failure is typically indexed by the conditional probability that a target word is not recognized given that it is recalled (Watkins & Tulving, 1975). The measured recognition failure is inflated to some extent by unsystematic variability in performance from test to test, and to some extent by "priming" of target words on the recognition test (Bowyer & Humphreys, 1979; Donnelly, in press; Flexser & Tulving, 1982; Humphreys & Bowyer, 1980). But these sources of variability account for only a small part of the observed effect.

Recognition failure is a robust phenomenon. It has been demonstrated with a variety of to-be-remembered items (e.g., Neely & Payne, 1983; Nilsson et al., in press; Tulving & Watkins, 1977), with different kinds of contextual materials (e.g., Bartling & Thompson, 1977; Park & Tulving, reported in Tulving, 1983), in experimentally naive subjects (e.g., Tulving & Thomson, 1973) as well as subjects thoroughly familiar with the phenomenon and the paradigm (Rabinowitz, Mandler, & Barsalou, 1977; Wiseman & Tulving, 1975), with both yes/no and forced-choice recognition tests (e.g., Watkins & Tulving, 1975), with different kinds of distractor items (e.g., Watkins & Tulving, 1975) or even no distractors (Begg, 1979; Wallace, 1978), with and without lure cues in the cued recall tests (e.g., Tulving & Watkins, 1977; Watkins & Tulving, 1975), over retention intervals measured in minutes (e.g., Begg, 1979; Rabinowitz, 1984) as well as days and weeks (Donnelly, in press), and in situations in which recognition is higher than recall (e.g., Postman, 1975)

This research was supported by a grant from the Swedish Council for Research in the Humanities and Social Sciences to L.-G. Nilsson and Grant A8632 from the Natural Sciences and Engineering Research Council of Canada to E. Tulving. The research reported here was completed while L.-G. Nilsson held an International Scientific Exchange Award from the Natural Sciences and Engineering Research Council of Canada at the University of Toronto.

Requests for reprints should be sent to Lars-Göran Nilsson, Department of Psychology, University of Umeå, S-90187 Umeå, Sweden, or to Endel Tulving, Department of Psychology, University of Toronto, Toronto, Ontario M5S 1A1, Canada.

as well as those in which recall is higher than recognition (e.g., Wiseman & Tulving, 1976).

Even more remarkable than the fact that recognition failure occurs under all these conditions is the systematic relation between the extent of recognition failure and the overall recognition hit rate observed in all these experiments. It is described by a quadratic function (Tulving & Wiseman, 1975) of the following form:

$$P(Rn|Rc) = P(Rn) + .5[P(Rn) - P(Rn)^2], \quad (1)$$

where $Rn|Rc$ refers to recognition conditionalized on recall—the complement of recognition failure—and Rn refers to recognition hit rate. Almost invariably, outcomes of recognition-failure experiments conform to this function, in the sense that the observed probabilities of recognition failure in different experimental conditions show relatively little deviation from the probabilities “predicted” by the function. The largely invariant relation between recognition failure and overall recognition described by Equation 1 obtains regardless of the determinants of the level of overall recognition hit rate in a given experimental condition, and regardless of the determinants of variations in the overall level of recall.

The quadratic function given by Equation 1 was fitted to the data from a number of different experiments by Tulving and Wiseman (1975). The close adherence of the data to the function subsequently observed in many additional experiments has been described and discussed by Flexser and Tulving (1978, 1982) and Tulving (1983, chap. 13). The wide generality of conditions under which Tulving and Wiseman’s (1975) empirically determined function (hereafter referred to as the Tulving-Wiseman function) has predicted the extent of recognition failure on the basis of overall recognition hit rate defines the boundary conditions of the general “law” mentioned in the opening paragraph of this article. The term “largely independent” in the statement of the law corresponds to the value of the single constant in the Tulving-Wiseman function, namely .5 (Equation 1). Complete independence between recognition and recall would be represented by a constant of zero in the function.

Given the general scarcity of laws in the science of learning and memory (Cohen, 1985) and the widely accepted inevitability of all sorts of interactions in the outcomes of learning and memory experiments (e.g., Jenkins, 1979), the appearance of any persistent regularity in experimental findings, and the possibility that such a regularity represents a law, should be met with more than passing interest by students in the particular field of inquiry. Researchers who love order in their universe should be pleased with every new manifestation of the phenomena that reveal such order, whereas those who are troubled by it should rush to discover exceptions to it.

Not surprisingly, exceptions to the Tulving-Wiseman function do exist (Tulving, 1983, p. 289). Some experiments whose designs have conformed to the recognition-failure paradigm either have found no recognition failure at all or have found considerably less recognition failure

than expected on the basis of the Tulving-Wiseman function. Since the existence of these exceptions has a clear bearing on the status of the law we are talking about, it is of considerable importance to examine the exceptions and to seek experimental and theoretical clarification of them and of their relation to the putative law. We are engaged in a research program aimed at such clarification (Nilsson & Tulving, 1986; Nilsson et al., in press). The present article describes a part of it.

The exceptions to the Tulving-Wiseman function known at this stage of research seem to fall into two categories. Findings reported by Bartling and Thompson (1977), Begg (1979), Fisher (1979), and Gardiner and Tulving (1980) form one of these categories. Exceptions in these cases seem to be attributable to insufficient association of the to-be-remembered item with its list-context. The matter has been discussed by Gardiner and Tulving (1980).

The second category of exceptions is the object of central interest in the present article. In this case, exceptions to the function have been demonstrated in experiments in which there has been no difficulty with the relational encoding of the target item and its context. Relevant data have been reported by Nilsson and Shaps (1980, 1981), based on six separate experiments. The findings were essentially replicated in two unpublished experiments done at Toronto, one by Pamela Auble, and the other by Norman Park (Tulving, 1983, p. 289). All these experiments followed the general recognition-failure paradigm, but the study materials were different from those used in all previous work: the context items of the to-be-remembered words were the names of different conceptual categories and the to-be-remembered words were the names of instances of the same categories (e.g., *flower-ROSE, four-legged animal-HORSE*). We will refer to such pairs as *categorized pairs*, or *category-INSTANCE* pairs. No or little recognition failure was found in any of Nilsson and Shaps’s (1980, 1981) experiments. Their results, therefore, seem to violate the law under scrutiny. If left unchallenged, these results alone would be sufficient to reject the law.

A Tentative Explanation of the Exception

The purpose of the present article is to describe three related experiments that were designed to test a hypothetical explanation of Nilsson and Shaps’s (1980, 1981) findings of exceptions with categorized pairs. The hypothesis was that Nilsson and Shaps’s results did not violate the law as stated in the opening paragraph of this article, because their experimental conditions fell outside the boundaries of the domain within which the law holds.

Our hypothetical explanation of Nilsson and Shaps’s (1980, 1981) findings can be stated at two levels, empirical and theoretical. The empirical hypothesis holds that the deviations from the Tulving-Wiseman function of the sort obtained by Nilsson and Shaps (1980, 1981) with *category-INSTANCE* pairs are attributable to the use of *homogeneous* lists in which all pairs of study materials are of the same kind. This hypothesis is tested in our ex-

periments by measuring the extent to which recognition failure occurs with the *category*-*INSTANCE* pairs presented for study in *nonhomogeneous* lists in which only some of the pairs are of this kind. Our theoretical hypothesis holds that the use of homogeneous lists effectively converts the nominal noncued recognition test into a functional cued recognition test (cf. Underwood, 1963). Since recognition failure does not occur under the conditions of cued recognition (Wallace, 1978), it does not occur with homogeneous *category*-*INSTANCE* lists.

Our theoretical hypothesis is implicit in several different theoretical accounts of recognition failure. All entail the same basic idea that recognition failure is absent when there is a great deal of informational overlap between recognition and recall cues, or when the information in recall cues is included in the (functional) recognition cues. Let us briefly consider five such accounts.

The first explanation is the general model of recognition failure proposed by Flexser and Tulving (1978). This model incorporates a parameter, k , that specifies the relation between component features "useful for retrieval" in the recognition cue and in the recall cue. When these features in one cue are uncorrelated with those in the other—a situation described by Flexser and Tulving's (1978) special model—the value of k is near zero and recognition failure conforming to the Tulving-Wiseman function occurs. At the other extreme, when all useful-for-retrieval features of the recall cue are included among the useful-for-retrieval features of the recognition cue, the value of k is 1. In this latter case, all recallable items are necessarily recognized, provided that the recognition criterion is not higher than the recall criterion (cf. Kintsch, 1978). Thus, according to Flexser and Tulving's (1978) model, Nilsson and Shaps's (1981) findings could reflect a situation in which no useful retrieval information is available to the subjects at recall that is not available at recognition.

Second, the explanation that Nilsson and Shaps (1981) offered for their findings assumed that, after studying a list consisting of categorized pairs, the subjects would be well aware of the structure of the list. Guided by this knowledge, the subjects would adopt the strategy at the time of the recognition test of reconstructing the category name associated with each test item and responding on the basis of this more comprehensive functional cue. Thus, after having seen a pair such as *flower*-*ROSE* in the study list, the subject would implicitly convert the nominal recognition cue *rose* into the functional cue of *flower-rose* and make the recognition judgment about it. As a consequence of this strategy, when the category name is presented as a retrieval cue in the subsequent cued recall test, it constitutes only a part of the total retrieval information that the subject has already used in the immediately preceding recognition test. Since no grounds exist for expecting that a partial cue of this kind is more effective than a whole cue, subjects would not recall words other than those they have already recognized. Thus, according to Nilsson and Shaps (1981), no recognition

failure can be expected when subjects adopt a reconstruction strategy that maximizes the functional cue overlap between the recognition and recall tests. Conversely, when no such reconstruction is possible, recognition failure would be expected to occur (Nilsson & Shaps, 1981).

A third explanation of exceptions to the function with homogenized categorical materials is implicit in the idea that recognition failure is a consequence of associative asymmetry between the A and B members (contextual and to-be-remembered components) of study pairs: recognition failure of the B member of the study pair occurs when the subject cannot implicitly retrieve A when B is given as the cue (Bartling & Thompson, 1977; Rabinowitz et al., 1977; Salzberg, 1976; Tulving & Thomson, 1973). If the subject can retrieve A given B, all recallable words are recognized. Therefore, in an experiment such as that of Nilsson and Shaps (1981), in which subjects can and probably do implicitly generate the category name when faced with a category instance in the recognition test, no recognition failure should occur.

A fourth explanation is implicit in Jones's (1978, 1983) account of recognition failure. Jones's argument was that cued recall is mediated by either or both of two mechanisms, "direct access" and "schema based." Recognition failure of recallable words can occur only if the cue in the recall test provides direct access to the encoded and stored representation of the target. It cannot occur if recall is "schema based" since both recall and recognition require the same intact schema information. In Jones's (1983) conceptualization, therefore, Nilsson and Shaps's (1981) experiments represent a situation in which recall is entirely schema based. Since each target word was initially encoded in relation to previously acquired knowledge—the category to which the word belongs—this schema served to mediate retrieval at the time of recall (and recognition).

Finally, Wallace (1978) reported three experiments in which he explored the relation between memory performances in several kinds of successive tests. One of the comparisons made was that between the standard recognition-failure sequence of noncued recognition followed by cued recall tests and cued recognition followed by cued recall. Wallace found less recognition failure in the latter than in the former condition and argued that recognition failure is simply a special case of much more general context effects. Recognition failure occurs if the contextual cues presented to the subjects in the recall test are "different and more appropriate" than those presented in the recognition test. In Wallace's formulation, the exception to the function with category pairs occurs because the study-list cues are either implicitly or explicitly present in both the recognition and the recall tests.

Despite the differences in terminology, all these theoretical accounts of recognition failure converge on the idea that recognition failure occurs when the recall cues contain relevant retrieval information not contained in recognition cues, and that it does not occur when recall cues do not contain any additional relevant information. We

hypothesize that the latter situation came about in Nilsson and Shaps's (1980, 1981) experiments because of the homogeneous nature of the lists that were used. Such lists made it possible for the subjects to convert the nominally noncued recognition test into a functional cued recognition test, rendering the subsequently presented recall cue redundant with the recognition cue.

Furthermore, according to most of these theoretical accounts, it is not the type of material used by Nilsson and Shaps (1981) that is directly responsible for the exception, but rather a particular strategic component of the processing of the material that the use of homogeneous lists allowed. If this idea is correct, then it follows that recognition-failure results conforming to the Tulving-Wiseman function should be found with categorized study-list pairs under conditions in which the reliance on this strategic component is precluded.

Such conditions can be created through the use of heterogeneous lists in which only some of the study pairs consist of category names and appropriate instances. To the extent that the use of such lists precludes the invariant informational overlap between recognition and recall cues, recognition failure of categorical materials should approximate the levels found with other kinds of materials studied in heterogeneous lists.

We now describe three experiments in which we examined the extent to which, if any, *category-INSTANCE* pairs would exhibit recognition failure when these pairs appear in study lists containing other kinds of pairs as well. We compare our results with those reported for homogeneous lists by Nilsson and Shaps (1980, 1981). We did not include any homogeneous lists in these new experiments, because we consider Nilsson and Shaps's findings, based on a large number of observations, to be firm and reliable. There were altogether 11 different experimental conditions in Nilsson and Shaps's experiments, incorporating a total of 4,792 observations. With the overall recognition hit rate averaging .86 in the 11 conditions, the expected number of subject-items recalled but not recognized in these 11 conditions would have been 382, according to the Tulving-Wiseman function. In fact, the observed number was very much smaller, namely 81. Our aim was to see whether a "normal" amount of recognition failure for *category-INSTANCE* pairs could be obtained with nonhomogeneous lists.

METHOD

The three experiments reported here were designed with the stated considerations in mind. The critical conditions in the three experiments were represented by *category-INSTANCE* pairs of the same kind that had been tested by Nilsson and Shaps (1980, 1981). In these conditions, the to-be-remembered pairs consisted of a category name as the cue (e.g., *flower*) and a category instance as the target item (ROSE or POPPY). These critical pairs formed one third of the experimental study pairs. Another third of the pairs in the study list were *instance-INSTANCE* pairs, composed of one category instance as the cue and another instance from the same category as the target (e.g., *tulip-ROSE*). The remaining one third of the

study-list pairs were *associate-INSTANCE* pairs. The cues in these pairs consisted of noncategorical semantic associates of the target words, and targets consisted of category instances comparable to those in the rest of the list. Thus, all target words in the study list were instances of categories, whereas the cue words, and their relation to target words, differed for three types of pairs. This arrangement made it impossible for the subjects to determine the type of cue word associated with a target word simply on the basis of the type of target word.

In each of these three within-list conditions, half of the target words were high-frequency instances of the category (e.g., *flower-ROSE*) and the other half were low-frequency instances (e.g., *flower-POPPY*). The purpose of this classification was simply to imitate Nilsson and Shaps's (1980, 1981) experiments, in which the frequency of category instances in the norms (high vs. low) were included in the design and the analysis of the results.

Experiments 1 and 2 were conducted at the University of Umeå in Sweden, and Experiment 3 was conducted at the same time at the University of Toronto in Canada. The three experiments were all alike in some respects, particularly in the use of the study lists as briefly outlined, but they differed in some minor methodological details. The reason for doing three experiments in two laboratories in different countries, with subjects speaking different languages, was simply to assure the replicability of the results across minor variations in methodological details.

General Design, Materials, and Procedure

In the study lists of all three experiments, there were 36 critical target words appearing as right-hand members of 36 word pairs. Each of the six experimental conditions was represented in the list by six different word pairs. Three types of left-hand members (category names, category instances, and associates) were orthogonally combined with high- or low-frequency targets to form the six types of word pairs. High-frequency targets within each category had frequencies of 38% or more, according to the norms of Battig and Montague (1969) and Nilsson (1973), whereas low-frequency targets had frequencies of 16% or less. Table 1 presents a schematic summary of the list structure. Left-hand members of study pairs are typed in lowercase letters, target words in uppercase letters, and retrieval cues in italicized lowercase letters.

The design of each of the three experiments corresponded to the structure of the study list. Thus the orthogonal combinations of three types of study contexts and two levels of normative frequency of targets constitute a 2×3 within-subjects design. It should be noted, however, that neither of the two independent variables taken separately, nor their interaction, was of primary concern here.

The critical conditions of primary interest in all three experiments were the two topmost conditions listed in Table 1, in which the subjects studied *category-INSTANCE* pairs and in which they were tested with the same (intralist) *category* cues in recall. The other four conditions in the design served two purposes: (1) the presence of other types of study-list pairs created the nonhomogeneous list dictated by the logic of the design, and (2) the recognition-failure data from them, expected to conform to the Tulving-Wiseman function, provided a convenient set for comparison with the recognition-failure data from the two critical conditions.

The following five successive steps in the procedure were common to all three experiments: (1) The subjects were instructed that they were to study the word pairs of the to-be-presented list in such a way that they would be able to recall the second member of each pair when the first member was presented as a retrieval cue. (2) The study list of word pairs was presented. (3) An interpolated task was given. (4) The yes/no recognition test was given that included old test items corresponding to the target words from critical study pairs. (5) The subjects were given retrieval cues for the recall of the target words from the critical study pairs. The subjects in each of the three experiments were tested individually.

Table 1
Basic Structure of Study List and Cue Information Given in
Recognition and Recall Tests in Three Experiments,
with a Sample of Materials

Context at Study	Target and Frequency	Copy Cue in Recognition Test	Contextual Cue in Recall Test
Category Name: Body part	High Instance LEG	LEG	<i>Body part</i>
Category Name: Body part	Low Instance LIP	LIP	<i>Body part</i>
Category Instance: Arm	High Instance LEG	LEG	<i>Body part*</i> <i>Arm†</i>
Category Instance: Arm	Low Instance LIP	LIP	<i>Body part*</i> <i>Arm†</i>
Associate: Long	High Instance LEG	LEG	<i>Body part*</i> <i>Long†</i>
Associate: Long	Low Instance LIP	LIP	<i>Body part*</i> <i>Long†</i>

*Experiments 1 and 3. †Experiment 2.

Experiments 1 and 2

Experiments 1 and 2 differed from each other in only one important respect: In Experiment 1, cues in the recall test for all targets consisted of category names of the target words, whereas in Experiment 2 recall cues were left-hand members of the study-list pairs. Thus, recall was cued with extralist cues (cues not explicitly present at study) in four conditions in Experiment 1, with intralist cues in two conditions in Experiment 1, and with intralist cues in all six conditions in Experiment 2. The purpose of this difference in design between the two experiments was to explore the relevance of the homogeneity of recall cues to the results of the experiments.

Thirty subjects, undergraduate students at the University of Umeå, participated in Experiment 1, and 24 subjects from the same source took part in Experiment 2. In all other respects, the method used was identical in these two experiments.

The study list consisted of 36 critical word pairs, 2 filler pairs at the beginning, and 2 at the end of the list. Target words were selected from Swedish norms (Nilsson, 1973). To balance words across the experimental conditions, three versions of the study list were constructed. Across different versions, a given critical target word appeared once with its category name, once with another category instance, and once with a semantic associate. High- and low-frequency target words of a given category were accompanied, in different versions of the list, by the same instance of the category and the same semantic associate. Thus, with the data pooled over the three lists, any differences in performance in different conditions could be attributed to the relation between the study context of the target word and the retrieval cue, and not to differences among target words as such.

The list was presented on a computer screen for study on a single trial. Pairs appeared in succession at a rate of 2 sec/pair with a 1-sec blank interval between pairs. The 36 critical pairs appeared in a different random order for each subject. The interpolated task, given immediately after study, consisted of one verbal and one non-verbal test taken from an intelligence test battery and lasted for 20 min. In the recognition test, the 36 targets from the study list were presented in a new randomized order, which was again different for each subject. No lures were contained in the recognition test. Previous research (Wallace & Page, 1982) has shown that overall recognition hit rates are little affected by the presentation of distractors. The subjects were instructed to identify the words on the recognition test that they remembered having seen in the study list. They were not told that only the targets (and no lures) would appear in the recognition test. The subjects indicated their recognition choices by pressing a key on the keyboard of the computer terminal. By pressing another key they indicated that they did not

recognize a given test word. Each test word disappeared from the screen as soon as the subject responded. One second later the next word appeared on the screen. In the cued recall test that followed, 36 extralist or intralist cues as described earlier were presented on the computer screen, in a different random order for each subject. Subjects were instructed to try to recall the appropriate target word as each cue appeared. They responded by typing the target words on the keyboard of the computer terminal. A given cue disappeared from the screen as soon as the subject had completed the typing of a word or indicated that he/she could not respond. One second later the next cue appeared on the screen.

Experiment 3

The procedure of Experiment 3, conducted as an undergraduate research project at Toronto, differed from that of Experiments 1 and 2 in that it was less mechanized. In Experiment 3, study and a cued recall test of two practice lists preceded the study of the main experimental list. This main list contained 36 critical word pairs, composed of six subsets as shown in Table 1, together with 34 filler pairs for half of the subjects and 80 filler pairs for the remaining half. (The number of filler pairs was increased halfway through the experiment in an attempt to reduce recognition performance. The scores did not, however, differ greatly between the two sets of subjects, and the data from all subjects were, therefore, treated as coming from one single group.) The order of the critical word pairs was random. Half the filler items appeared at the beginning of the study list and half at the end. The words of the critical pairs and the 34 filler items were selected from the Battig and Montague (1969) norms. The remaining 46 buffers for the second half of the subjects were obtained from the lists used by Tulving and Thomson (1973) and Wiseman and Tulving (1975), or were generated specially for the experiment. To balance words across the experimental conditions, we constructed six versions of the experimental list. Across three different versions, a given critical target word appeared once with its category name, once with another category instance, and once with a semantic associate. High- and low-frequency target words of a given category were paired with different instances and different semantic associates.

The word pairs of the study list were presented by means of a "window mask" (i.e., a long piece of cardboard with a small rectangular opening showing one pair at a time). At study, the subject, guided by auditory signals, moved the mask down the list at a rate of 3 sec/pair.

After the subjects had seen the main study list, there were two brief interpolated tasks. In one subjects were asked to rate cartoons for "funny-ness," and in the other they were asked to generate one

Table 2
Mean Proportions of Responses in the Four Cells of a 2x2 Classification of Recognition and Recall Successes and Failures in Six Conditions in Each of the Three Experiments

Encoding Condition		Rn,Rc			Rn,Rc			Rn,Rc			Rn,Rc		
Type of Cue	Frequency of Target	Experiment			Experiment			Experiment			Experiment		
		1	2	3	1	2	3	1	2	3	1	2	3
Category Name	High	.56	.72	.57	.21	.08	.14	.12	.12	.12	.12	.09	.17
Category Name	Low	.48	.58	.49	.32	.24	.33	.07	.08	.02	.13	.10	.16
Category Instance	High	.47	.60	.40	.24	.10	.30	.16	.14	.09	.13	.15	.21
Category Instance	Low	.38	.56	.32	.41	.25	.43	.06	.03	.03	.16	.16	.22
Associate	High	.50	.52	.35	.23	.19	.31	.09	.13	.09	.18	.16	.25
Associate	Low	.37	.45	.22	.40	.28	.51	.07	.07	.03	.16	.19	.24
Mean		.46	.57	.39	.30	.19	.34	.10	.10	.06	.15	.14	.21

Note—Failure to recognize or recall is indicated by a bar, Rn and Rc, respectively.

free associate to each of 24 words taken from the practice lists. These interpolated tasks lasted approximately 15 min. In the recognition test that followed next, the 36 critical targets from the study list were randomly placed among 72 lures that consisted of one high-frequency and one low-frequency instance of each of the 36 categories corresponding to the study-list targets. The recognition test items were presented on a sheet of paper, and the subjects were instructed to make an "old" or a "new" judgment about each. In the cued recall test, 36 category names were presented as cues on a sheet in a randomly determined order; the subjects were asked to write down the targets presented in the study list.

Twenty-four subjects, undergraduate students at the University of Toronto, participated in the experiment. Subgroups of 4 subjects studied different versions of the study list. All subjects took the same recognition test and the same cued recall test.

RESULTS AND DISCUSSION

We begin by presenting the basic summary data from all three experiments in Table 2, consisting of weighted mean proportions of target words recognized or not recognized, and recalled or not recalled, conforming to the 2x2 contingency tables typically used in the analyses of recognition-failure data (Tulving & Wiseman, 1975; Watkins & Tulving, 1975). These data give a general idea about the comparability of the results from the three experiments, but are not of much interest otherwise.²

The theoretically interesting data have to do with the extent to which observed probabilities of recognition failure in the critical conditions, as well as in the other conditions in our experiments, deviate from the Tulving-Wiseman function. The relevant conditional probabilities, as well as the probabilities of overall recognition and recall, can be calculated from the data presented in Table 2. When these calculations are performed, four observations relevant to the objectives of the experiments emerge: (1) the three experiments yielded roughly similar results, (2) a modest amount of recognition failure occurred in all 18 experimental conditions of the three experiments, (3) the observed probabilities of recognition failure show reasonably good agreement with the predicted probabilities, and (4) the differences between the observed and the predicted probabilities of recognition failure do not seem to be systematically related to either the type of study pair or the category frequency of the target word.

Figure 1 depicts the proportions of recallable words recognized (the complement of recognition failure) as a function of the overall recognition hit rate, together with the comparable data reported by Nilsson and Shaps (1980, 1981), and shows the extent to which both sets of the data conform, or do not conform, to the Tulving-Wiseman function, indicated by the curved line in the graph. Filled circles at the upper edge of the graph represent data from 11 conditions in the experiments of Nilsson and Shaps (1980, 1981), triangles represent the data for the critical category-INSTANCE pairs from our three experiments, and open circles represent the data for the other pairs.

The data in Figure 1 tell a simple and straightforward story: When category-INSTANCE pairs appear in homogeneous lists, there is very little recognition failure of target words from these pairs, and whatever recognition failure does occur deviates considerably from the levels predicted by the Tulving-Wiseman function. When the same pairs appear in nonhomogeneous lists, the target words from these pairs show about the same amount of recognition failure as has been shown by numerous previous experiments in which a variety of different materials

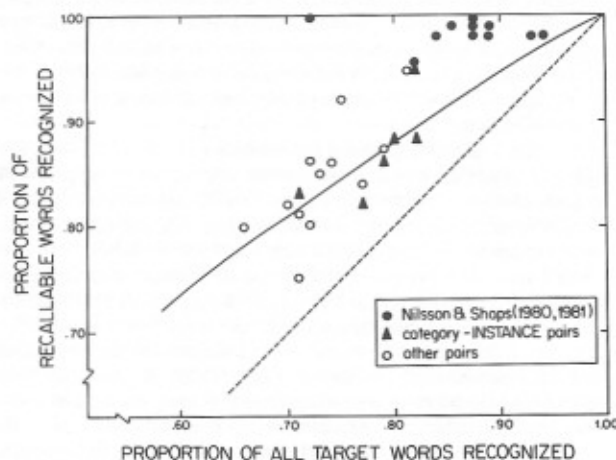


Figure 1. Proportion of recallable words recognized as a function of the proportion of recognition of all target words for the 11 conditions in six experiments reported by Nilsson and Shaps (1980, 1981) and the 18 conditions of the present three experiments.

have been used. The Tulving-Wiseman function summarizes the aggregate of the data from all the previous experiments. The relatively small and unsystematic deviation of the observed data from those predicted by that function tells us that the amount of recognition failure observed for the target words from the critical *category-INSTANCE* pairs was as "normal" as it was for the target words from our other kinds of pairs; that is, our data are not distinguishable from the data produced in a large number of previous recognition-failure experiments.

The correspondence between the present data and the data from previous experiments (Tulving, 1983, Figure 13.1) can be judged by inspection of the frequency distributions of deviations from the Tulving-Wiseman function for the two sets of experiments, as presented in Figure 2. Deviations from the function were calculated by subtracting the observed probability of recognition failure from that predicted. Thus, positive deviations in this graph indicate less recognition failure than expected, and negative values indicate more recognition failure than expected. Figure 2 shows that the two distributions are quite similar.

The results of the three experiments thus provide support for our empirical hypothesis: Recognition failure for categorized pairs does not occur when these pairs are presented for study in homogeneous lists, but does occur to the normal extent when they are presented in non-homogeneous lists. The test of the empirical hypothesis relied on a comparison between the results of our three experiments and the results of Nilsson and Shaps's (1980, 1981) two experiments, but was otherwise straightforward.

In designing the experiments, we were guided by what we referred to as the theoretical hypothesis, which was that homogeneous lists create a situation in which the nominal noncued recognition test is converted into a functional cued recognition test (cf. Underwood, 1963; Wallace, 1978). Inasmuch as recognition failure would not be expected to occur in a cued recognition test, but would be expected to occur in a noncued recognition test, for reasons discussed directly or indirectly by a number of theorists (e.g., Bartling & Thompson, 1977; Flexser &

Tulving, 1978, 1982; Gillund & Shiffrin, 1984; Jones, 1978, 1983; Nilsson & Shaps, 1980, 1981; Rabinowitz et al., 1977; Salzberg, 1976; Tulving & Thomson, 1973; Wallace, 1978), we expected that in our experiments the *category-INSTANCE* pairs would show normal amounts of recognition failure. This is what we found.

The implications of our findings for the theoretical hypothesis, like the implications of all experimental findings for all such hypotheses, are less compelling than the implications of the findings for the empirical hypothesis. Although the results of the experiments are in accord with the hypothesis, and hence can be thought to support it, they do not rule out alternative theoretical explanations that may also apply to the differences in recognition failure observed with homogeneous and heterogeneous lists. It is possible that the processes of encoding, or retrieval, or both, engendered by homogeneous and heterogeneous lists, differ in ways other than those that we have tried to capture and explicate under the rubric of "functional" retrieval cues. Until such time, however, that a rival interpretation is put forth and tested, we have some reason to believe in the tenability of our hypothesis.

In terms of our speculative account of the reasons for the exceptions to the law shown by Nilsson and Shaps's (1980, 1981) findings, we suggest that those findings did not conform to the law because the conditions under which they were obtained fell outside the boundaries of the domain within which the law holds. When homogeneous *category-INSTANCE* lists are used that allow the subjects to adopt the reconstructive strategy, recognition of the target item *T* does not occur in the absence of the contextual information *C*. The law states the conditions under which the relation between recognition and cued recall manifests itself in a particular manner. If these conditions are not fulfilled, as we argue they were not in Nilsson and Shaps's (1980, 1981) experiments, the predicted relation need not be, and in Nilsson and Shaps's experiments was not, observed.

Finally, the explanation we have offered for Nilsson and Shaps's (1980, 1981) exceptional results may also hold for similar results reported by Muter (1984). Muter tested his subjects' ability to recognize and subsequently to recall unique names of well-known people (such as Kierkegaard and Atatürk) in a semantic-memory analogue of the typical episodic recognition-failure experiment. He found no recognition failure: the subjects almost invariably recognized the names they could recall. Muter's exception to the Tulving-Wiseman function cannot be attributed to insufficient association of cue and target items, but it is possible that it resulted from a similar overlap between retrieval information in the two tasks that we have dealt with in the present article. We have discussed Muter's "exception" at greater length elsewhere (Nilsson et al., in press).

The data we have presented in this article eliminate one reason for doubting the generality of the law of human memory with which we started our article. Although a few other sources of misgivings may still exist, it seems reasonable to expect that they, too, will fall by the way-

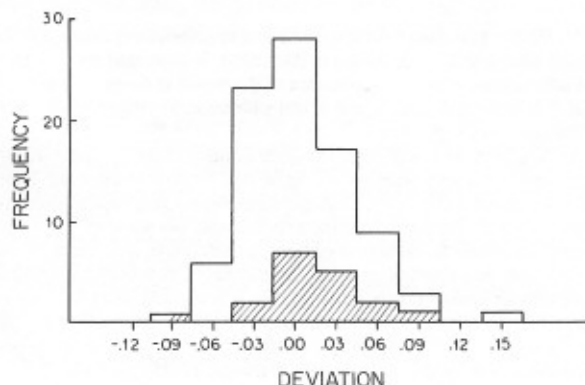


Figure 2. Frequency distributions of deviations from Tulving and Wiseman's (1975) function for the present experiments (shaded area) and the experiments summarized by Tulving (1983, Figure 13.1).

side as we gain better empirical and theoretical understanding of the remaining demonstrated exceptions to the law of relative independence of retrieval processes in recognition and cued recall.

REFERENCES

- BARTLING, C. A., & THOMPSON, C. P. (1977). Encoding specificity: Retrieval asymmetry in the recognition failure paradigm. *Journal of Experimental Psychology: Human Learning & Memory*, 3, 690-700.
- BATTIG, N. F., & MONTAGUE, W. E. (1969). Category norms for verbal items. *Journal of Experimental Psychology Monograph*, 80(3), 1-46.
- BEGG, I. (1979). Trace loss and the recognition failure of unrecalled words. *Memory & Cognition*, 7, 113-123.
- BOWYER, P. A., & HUMPHREYS, M. S. (1979). Effect of a recognition test on a subsequent cued-recall test. *Journal of Experimental Psychology: Human Learning & Memory*, 5, 348-359.
- COHEN, R. L. (1985). On the generality of the laws of memory. In L.-G. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 247-277). Hillsdale, NJ: Erlbaum.
- DONNELLY, R. (in press). Priming effects in successive episodic tasks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.
- FISHER, R. P. (1979). Retrieval operations in cued recall and recognition. *Memory & Cognition*, 7, 224-231.
- FLEXNER, A. J., & TULVING, E. (1978). Retrieval independence in recognition and recall. *Psychological Review*, 85, 153-171.
- FLEXNER, A. J., & TULVING, E. (1982). Priming and recognition failure. *Journal of Verbal Learning & Verbal Behavior*, 21, 237-248.
- GARDINER, J. M., & TULVING, E. (1980). Exceptions to recognition failure of recallable words. *Journal of Verbal Learning & Verbal Behavior*, 19, 194-209.
- GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1-67.
- HUMPHREYS, M. S., & BOWYER, P. A. (1980). Sequential testing effects and the relationship between recognition and recognition failure. *Memory & Cognition*, 8, 271-277.
- JENKINS, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing and human memory* (pp. 429-446). Hillsdale, NJ: Erlbaum.
- JONES, G. V. (1978). Recognition failure and dual mechanisms in recall. *Psychological Review*, 85, 464-469.
- JONES, G. V. (1983). Structure of the recall process. *Philosophical Transactions of the Royal Society 302B* (pp. 373-385). Reprinted in D. E. Broadbent (Ed.), *Functional aspects of human memory*. (pp. 135-147). London: Royal Society.
- KINTSCH, W. (1978). More on recognition failure of recallable words: Implications for generation-recognition models. *Psychological Review*, 85, 470-473.
- LIEURY, A. (1979). La memoire episodique est-elle emboitee dans la memoire semantique? [Is episodic memory embedded in semantic memory?] *L'Année Psychologique*, 79, 123-142.
- MUTER, P. (1984). Recognition and recall of words with a single meaning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 10, 198-202.
- NEELY, J. H., & PAYNE, D. G. (1983). A direct comparison of recognition failure rates for recallable names in episodic and semantic memory tests. *Memory & Cognition*, 11, 161-171.
- NILSSON, L.-G. (1973). *Category norms for verbal material* (Report No. 135). Uppsala, Sweden: University of Uppsala, Department of Psychology.
- NILSSON, L.-G., LAW, J., & TULVING, E. (in press). Recognition failure of recallable unique names. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.
- NILSSON, L.-G., & SHAPS, L. P. (1980). A functional view of memory. In F. Klix & J. Hoffman (Eds.), *Cognition and memory: Interdisciplinary research of human memory activities* (pp. 40-46). Berlin: Deutscher Verlag der Wissenschaften.
- NILSSON, L.-G., & SHAPS, L. P. (1981). A reconstructive processing interpretation of the recognition failure phenomenon. *Acta Psychologica*, 47, 25-37.
- NILSSON, L.-G., & TULVING, E. (1986, November). *Exceptions to recognition failure*. Paper presented at the Psychonomic Society meeting, New Orleans.
- POSTMAN, L. (1975). Tests of the generality of the principle of encoding specificity. *Memory & Cognition*, 3, 663-672.
- RABINOWITZ, J. C. (1984). Aging and recognition failure. *Journal of Gerontology*, 39, 65-71.
- RABINOWITZ, J. C., MANDLER, G., & BARSALOU, L. W. (1977). Recognition failure: Another case of retrieval failure. *Journal of Verbal Learning & Verbal Behavior*, 16, 639-663.
- SALZBERG, P. M. (1976). On the generality of encoding specificity. *Journal of Experimental Psychology*, 2, 586-596.
- TAJIKI, H. (1978a). Features of recognition tasks in encoding specificity: The function of context in recognition tasks. *Japanese Psychological Research*, 20, 93-100.
- TAJIKI, H. (1978b). Features of recognition tasks in encoding specificity: The function of free associate tasks and non-verbal tasks. *The Japanese Journal of Psychology*, 48, 344-347.
- TAJIKI, H. (1979). Memory processes in recall and recognition. *Psychologia*, 33, 146-154.
- TULVING, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.
- TULVING, E., & THOMSON, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-373.
- TULVING, E., & WATKINS, O. C. (1977). Recognition failure of words with a single meaning. *Memory & Cognition*, 5, 513-522.
- TULVING, E., & WISEMAN, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, 6, 79-82.
- UNDERWOOD, B. J. (1963). Stimulus selection in verbal learning. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning* (pp. 33-48). New York: McGraw-Hill.
- WALLACE, W. P. (1978). Recognition failure of recallable words and recognizable words. *Journal of Experimental Psychology: Human Learning & Memory*, 4, 441-452.
- WALLACE, W. P., & PAGE, K. A. (1982). Recognition test trials without distractors: A comparison of test trials and study trials in recognition and recall. *Bulletin of the Psychonomic Society*, 20, 245-248.
- WATKINS, M. J., & TULVING, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, 104, 5-29.
- WISEMAN, S., & TULVING, E. (1975). A test of confusion theory of encoding specificity. *Journal of Verbal Learning & Verbal Behavior*, 14, 370-381.
- WISEMAN, S., & TULVING, E. (1976). Encoding specificity: Relation between recall superiority and recognition failure. *Journal of Experimental Psychology: Human Learning & Memory*, 2, 349-361.

NOTES

1. This expression, "retrieval of the to-be-remembered words," like many other similar expressions in this article, is shorthand for "retrieval of information about the appearance of the words in the particular study list." In this article we are concerned with remembering of word-events (Tulving, 1983, p. 146).

2. The proportions of cases in which subjects failed to recognize target words that they managed to recall can be seen in the third group of columns of Table 2. These proportions show some variability across experimental conditions in three experiments, but some of this variability is attributable, on logical grounds, to differences in levels of overall recall and, on empirical grounds, to differences in overall recognition. The former source of variability is taken into account when recognition failure is expressed in relation to recall, as shown in Equation 1 in the introduction. The latter source, differences in recognition, is largely accounted for by the Tulving and Wiseman (1975) function.

(Manuscript received May 5, 1986;

revision accepted for publication December 22, 1986.)