JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 19, 194–209 (1980)

Exceptions to Recognition Failure of Recallable Words

JOHN M. GARDINER

The City University, London

AND

ENDEL TULVING

University of Toronto

Results of many experiments that conform with the recognition failure paradigm have yielded a systematic relation between the proportion of all words that are recognized and the proportion of recalled words that are recognized. In two experiments described here, the proportion of recalled words recognized was found to be much higher than expected on the basis of this relation when lists composed of pairs of abstract words or digit–word pairs were presented under typical study conditions of the paradigm. Under special study conditions, however, it was found that the proportion of recalled words that are recognized approximated more closely to expected values. Thus, exceptions to the general pattern of results were obtained and, more important, shown to depend primarily on encoding operations rather than on properties of materials as such. Theoretical explanations of recognition failure were evaluated with respect to their ability to account for these exceptions, and it was concluded that no theory can adequately explain them.

In certain circumstances, it can be shown that people can recall a previously studied word which they cannot recognize. This recognition failure phenomenon (recognition failure of recallable words) was first described in detail by Tulving and Thomson (1973). They reported three experiments in which subjects studied a series of word pairs (A–B) in the expectation that they would be tested for the recall of each B item given the A item as a cue. Before the expected recall test, however, subjects were given a recognition test for B items: literal copies of the B items were represented in the absence of the A items, along with a large number of lure items. After the recognition test, all the A items were re-presented in the cued recall test that subjects had been led to expect. Recognition failure of recallable words occurred, in the sense that subjects frequently recalled B items which they had not recognized in the preceding test.

Recognition failure of recallable words has been observed now in many experiments which, though varying in a number of procedural details, have all conformed with the basic experimental paradigm just described. Flexser and Tulving (1978), for instance, summarized recognition failure data from 33 different experiments and since the publication of their paper other experiments have appeared that could be added to the list (e.g., Begg, 1979; Tajika, 1978; 1979; Wallace, 1978). The available data show large variations in overall recall and recognition rates across experiments (Flexser and Tulving, 1978, Fig. 2), yet recognition failure has been observed in every single experiment. The incidence of recognition failure may be expressed directly as the conditional probability that B members of studied A–B pairs are not recognized given that they are recalled (Watkins & Tulving, 1975). For certain purposes it is more convenient to deal with the complement of this measure, namely, the conditional probabil-

This research was carried out while the first author was on leave at the University of Toronto, and supported by the National Research Council of Canada Grant A8632 to Endel Tulving. We thank Arthur J. Flexser and Gregory V. Jones for useful discussion. Reprint requests may be sent to John M. Gardiner, The Psychology Unit, Department of Social Science and Humanities, The City University, Northampton Square, London EC1V OHB, England.

0022-5371/80/020194-16$02.00/0

ity that B items are recognized given that they are recalled. Adopting the latter expression, Tulving & Wiseman (1975) plotted incidence of recognition failure against the overall level of recognition across a number of experimental conditions from a number of experiments, and they found that the scatter plot yielded a highly systematic relation. This relation between recognition of recallable words and overall recognition turned out to be quite well described by the following equation:

$$p(Rn/Rc) = p(Rn) + c[p(Rn) - p(Rn)^2] \quad [1]$$

where the single constant c in the function best fitting the data was found to be 0.5. Flexser and Tulving (1978, Fig. 1) presented an updated version of this function based on data from 89 experimental conditions in 33 different experiments. No large or systematic deviations of data points from the function described by Equation [1] were obtained. The systematic relation between recognition and recognition given recall thus appears to be largely invariant with experimental conditions that produce large variations in overall recognition and recall in the set of experiments surveyed.

This paper is concerned with exceptions to the phenomenon of recognition failure and, in particular, with exceptions relating to the nature of the stimulus material. Exceptions to recognition failure may be defined in two rather different ways. The first type of exception does not take into account the systematic relation between the incidence of recognition failure and the overall level of recognition: it is defined simply by the absolute magnitude of the effect. Thus, exceptions here refer to cases where the conditional probability that B members of studied A–B pairs are not recognized given that they are recalled turns out to be rather small, or even nonexistent. For example, several commentators have suggested or implied that the phenomenon may be restricted to certain kinds of materials (e.g., Bahrick, 1979; Light, Kimble, & Pellegrino, 1975; Martin, 1975; Murdock, 1976). And results of at least two studies (Reder, Anderson, & Bjork, 1974; Salzberg, 1976) have been widely interpreted as limiting the generality of the phenomenon in this way. Such claims have been based on the observation that in the recognition failure paradigm recall is sometimes higher than recognition, and sometimes recognition is higher than recall (see too, e.g., Postman, 1975). Tulving and Watkins (1977) showed, however, that when plotted in the manner described by Tulving and Wiseman (1975), the data obtained by Reder et al. (1974) were indistinguishable from results of other experiments demonstrating recognition failure. Similar analyses of data reported by Salzberg (1976) led to the same conclusion: although different kinds of materials used in experiments did produce differences in the incidence of recognition failure, these differences were well predicted by overall levels of recognition, and the incidence of recognition failure did not deviate significantly from the Tulving and Wiseman function. Given the apparent empirical generality of the recognition failure data as described by the Tulving and Wiseman function, it is clear that small amounts of recognition failure could be readily obtained simply by doing experiments in which the overall level of recognition is high. Furthermore, given the apparent invariance of the relation between recognition failure and overall level of recognition, it would not matter exactly how these high levels of recognition were achieved: the amount of recognition failure would be expected to be small in every case.

Hence the exceptions to recognition failure with which we are concerned are not defined by the overall conditional probability that recognition fails given that recall succeeds. Rather, by exceptions to recognition failure we mean deviations from the Tulving and Wiseman (1975) function. Moreover, since some natural variation of data points around the function occurs (see, e.g., Flexser & Tulving, 1978), we must refine this definition of exceptions to recognition failure further by stipulating that it entails relatively large deviations from the function.

There have been some scattered reports suggesting that the use of certain kinds of materials may, indeed, produce data points that deviate considerably from the Tulving and Wiseman function and which in this sense can be thought to constitute exceptions to the generality of the phenomenon of recognition failure. For example, Tulving and Watkins (Note 1) described an experiment in which the A and B items were abstract words or the A item was a two-digit number and the B item was a highly familiar word. In both cases the observed incidence of recognition failure deviated substantially from the Tulving and Wiseman function. Some deviation from the function is also apparent, where the A item was an adjective and the B item a noun, or both A and B items were adjectives, in data reported by Bartling and Thompson (1977). Begg (1979) varied word frequency and concreteness of A and B items and, under conditions of rote processing, obtained several instances where the observed incidence of recognition failure was actually 0.00. (A recognition failure rate of 0.00 was also observed in one condition of Tajika's (1977) study, but since the overall recognition level in that condition was .98, ceiling effects cannot be excluded as a reason for the observation and it cannot to be regarded as an exception). And Flexser (Note 2) has shown that marked deviations from the Tulving and Wiseman function occur when subjects form separate images of A and B items high in imagery value.[1]

The purpose of the present experiments is twofold. First, the experiments were designed to replicate the previous results of Tulving and Watkins (Note 1) and to show that with A−B pairs composed of abstract words or digits and words, recognition failure data points deviate appreciably from the Tulving and Wiseman function. Second, the experiments test two hypotheses as to why certain kinds of experimental materials produce exceptions to the general pattern of recognition failure data. One hypothesis holds that it is the properties of certain kinds of items in semantic memory, or perhaps as defined by linguistic analysis, that are responsible for exceptional findings in the recognition failure paradigm. This hypothesis is referred to as the "properties of materials" hypothesis. The alternate hypothesis, referred to as the "levels of processing" hypothesis, holds that it is the absence of adequate encoding operations for these kinds of materials that produces the observed exceptions.

Since we already know from the evidence just mentioned that exceptions to recognition failure do indeed occur with certain kinds of materials, we would not wish to argue that the properties of to-be-remembered items are unimportant in determining the phenomenon or its exact extent. Nor do we mean to suggest that properties of materials may not influence the mental activity of the learner. This means that our two hypotheses are not mutually exclusive, in the sense that one of them must be "right" and the other "wrong." Rather, in a sense which we will specify in a moment, we assume that the level of processing of the to-be-remembered material (or the sort of encoding operations performed on it at the time of study) may be the *primary* determinant of the amount of recognition failure, and that the linguistic properties of the to-be-remembered material may be only a *secondary* determinant. Fortunately, all that has been learnt from research within the levels of processing framework provides, in principle, a way of distinguishing between these possibilities.

[1] There are also some unpublished observations, by Lars-Göran Nilsson at the University of Uppsala and by Norman Park at the University of Toronto, showing that when the A member of the A−B pair is the name of a conceptual category, recognition failure is very low and deviates drastically from the Tulving and Wiseman function.

This research has shown that with all the more "traditional" learning variables—including kinds of material—held constant, huge differences in memory performance occur depending solely on the mental task given to the learner (see, e.g., Craik & Tulving, 1975). Our two hypotheses are directed at the question of whether exceptions to recognition failure depend, in a similar way, on encoding processes. And in keeping with the levels of processing approach, we can test the two hypotheses by arranging to hold constant the linguistic properties of the A–B items and then observing recognition failure rates following different kinds of encoding operations used by subjects at the time of study. The hypothesis that exceptions to recognition failure are primarily attributable to linguistic properties of the material, or the manner of its representation in semantic memory, would be rejected if it can be demonstrated that the incidence of recognition failure, relative to the Tulving and Wiseman function, depends on the level of processing of the material at the time of study. The levels of processing hypothesis, on the other hand, would be rejected if it were observed that the nature of encoding operations has little or no bearing on the amount of recognition failure obtained with certain kinds of materials, and that deviations from the Tulving and Wiseman function occur regardless of study conditions.

Accordingly, in each of the following experiments, subjects studied the same lists of abstract word pairs or digit-word pairs under one of two conditions designed to manipulate encoding processes, mainly through the use of different sets of instructions. In one condition, study conditions generally were similar to those used in most previous experiments in the recognition failure paradigm (including those described by Tulving and Watkins, Note 1). In the second, study conditions, especially the instructions given to subjects, were changed so as to foster deeper, or perhaps more elaborate and integrative, levels of processing. The latter set of instructions were suggested by results of an additional test reported by Tulving and Watkins (Note 1), which showed that subjects who *free* recalled B items recalled just as many as subjects who were given the usual cued recall test. Given that these manipulations turn out to be effective, the critical questions are (1) whether, under typical study conditions, the incidence of recognition failure deviates markedly from the Tulving and Wiseman function, and, if so, (2) whether such exceptions to recognition failure are affected by depth or integrity of encoding.

Experiment 1

Method

Subjects. The subjects were 40 undergraduate students at the University of Toronto who participated in the experiment for a course credit. They were tested either individually or in small groups. The subjects were assigned arbitrarily to one of two groups with 20 subjects in each.

Design. All subjects were presented with one set-establishing list composed of 16 pairs of familiar, nominally unrelated, words drawn from the Toronto word pool. Presentation of this list was followed by a recall test in which the A item of each study pair was re-presented as a cue for recall of the B item. The subjects were then presented with two critical lists composed respectively of pairs of abstract words and digit–word pairs. Following presentation of the second list, the subjects were engaged in a period of distractor activity. Each of the two critical lists was then tested successively for recognition of each B item and for recall of each B item given the A item. Lastly, the subjects were presented with copies of the original study lists, and asked to indicate which A–B pairs they had integrated during presentation and to describe briefly the manner in which they had integrated them. Pilot work had indicated that these arrangements of study and test trials reduced the risk of obtaining high levels of overall performance without appearing to produce any undesirable bias.

The two groups were distinguished only by the instructions given to subjects in each. Subjects in one, which we refer to as the Standard group, were given instructions similar to those usually given in the recognition failure paradigm. Subjects in the other, which we refer to as the Special group, were given explicit, detailed suggestions as to how they might study each A–B pair so as to render the A member of the pair an effective recall cue.

Materials and procedure. Two critical lists of each type were used in the experiment. For the abstract word lists, each list was formed by selecting from a pool of 144 words having a rating of less than 4.00 on the concreteness dimension in Paivio, Yuille, and Madigan's (1969) norms. Each A–B pair was chosen so as to avoid any obvious association between the two words. The A items of one list served as the B items of the other. Examples are, from one list: *Honor* ANXIETY, *Moment* ABILITY, *Blasphemy* CHANCE; from the other, *Time* HONOR, *Greed* MOMENT, *Obsession* BLASPHEMY. Each list was composed of 16 such abstract word pairs. Two separate recognition test lists were constructed, one for each presentation list, by adding 32 lure items from the same word pool to the 16 list items. A different set of lures was used in each recognition test.

For the two lists of digit–word pairs, one set of 16 two-digit numbers served as list cues for both lists. These numbers were randomly paired with 16 high-frequency words in each list. List words were taken from those used by Tulving and Thomson (1973). Examples are, from one list: *49* WET, *74* OPEN, *27* BEING; from the other: *49* SWEET, *74* HARD, *27* DIRTY. One recognition test list was constructed. It comprised target words from both presentation lists and a further 16 lure items taken from the same source.

Half of the subjects in each group were presented first with an abstract word list, half with a digit–word list. The sequence of events for half the subjects was, in detail: (i) presentation of abstract word list, (ii) presentation of digit–word list, (iii) interpolated activity, (iv) recognition then recall tests, abstract word list, (v) recognition then recall, digit–word list, (vi) re-presentation of abstract word study list, then digit–word study list. The other subjects in each group received the same test sequence but with list order reversed. The identity of a list was confounded with presentation and test order. Presentation order of list items was constant, as was test order of items in recognition and recall, though, of course, different constant orders were used in each case.

All subjects were informed that they would be presented with several sets of item pairs, and that, in each set, the left-hand member of a pair would be re-presented at test as a cue for recall of the right-hand member. Lists were presented via an overhead projector, with list cues shown in lowercase and list words in uppercase. The subjects were told about the general nature of each list before it was presented. The first set-establishing list was presented at a rate of about one A–B pair every 3 seconds. After the recall test of this list, all subjects were told that they would be presented with two further lists and that these lists would not be tested immediately.

The two critical lists were presented at the rate of about one A–B pair every 10 seconds. Apart from being informed as to the general nature of the list materials, subjects in the Standard group were given no further instructions. Subjects in the Special group, however, were given very detailed instructions, prior to each critical list presentation, as to how they should study each A–B pair. In the case of the digit–word pair lists, the subjects were told that ordinarily, the digits were no help at all in remembering which words had occurred in the list. They were then told that the digits would help provided that they thought about them in a particular way. Subjects were told that the trick was to think of each pair of digits as standing for, or representing, something more meaningful than merely a sample of arithmetic. For instance, the number 25 could represent a distance (25 miles), a time (25 minutes, or 5 minutes past two), a date (Christmas Day), someone's age (25 years), or waistline (25 inches), a year (1925), a sum of money ($25.00), a birthday (July 25th), a height (two feet five inches), a temperature (25°C), a score in a game (Blue Jays 2; Argonauts 5), a weight (25 tons), *and so on.* After subjects were given many examples of different things which numbers can represent, they were told that there was a second trick essential to making them effective cues for recall, and that was to integrate, or link, the numbers with the list word paired with them. The particular example given was "63 HOW". Subjects were told that, on encountering such a pair in the study list, they might think of 63 as the year 1963. They might then remember that 1963 was the year in which John F. Kennedy was assassinated, and if they did so, they would undoubtedly remember *how* he was assassinated. Though this example turned out to be news for quite a few subjects, nonetheless, it served to demonstrate the importance of finding some integrative link between the cue item and the word paired with it. The subjects were additionally told that, best of all, would be cases in which they could relate their encoding of the digit–word pair to some personal experience, either one pertaining directly to themselves, or one relating to someone else known to them. Alternative strategies were also suggested, though less strongly recommended, such as imagining a particular scene or situation which might serve to integrate the list cue and the list word paired with it, or even simply trying to create an image in which both the digits and the list word figured (e.g., a poster or billboard advertising some article or consumer service.)

The tenor of instructions given to subjects in the Special group with respect to the list of abstract word pairs was similar. Subjects were told that ordinarily, the abstract word cues did not help much, but that by thinking of each study list pair in certain ways, they might really aid recall. Several ways of achieving this were suggested. For instance, subjects were recommended to try to create associations between the cue and the list word. It was suggested that they might, for some pairs of items, think of a sentence which incorporated the two. It was suggested, too, that for other pairs, simply the initial letters of the two words, or some characteristic of the sound of the two words, might serve to integrate and link them together. The notion of a "portmanteau word" was also mentioned: that is, that for some word pairs, it might be possible to take parts of each word and put them together to form a third, new word. Those foregoing strategies were, however, less strongly recommended than trying to visualize, or image, a scene or situation (such as a courtroom) in which the two words might be related. More emphasis still was given to the possibility that the two words might together aptly fit some description of an event experienced by the subject, or something known from literature, movies, television series, and the like.

It should be stressed that, apart from the nature of the instructions given to subjects prior to the presentation of the critical lists, subjects in both the Standard and the Special groups were treated alike. Materials, and so forth, were yoked across groups. The subjects in each group were given as much time as they felt they needed to complete any particular test. In recognition tests, subjects were instructed to work through the set of words at their own pace and draw a circle around any word they remembered as a list word. In recall, subjects were asked to write down as many list words as they could, if possible, next to the appropriate list cue. In the case of the last test administered, in which copies of the originally presented study lists were presented, the subjects were enjoined to report only those integrative links between list cues and list words which they remembered thinking of during the study phase. All instructions in the course of the experiment were given orally.

The task interpolated between presentation and test of critical lists had the sole function of reducing the overall level of performance in the recognition tests. In this task, subjects were given, in alphabetical order, a list of the names of 96 different countries and they were asked to write down, next to the name of each country, what they thought the average person would first think of in connection with the particular country. This task was terminated after 10 minutes. And the entire experimental session lasted for about 40 minutes.

Results and Discussion

Data from the critical lists were scored in two ways, first, strictly, which means that a word was counted as having been recalled only if it was paired with the appropriate list cue, second, leniently, which means that a word was counted as having been recalled irrespective of whether it was paired with the appropriate list cue. Summaries of the data with respect to a fourfold classification of recognition and recall scores are presented in Table 1. The first two columns in the table give recognition hit rates and cued recall probabilities for each list type. Data were collapsed over the two lists of each type, and hence also over presentation and test order, since no large or systematic variation due to list or order was observed. The last two columns in the table give, respectively, the observed incidence of recognition failure and the amount of recognition failure predicted by the Tulving and Wiseman (1975) function. False positive rates in the recognition tests for the Standard group were 8% abstract word lists and 5% digit–word lists. The corresponding rates for the Special group were 7 and 3%. Data from the set-establishing list were not scored.

It is apparent from Table 1 that subjects in the Special group both recognized and recalled far more list words than did subjects in the Standard group. Performance in the Special group was reliably superior, $F(1,38) = 11.89$, $p < .01$, $MS_e = 18.77$. Also, recognition significantly exceeded recall, $F(1,38) = 157.02$, $p < .001$, $MS_e =$

TABLE 1

PROPORTIONS OF ITEMS IN DIFFERENT SCORING CATEGORIES AND PROPORTIONS OF RECOGNITION FAILURE OF RECALLABLE WORDS, EXPERIMENT 1

				Recognized				Not recognized				Recognition failure		
		Cued recall		Recalled		Not recalled		Recalled		Not recalled		Observed		
	Recognition hits	S[a]	L	S	L	S	L	S	L	S	L	S	L	Predicted
Standard group														
Abstract lists	.58	.14	.18	.12	.16	.47	.43	.03	.03	.39	.39	.18	.15	.30
Digit lists	.51	.18	.23	.17	.21	.35	.30	.01	.02	.48	.47	.07	.08	.36
Special group														
Abstract lists	.68	.28	.30	.23	.26	.45	.42	.04	.05	.28	.28	.16	.15	.21
Digit lists	.69	.37	.46	.33	.41	.36	.28	.04	.05	.27	.26	.11	.12	.20

[a] S = strict; L = lenient.

9.21; there were no differences between the two types of list, $F < 1$ and F ratios from interaction terms were less than one.[2] Additional, corroborative evidence on the effectiveness of the instructional manipulation is provided by subjective reports about which A–B pairs were integrated during the study phase of the experiment. In the Special group, subjects reported integrating 43% of the abstract word pairs and 50% of the digit–word pairs. The comparable percentages in the Standard group were 26 and 24%, respectively. These figures imply that, although subjects given special instructions were more likely to integrate each A–B pair, they were far from completely successful in so doing. Conversely, even in the absence of special instructions, it would seem that subjects spontaneously adopted the strategy of attempting to integrate at least some A–B pairs.

The data of primary interest relate to the incidence of recognition failure shown in the table. Although recognition failure occurred in the Standard group, the rates observed are substantially smaller than those predicted by the recognition failure function. The incidence of recognition failure observed in the Special group, however, approximates more closely to expected rates. Differences between observed and expected recognition failure rates were assessed directly, using macrosubject data in order to mitigate against floor and ceiling problems. For each successively tested pair of subjects in each group, the difference between expected and observed recognition failure rates was computed separately for abstract and for digit–word lists. Differences between observed and expected recognition failure rates were significantly larger in the Standard group, $F(1,18) = 6.06$, $p < .025$, $MS_e = 0.29$. There was no significant effect of list type, $F(1,18) = 3.37$, $MS_e = .029$ nor was the interaction between these two factors reliable, $F(1,18) = 1.23$. These results thus confirm the earlier observation reported by Tulving and Watkins (Note 1) that, with lists of this kind, exceptions to the Tulving and Wiseman function occur. However, results also show that, for the same lists, deviations from the function are significantly reduced with special encoding instructions.

In general, these findings are more consistent with the levels of processing hypothesis than with the properties of materials hypothesis. However, the overall pattern of results is less clear cut than one might wish. By the view that encoding operations, rather than materials per se, are primary determinants of exceptions to recognition failure, this may be attributable to the possibility that subjects in the Special group were somewhat less than completely successful in integrating A–B pairs on the one hand, and on the other, that some degree of integration seemed to occur in the Standard group. Hence Experiment 2 was designed to replicate the principal findings and, moreover, the experiment incorporated a number of procedural changes with a view to obtaining more clear-cut results. In the first experiment, all conditions apart from instructions were yoked across the two groups. In the following experiment a number of factors, including study time and amount of practice, were deliberately confounded with group condition. Also, subjects in the Special group were required to copy down each A–B pair and to note down any integrating links that occurred to them during the study phase. These changes were made in order to try to increase the likelihood of successful integration in the Special group, and to decrease its likelihood in the Standard group.

Experiment 2

Method

Subjects. The subjects were 40 undergraduate students at the University of Toronto. They participated in the experiment either for pay or for a course credit. None had taken part in Experiment 1. The sub-

[2] All statistical analyses reported in results of this and the following experiment are based on data obtained with strict scoring.

jects were assigned arbitrarily to one of two groups with 20 subjects in each, and they were tested either individually or two at a time.

Design. All subjects were presented with the same set-establishing list that was used in Experiment 1 and tested immediately for cued recall. Following this, subjects in the Standard group were presented with two critical lists, composed respectively, of abstract words and digit–word pairs. Each of the two lists was then tested successively for recognition and cued recall of list words, without any distractor activity interpolated between study and test. After the second cued recall test, subjects were asked to look through each recall protocol and indicate any A–B pairs for which they had thought of some integrating link at the time of study.

Following the presentation and recall of the initial set-establishing list, subjects in the Special group were then given two further practice trials, one with each list type, and without any further instructions. The point of these trials was to bring home to subjects the nature of the critical material and hence lend weight to the special instructions then given. Each of the two critical lists was presented twice in succession. During each presentation, subjects worked through a small booklet, copying each A–B pair down on a separate page and also writing down any integrating links which occurred to them at the time. Recognition and recall tests of critical list words were administered in a second session about 24 hours later.

Materials and procedure. Materials for the critical lists, at both presentation and test, and instructions given to subjects, were the same as those used in Experiment 1. The two additional practice lists for subjects in the Special group were constructed from the same source and in the same way as were critical lists. The lists were presented in the same manner as in the previous experiment, except with respect to study time. The A–B pairs in the initial set-establishing list, and in the two additional practice lists, were presented at about a 2-second rate. The critical lists for subjects in the Standard group were also presented at this rate. Subjects in the Special group were, however, presented with the critical lists at a slower and more flexible rate. A single list took typically between 3 and 5 minutes to present. To some extent the exact timing was at the discretion of the subject and depended upon when a person completed the copying of an A–B pair and noting down any integrating relation. If, after about 10 seconds, it appeared that the subject had not integrated the A–B pair, the experimenter presented the next pair. Prior to the critical lists being presented, subjects were told that each list would be shown twice in succession and that lists would not be tested until the following day. On the second presentation of each list, subjects were told that in addition to copying each A–B pair, they should write down again any integrating relation they had thought of first time round and try to integrate pairs which they had not integrated before.

Subjects were again allowed as much time as they felt they needed to complete any particular test. For subjects in the Standard group, the single experimental session lasted about 20 minutes. For subjects in the Special group, the first session lasted between 40 and 50 minutes, the second about 10 minutes. The sole purpose of the 24-hour retention interval, in the latter case, was to prevent recognition performance attaining ceiling levels.

Results and Discussion

Data from the critical lists were scored both strictly and leniently, and the principal data are summarized in Table 2. As in Experiment 1, data were collapsed over the two lists of each type, and hence also over presentation and test order, since no apparent effect due to list identity or order was observed. False positive positive rates in the recognition tests for the Standard group

TABLE 2

Proportions of Items in Different Scoring Categories and Proportions of Recognition Failure of Recallable Words, Experiment 2

				Recognized				Not recognized				Recognition failure		
		Cued recall		Recalled		Not recalled		Recalled		Not recalled		Observed		
	Recognition hits	S[a]	L	S	L	S	L	S	L	S	L	S	L	Predicted
Standard group														
Abstract lists	.58	.11	.22	.09	.20	.49	.38	.02	.03	.41	.40	.18	.11	.30
Digit lists	.53	.13	.29	.11	.26	.42	.27	.02	.03	.45	.44	.15	.11	.34
Special group														
Abstract lists	.69	.48	.50	.40	.42	.29	.27	.08	.08	.23	.23	.16	.16	.20
Digit lists	.74	.52	.61	.45	.53	.29	.21	.07	.08	.20	.18	.13	.13	.16

[a] S = strict; L = lenient.

were 7% abstract word lists, and 6% digit–word lists. The corresponding rates for the Special group were 4% and less than 1%. Data from practice and set-establishing lists were not scored.

The percentage of A–B pairs which subjects in the Special group reported integrating during the study phase was 92% with abstract lists, and 85% with digit lists. These percentages are for both presentations of a list combined, and by far the majority of the integrating links reported were reported during the first of the two presentation trials. It is clear that subjects in this group were quite successful in integrating each A–B pair. By contrast, very little spontaneous integration was reported on the recall protocols of subjects in the Standard group. Based on the total number of A–B pairs presented (as are the above percentages) the corresponding figures are less than 1% for each list type. It should be noted, however, that if only because these subjective reports were obtained at different points in the procedure within each group, the percentage scores are not strictly comparable in any literal sense. At best, these data simply provide corroborative evidence for rather gross differences in the extent to which subjects in each group integrated A–B pairs.

Consider now overall performance in recognition and cued recall. Despite the 24-hour retention interval in the Special group, both recognition and recall scores are substantially higher than those in the Standard group, $F(1,38) = 54.18, p < .001, MS_e = 13.73$. Also, recognition was significantly superior to recall, $F(1,38) = 184.71, p < .001, MS_e = 5.94$, and there was no difference between the two list types, $F < 1$. And, in contrast to the results of Experiment 1, the interaction between group conditions and test mode was reliable, $F(1,38) = 20.33, p < .001$, indicating that between-group differences were higher in recall than in recognition. No other interaction term approached statistical significance.

More important, it is apparent from the last two columns of the table that once again the recognition failure rates observed in the Standard group deviate considerably from the rates expected from the recognition failure function. The recognition failure rates observed in the Special group are, by contrast, quite close to those predicted by the function. The extent to which recognition failure data points approximate more closely to the function in the Special group was assessed by means of a similar macrosubject analysis to that used in Experiment 1. For each successively tested pair of subjects in each group, the difference between expected and observed recognition failure rates was computed separately for abstract and for digit–word lists. Differences between observed and expected recognition failure rates were significantly greater in the Standard group, $F(1,18) = 69.76, p < .001, MS_e = .004$. There was no significant effect of list type, $F(1,18) = 1.17, MS_e = .044$, nor was the interaction between these two factors reliable, $F(1,18) = 1.04$. These results replicate those of Experiment 1. It is clear that when lists composed of abstract word pairs or digit–word pairs are presented under standard study conditions in the recognition failure paradigm, exceptions to recognition failure occur, in the sense that recognition failure data points deviate markedly from the Tulving and Wiseman function. It is clear, too, that these exceptions depend primarily on encoding operations that occur during the study period, rather than on properties of to-be-remembered items as such: under special study conditions, deviations from the recognition failure function were, if not eliminated, sharply reduced. Recognition failure data points fell either relatively distant from or relatively close to the recognition failure function, for each list type, depending on the extent to which subjects apparently succeeded in integrating the A and B items of each A–B pair.

General Discussion

The two experiments described here replicate Tulving and Watkins' (Note 1) earlier observation that when abstract word pairs or digit–word pairs are studied under conditions typical of the recognition failure paradigm, the incidence of recognition failure deviates considerably from that expected on the basis of the Tulving and Wiseman (1975) recognition failure function. Results also show that under study conditions in which, for the same materials, deeper, more integrative encoding is achieved, recognition failure rates approximate much more closely to the Tulving and Wiseman function. These findings, therefore, demonstrate exceptions to recognition failure, in the sense of marked deviations from the recognition failure function, and show that those exceptions are determined, at least in part, by the nature of encoding operations carried out at the time of study. In neither of the two experiments we report were the special encoding instructions completely successful in bringing recognition failure data points directly on the recognition failure function. A reasonable interpretation of this failure is that the kinds of processing necessary for typical results to occur were not achieved by some subjects with some to-be-remembered units.

These findings, as well as those reported elsewhere (e.g., Begg, 1979; Flexser, Note 2; Tulving and Watkins, Note 1), confirm that systematic deviations from the Tulving and Wiseman function can be readily produced through the use of certain kinds of to-be-remembered materials. Our experiments, however, as well as those described by Begg (1979) and Flexser (Note 2), also suggest that with the very same materials the levels of recognition failure predicted by the Tulving and Wiseman function could be readily achieved if subjects are induced to form special, integrative encoding operations on the to-be-remembered material at the time of study. Since in these experiments the kinds of materials are held constant across different study conditions, such results provide little or no support for the properties of materials hypothesis, that is, for the notion that these exceptions to recognition failure are determined primarily by linguistic characteristics of the A–B pairs or the manner of their representation in semantic memory. Indeed, the fact that the two types of list used in the present experiments, each—presumably—having rather dissimilar linguistic or semantic memory characteristics, produce recognition failure data points whose relation to the Tulving and Wiseman function is statistically indistinguishable, in itself argues against the properties of materials hypothesis. By the same token, that variant of generate-recognize theory sometimes called the semantic interpretation, or the multinode version, is still further weakened by our findings (for more discussion see, e.g., Tulving & Watkins, 1977; Watkins & Gardiner, 1979).

Thus, together with other relevant findings mentioned above, our results favor the levels of processing hypothesis. But it should be emphasized that in rejecting the properties of materials hypothesis we do not wish to argue that the nature of the to-be-remembered materials has no effect whatsoever upon the likelihood of obtaining systematic deviations from the Tulving and Wiseman function. Rather, by accepting the levels of processing hypothesis, we mean to argue merely that the nature of the to-be-remembered material is not a *primary* determinant of the incidence of recognition failure in relation to the recognition failure function. The primary determinant is the adequacy or integrity of the encoding operations carried out at the time of study. This means simply that certain kinds of materials may, when presented under conventional study conditions, normally lead to a lack of integrity, or a deficiency in encoding A–B units, but that any such deficiency or lack of integrity is not an inevitable consequence of the use of those materials. Nature of to-be-remembered material may, and frequently does, influence nature of encoding, but it is the nature of encoding that directly determines exceptions to recognition failure.

In the introduction to the paper we pointed out that the concept of exceptions could be defined either simply in terms of the overall conditional probability that recognition fails given that recall succeeds or in terms of deviations from the Tulving and Wiseman recognition failure function. It is important to note that if one wished to opt for the former rather than the latter definition, and consider only the absolute incidence of recognition failure, very different conclusions would be drawn from the experimental results. Averaged over both experiments, and both types of list, the overall recognition failure rate turns out to be 0.14 under standard study conditions . . . and 0.14 under special study conditions. Hence it would be concluded (1) that experiments did not demonstrate exceptions to recognition failure and (2) that differences in encoding operations had no effect on the incidence of recognition failure. It would then follow that experimental results can be interpreted as being entirely consistent with the view that recognition failure is primarily determined by the linguistic properties of to-be-remembered materials. And so the levels of processing hypothesis would be rejected.

We do not, of course, accept this reasoning. As argued in the introduction, given the general pattern of recognition failure data summarized by the Tulving and Wiseman function we think it appropriate to conceptualize exceptions to recognition failure in terms of systematic deviations from that function. And there is an important corollary to this argument. Some theoretical accounts of recognition failure focus on explaining the relative magnitude of observed recognition failure rates in absolute terms. Theories of this kind do not attempt to account for the relation between recognition failure and overall levels of recognition—the Tulving and Wiseman function. By the same token, such theories cannot account for exceptions to recognition failure conceived as deviations from the function. For these reasons, the dual-code theory recently advanced by Mandler and his associates (e.g., Rabinowitz, Mandler, & Barsalou, 1977), and the rather similar explanation of recognition failure proposed by Bartling and Thompson (1977) are not directly relevant to our findings. Several other recent theoretical ideas, however, are explicitly directed at explanation of the recognition failure function (e.g., Begg, 1979; Flexser & Tulving, 1978; Jones, 1978; Kintsch, 1978). How well do these theories handle our results?

Consider first Jones' (1978) theory. It predicts that the value of the constant c in Equation 1 is given by

$$c = \frac{P(\mathrm{G})\,[1 - P(\mathrm{E})]}{P(Rc)} \qquad [2]$$

where P(G) represents the probability of *generating* the B item given the A item as a cue (the use of "extrinsic knowledge") and P(E) represents the probability of encoding "intrinsic knowledge" of the co-occurrence of A and B items as members of a studied pair. Since in the present experiments P(G) was held constant for a given type of material, $cP(Rc)$ should vary *inversely* with depth or integrity of encoding, that is, $cP(Rc)$ should be lower under special than standard study conditions. The data presented in Table 3 do not support this prediction. In eight possible comparisons, four based on strict scoring and four on lenient scoring, $cP(Rc)$ was higher under special than standard study conditions, that is, $cP(Rc)$ varied *directly* with depth or integrity of encoding. Thus, Jones' (1978) model does not give a good account of the present data.

TABLE 3
VALUES OF THE TULVING–WISEMAN CONSTANT c AND THE PRODUCT $cP(Rc)$

Condition (Experiment, list, instructions)	Constant c		Product $cP(Rc)$	
	S	L	S	L
Experiment 1				
Abstract–standard	.98	1.11	.14	.20
Abstract–special	.73	.78	.20	.23
Digit–standard	1.68	1.64	.30	.38
Digit–special	.94	.89	.35	.41
Experiment 2				
Abstract–standard	.98	1.27	.11	.28
Abstract–special	.70	.70	.34	.35
Digit–standard	1.28	1.45	.09	.20
Digit–special	.68	.68	.35	.41

Kintsch's (1978) variable-criterion generation/recognition theory of recognition failure could accommodate the present data by introducing the assumption that the relative differences between recognition and recall criteria in our experiments varied with study conditions. Specifically, the assumption would have to be that the recall criterion was much lower in relation to the recognition criterion in the special study condition than in the standard condition. The major problem with this assumption is that it is entirely post hoc: The reasons for such a relation between depth or integrity of encoding and the corresponding changes in recall and recognition criteria are unexplained. We conclude, therefore, that Kintsch's (1978) theory does not provide a good explanation of our experimental results.

How well does Begg's "vandal theory" cope with these data? The question is particularly appropriate since Begg's own experimental data that led him to postulate the "vandal theory" were derived from experiments quite similar to ours; moreover, the overall pattern of his data was very much the same as ours: data points showing large deviations from the recognition failure function under rote study conditions and very much closer to the function under integrative encoding conditions. The vandal theory directly attributes deviations from the recognition failure function to high degrees of trace loss. On this basis, then, one would have to argue that study conditions in our standard groups were more conducive to high rates of trace loss than study conditions in our special groups. Although such a statement agrees with the observed data, we have some doubts about the appropriateness of the interpretation. The notion of trace loss is itself somewhat vague, and it is not clear how different degrees of trace loss relate to differences in *encoding* processes: The inferential link between theory and data seems tenuous. We think that Begg's theory would receive better support if it could be shown that, when to-be-remembered materials and encoding conditions are held constant, recognition failure data points increasingly deviate from the recognition failure function with longer retention intervals. Since available data are at variance with this deduction from Begg's theory (e.g., Tulving & Watkins, 1977; see also Watkins & Gardiner, 1979), the notion of trace loss does not seem to provide an acceptable explanation of deviations of the data from the Tulving and Wiseman function.

Finally, what about Flexser and Tulving's (1978) retrieval independence theory? Systematic deviations of data points from the recognition failure function can occur in the Flexser and Tulving model, both as a consequence of particular combinations of parameter values in their special model, and as a result of changes in the k parameter in their general model. The parameter k in the general model may assume values between zero and unity, where a value of 0 represents retrieval independence, in the sense that the encoded features of the A item presented in the cued-recall test are uncorrelated with the encoded features of the B item in the recognition test, and where a value of 1 represents a situation in which all encoded features of A are included in the encoded features of B. Tests by Flexser and Tulving (1978) of their general model yielded simulated recognition failure data points indistinguishable from the actual pattern of experimental data when k assumed a small value near 0. As k approached unity, increasingly large deviations from the recognition failure function occurred (see Flexser & Tulving, 1978, Figure 8). It could be assumed then that the large deviations from the recognition failure function observed in the experiments described here depict a situation where the parameter k in the Flexser and Tulving general model assumes a large value, whereas closer proximity of data points to the function under special conditions corresponds to a situation described by a smaller value of k. But this assumption is post hoc. And it is not at all obvious why integration of A and B members should produce *smaller* degrees of feature overlap between A and B than would lack of integration. Thus retrieval independence theory, too, seems unable to give a ready interpretation of our findings.

It would appear, therefore, that none of these theories can readily explain our experimental results. Hence these results, together with the other similar findings we mentioned earlier, pose a new theoretical challenge. The challenge consists in accounting for large, systematic deviations from the recognition failure function and explaining why such deviations depend upon levels of processing. Of course it is not impossible that one or more of the theories directed at explanation of the recognition failure function might be modified or elaborated so as to bring them into better agreement with the exceptions with which we are concerned. It might be somewhat premature to attempt to achieve such a reconciliation of data with theory just yet, however. And there are a few problems entailed in meeting this theoretical challenge.

With high levels of overall recognition it becomes difficult, if not impossible, to *detect* deviations from the recognition failure function. The recognition failure data points observed in the present experiments were reasonably well below ceiling levels but, as we noted in discussing the concept of exceptions, this has not always been the case in previous experiments. Thus it is possible that in at least some conditions of previous experiments where high recognition scores were obtained, deviations from the recognition failure function might have gone undetected. This possibility may render less surprising a necessary and perhaps unreasonable corollary to our argument: That in all conditions of all previous experiments where the incidence of recognition failure conformed to the function, an adequate degree of integration was achieved. Lack of experimental sensitivity may also be reflected by the absence of a much more extreme pattern of recognition failure in Experiment 2, compared with that observed in Experiment 1. If the procedural changes incorporated into the design of Experiment 2 increased the likelihood of A–B pair integration under special encoding conditions and decreased its likelihood under standard encoding conditions, then, relative to Experiment 1, it would be expected that recognition failure data points should have fallen closer to, and more distant from the recognition failure function. The experimental results provided little support for this outcome. But these considerations highlight a more fundamental issue. At present we have no independent means of establishing the extent of A–B pair integration that may be regarded as sufficient for recognition failure data points to conform to the Tulving and Wiseman function. How might this issue be tackled? One possible approach is indicated not only by the discrepancies between recognition failure rates sometimes observed with strict and lenient scores, but also by a great deal of experimental data that suggest that probability of recognition failure in comparisons of *free* recall and recognition is quite low and deviates drastically from the function. It might not be entirely unprofitable to begin to explore the idea that exceptions to recognition failure occur under conditions in which a nominally cued recall test in some sense functionally corresponds with a test of free recall.

References

Bahrick, H. P. Broader methods and narrower theories for memory research: Comments on the papers by Eysenck and Cermak. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, N.J.: Erlbaum, 1979.

Bartling, C. A., & Thompson, C. P. Encoding specificity: Retrieval asymmetry in the recognition failure paradigm. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, 3, 690–700.

Begg, I. Trace loss and the recognition failure of unrecalled words. *Memory & Cognition*, 1979, 7, 113–123.

Craik, F. I. M., & Tulving, E. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 1975, 1, 268–294.

Flexser, A. J., & Tulving, E. Retrieval independence in recognition and recall. *Psychological Review, 1978*, 85, 153–171.

Jones, G. V. Recognition failure and dual mechanisms in recall. *Psychological Review*, 1978, 85, 464–469.

Kintsch, W. More on recognition failure of recallable words: Implications for generation-recognition models. *Psychological Review*, 1978, 85, 470–473.

Light, L. L., Kimble, G. A., & Pellegrino, J. W. Comments on "Episodic memory: When recognition fails," by Watkins & Tulving. *Journal of Experimental Psychology: General*, 1975, 104, 30–36.

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

Murdock, B. B., Jr. Methodology in the study of human memory. In W. E. Estes (Ed.), *Handbook of learning and cognitive processes*. Vol. 4. Hillsdale, N.J.: Erlbaum, 1976.

Paivio, A., Yuille, J. C., & Madigan, S. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph*, 1968, 76 (1, Pt. 2).

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

Rabinowitz, J. C., Mandler, G., & Barsalou, L. W. Recognition failure: Another case of retrieval failure. *Journal of Verbal Learning and Verbal Behavior*, 1977, 16, 639–663.

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

Salzberg, P. M. On the generality of encoding specificity. *Journal of Experimental Psychology: Human Learning and Memory*, 1976, 2, 586–596.

Tajika, H. Features of recognition tasks in encoding specificity: Types of frequency associates in extralist cue words and types of recognition tasks. *Psychologia—An International Journal of Psychology in the Orient*, 1977, 20, 151–158.

Tajika, H. [Features of recognition tasks in encoding specificity: The function of free associates tasks and non-verbal tasks.] *The Japanese Journal of Psychology*, 1978, 48, 344–347.

Tajika, H. Memory processes in recall and recognition. *Psychologia—An International Journal of Psychology in the Orient*, 1979, 22. 146–154.

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

Tulving, E., & Watkins, O. C. Recognition failure of words with a single meaning. *Memory & Cognition*, 1977, 5, 513–522.

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

Wallace, W. P. Recognition failure of recallable words and recognizable words. *Journal of Experimental Psychology: Human Learning and Memory*, 1978, 4, 441–452.

Watkins, M. J., & Gardiner, J. M. An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior*, 1979, 18, 687–704.

Watkins, M. J., & Tulving, E. Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, 1975, 1, 5–29.

Reference Notes

1. Tulving, E., & Watkins, O. C. *Encoding operations and recognition failure*. Paper presented at the 17th annual meeting of the Psychonomic Society, St. Louis, Missouri, November, 1976.
2. Flexser, A. J. *Can we manipulate the degree of interdependence between recognition and recall?* Unpublished manuscript, University of Toronto, 1977.

(Received August 28, 1979)