Similarity Relations in Recognition

ENDEL TULVING

University of Toronto

Under certain conditions, subjects in a forced-choice recognition task can discriminate between targets and distractors more accurately when the targets and distractors are similar than when they are dissimilar. This reversal of the conventional result is demonstrated in two picture-recognition experiments. The results of the experiments suggest that two kinds of similarity relations—perceptual and ecphoric similarity—must be specified in descriptions of the phenomena of forced-choice recognition memory.

Similarity plays an important role in determining recognition memory. The similarity between the old and the new test items has been referred to as "a very powerful variable" (Kintsch, 1970, p. 221) or as "the most important of stimulus variables affecting perception and recognition alike" (Shepard & Podgorny, 1978, p. 208). Many experimental demonstrations of the relation between recognition performance and the similarity between and among test items have been reported in the literature (e.g., Bahrick, Clark, & Bahrick, 1967; Bower & Glass, 1976; Jörg & Hörmann, 1978; Klein & Arbuckle, 1970; McNulty, 1965; Nagae, 1980; Shepard & Chang, 1963; Weaver & Stanny, 1978; Wyant, Banks, Berger, & Wright, 1972). These demonstrations have involved different kinds of materials—letter strings, words, schematic drawings, and pictures—and both the Yes/No and forced-choice recognition tests. The outcome of the demonstrations has always been the same: Recognition accuracy is inversely related to the similarity between the old and the new test items. Reviewers of the literature, without exception, agree in reporting the empirical generalization. Thus, for instance, we are told that "the probability of correct choice on a multiple-choice recognition task decreases the greater the similarity of the incorrect alternatives to the correct alternatives" (Wickelgren, 1977, p. 404), that "accuracy of recognition memory has been found to decline with increasing similarity . . . between the two alternatives in each forced-choice test" (Shepard & Podgorny, 1978, p. 204), that "a recognition test may be made very difficult by using new items that closely resemble the old" (Woodworth, 1938), and that "we can make any recognition test as difficult as we want simply by making the distractors extremely similar to the correct alternative" (Glass, Holyoak, & Santa, 1979, p. 65). In the psychological world of uncertain facts such uniformity of findings and harmony among writers is a rare phenomenon.

The purpose of this paper is to propose a necessary qualification to the generalization that recognition accuracy is *inversely* related to the similarity between the old and the new test items. Two experiments are described whose results show that under certain conditions recognition accuracy varies *directly* with the test-item similarity. On the basis of these results it will be argued that two kinds of similarity relations must be specified for a more complete description of recognition memory tapped by forced-choice tests, and that the generally

This research was supported by Grant A8632 from the National Sciences and Engineering Research Council of Canada. A preliminary report of the experiments was given at the 1979 Meeting of the Psychonomic Society in Phoenix, Ariz. I am grateful to Ian Spence for consultation on data analysis and to Carol A. Macdonald for assistance with all phases of the research. Requests for reprints should be sent to Endel Tulving, Department of Psychology, University of Toronto, Toronto M5S 1A1, Canada.

0022-5371/81/050479-18$02.00/0

known inverse relation between recognition accuracy and test-item similarity represents a special case of a somewhat more complex situation.

The research reported here was motivated by three considerations. First, the empirical generalization about the effect of test-item similarity is worth thinking about, because normally we would not expect to find any simple effects that hold without exception over a wide range of conditions (Jenkins, 1979). One can suspect, therefore, that in the present instance, too, exceptions must exist even if so far they have not been identified. On the other hand, if determined attempts to find exceptions to the general rule fail, the case for the universality of the effect would be greatly strengthened.

Second, the test-item similarity effect in recognition memory could be thought of as a *retrieval* effect: With encoding conditions held constant, the outcome of the measurement operation depends on the conditions prevailing at the time of the test. On this view, it would be natural to seek understanding of such a retrieval effect in terms of processes operating at the time of the test. Yet, I have argued previously that both encoding and retrieval conditions must be stipulated when we describe data from memory experiments or make theoretical inferences from them, and that it is futile to try to understand remembering only in terms of retrieval processes (Tulving, 1979). Again, reconsideration of the test-item similarity effect in recognition seems to be called for. The effect either constitutes an exception or, along with other memory phenomena, conforms to the principle just mentioned.

The third consideration is perhaps the most important one. It has to do with the results of some experiments in which the typical test-item similarity effect failed to materialize. These experiments were done as part of a series concerned with the phenomenon of recognition failure of recallable words. In the initial experiments in which the phenomenon was observed (Tulving & Thomson, 1973), subjects generated their own recognition tests by producing free associates to stimulus words that were strongly related to the target words. This procedure resulted in recognition tests in which the distractor words were semantically similar not only to the target words but also to one another. In subsequent explorations of the generality of the phenomenon of recognition failure it seemed important, among other things, to evaluate the role played by the semantic similarity of the distractors. Therefore, we directly manipulated the semantic relatedness of the distractors in several experiments (Watkins & Tulving, 1975, Experiments 5 and 6; Wiseman & Tulving, 1976, Exp. 3) and compared recognition of target words for which the distractor items were semantically related with target words whose distractors were semantically unrelated. In addition to these experiments, Rabinowitz, Mandler, and Barsalou (1977, Experiment 2) also investigated the effect of associatively related and unrelated distractors.

The relevant data from these experiments are summarized in Table 1.[1] These are the hit-rate data from the forced-choice conditions in the experiments. (The data are rather similar for the Yes/No tests, but because of the method used in the construction of the recognition tests, these data are less compelling than the forced-choice data.)

The data in Table 1 provide no hint of superior recognition performance with semantically unrelated distractors. Indeed, in five out of six comparisons the hit rates are numerically higher for the related distractors, although the difference is not significant in any one of the individual experiments.

When we reported the Watkins and Tulving (1975) and the Wiseman and Tulving (1976) experiments, we did not discuss

[1] I thank Jan Rabinowitz for making available the forced-choice recognition data from Experiment 2 of Rabinowitz et al. (1977).

TABLE 1
FORCED-CHOICE RECOGNITION PERFORMANCE (HIT RATE) WITH SEMANTICALLY RELATED AND UNRELATED LURES IN PREVIOUSLY PUBLISHED EXPERIMENTS

Experiment	Number of alternatives	Lures	
		Related	Unrelated
Watkins & Tulving (1975)			
Expt 5	3	.81	.78
Expt 6	3	.67	.63
Wiseman & Tulving (1976)			
Expt 3, RN-RC	3	.69	.68
Expt 3, RC-RN	3	.76	.74
Expt 4	3	.71	.71
Rabinowitz et al. (1977)			
Expt 2	3	.72	.70

the curious absence of the effect of distractor similarity. Such absence, for our purposes at that time, simply meant that the phenomenon of recognition failure was not critically dependent on the nature of distractors in our initial subject-generated recognition tests.

In the Watkins and Tulving (1975) and the Wiseman and Tulving (1976) experiments we used a particular design in comparing related with unrelated distractors. It is quite possible that this design, schematically depicted in Table 2, was responsible for the absence of the typical test-item similarity effects. The letters A, B, C, D, E, and F refer to copies of target words, and the same letters with primes and double primes represent their corresponding semantically related words serving as distractors. Each row in the table represents a test set, consisting of a target and two distractors. In the odd-numbered rows, the targets are accompanied by two related distractors. In each even-numbered row, the target is accompanied by two unrelated distractors. Thus, in half the test sets the subject would select the target from among three related words and in the other half of test sets the selection would be made from among three unrelated test items.

TABLE 2
SCHEMATIC DESIGN OF WITHIN-LIST MANIPULATION OF TEST-ITEM SIMILARITY

Target	Distractors	
A	A′	A″
B	D′	F″
C	C′	C″
D	F′	B″
E	E′	E″
F	B′	D″

The feature of the design that may have been responsible for the absence of the test-item similarity effect may have been the fact that the distractors unrelated to the target in a particular test set were nevertheless related to other targets in the studied list. If we assume that learners confuse two similar test items not because they resemble each other but because both of them are similar to the information stored about one or more previously encountered targets, then it is not particularly surprising that the typical test-item similarity effect was absent in our experiments. In other words, if selection of test items is governed by the similarity of test items to the information stored (Underwood, 1965; Anisfeld & Knapp, 1968), then the within-test arrangement of test items into particular test sets should not greatly affect the tendency with which test items, both targets and distractors, are identified as "old." In experiments whose data are summarized in Table 1 the experimenters manipulated similarity of items in test sets, but did not manipulate the similarity between test items and the information stored. Conversely, in other experiments in which the typical test-item similarity effects have been demonstrated these two kinds of similarity relations have always been confounded.

The first experiment described here was designed to demonstrate that the test-item similarity effect depends on the similarity between test items and the information stored, rather than on the similarity between items in test sets. The expectation was that the typical inverse relation between test-item similarity and recognition performance would be obtained under conditions where the dissimilar distractors were dissimilar not only to the targets in their respective test sets but also dissimilar to other studied list items. It was also expected that no such effect would be obtained under conditions where the dissimilar distractors were related to other studied list items. The confirmation of these expectations would lend support to the interpretation just given for the experimental results summarized in Table 1.

Experiment 1

Subjects studied a large number of complex pictures and were then tested for some of the studied pictures in a series of two-alternative forced-choice recognition tests. There were three conditions, manipulated within a single series of tests. In one, the distractor within a test pair was similar to the target item; in the second, the distractor was dissimilar to the target but similar to another previously studied picture; in the third, the distractor was dissimilar to both the target and other pictures in the study series. The comparison of recognition performance in the first and second conditions corresponds to experiments in which test-item similarity effects were absent, whereas the comparison between the first and the third condition corresponds to the conventional experiments yielding typical test-item similarity effects. The expectation was, therefore, that no differences in recognition performance would be found between the first and second conditions, and that the performance in both of them would be lower than in the third condition.

Method

Design. All subjects were exposed to a study sequence of 160 pictures of which 48 pictures, presented as an undifferentiated block in the middle of the sequence, constituted critical items whose recognition was subsequently assessed. The first 56 and the last 56 pictures in the study sequence served only as buffer items and were not tested. The large number of buffer items was used in order to bring the recognition hit rate into the middle of the possible performance range.

Let us designate the studied pictures and their identical copies, serving as target items at test, as A, B, C . . . , pictures similar to them, and serving as distractors in the test as A′, B′, C′ . . . , and pictures dissimilar to any previously studied items, also serving as distractors, as X′, Y′, Z′. . . .

Three experimental conditions can then be defined as follows: (a) Condition of perceptual similarity: A–A′. The distractor in each pair (A′) is perceptually similar to the target (A). (b) Condition of referred similarity: A–B′. The distractor (B′) is not similar to the target (A) in the same test pair, but it is similar to another previously studied picture (B). (c) Condition of dissimilarity: A–X′. The distractor (X′) is similar to neither the target in the test pair (A) nor to any other previously studied picture, although it is similar to another picture (X) previously not seen by the subject.

In the recognition test, each subject was tested with 12 pairs of items in each of the three experimental conditions, or a total of 36 test pairs. Each pair consisted of a target and a distractor. Since the targets corresponding to the distractors in the A–B′ conditions did not appear in the recognition test, 48 critical items in the study sequence were necessary for the testing of 36 pairs as described.[2]

A basic pool of 72 target pictures and their corresponding similar distractors were divided into three subsets of 24. The targets (A, B, C, . . .) appearing in any particular study sequence were drawn from two of the three subsets, while the dissimilar distractors (X′, Y′, Z′, . . .) were drawn from the third subset. Each of the three subsets of 24 served equally frequently as a source of dissimilar distractors. Moreover, within each subset the pictures serving as targets for half the subjects served as distractors for the other half, and vice versa. The result of this balancing procedure was that the nominal identity of target pictures and the distractors was statistically constant in all three experimental conditions. Consequently subjects' selection of test alternatives could not be attributed to differences in preexperimental properties of the critical pictures used.

Pictures. Each picture serving as a target or a distractor in the experiment represented one-half of a colored complex picture that had appeared across two adjacent pages in a popular magazine. The two-page pictures represented outdoor scenes, landscapes, members of the animal kingdom, people in various situations and activities, and so on. Examples of the two halves of two-page pictures are shown in Figure 1. If the left half of a double picture was used in the critical part of the study sequence, then the right half represented the corresponding similar distractor, and vice versa. Thus, the test pair A–A′ in the perceptual similarity condition consisted of the two halves of one and the same two-page picture, whereas test pairs A–B′ in the referred similarity condition and A–X′ in the dissimilar condition consisted of halves of different two-page pictures. Examples of test pairs of pictures in the three experimental conditions are shown in Figure 2.

Subjects and procedure. Forty-eight undergraduate students of both sexes at the University of Toronto participated as subjects in the experiment. They signed up for participation in return for modest remuneration.

Subjects were tested individually or in pairs. Upon entering the experimental room they were given general information about the nature of the experiment. They were told that they would see a long series of pictures and that later on their recognition memory for the pictures would be tested by the two-alternative forced-choice method. A short demonstration, involving six pairs of pictures, and illustrating all three experimental conditions, was given as a part of the orientation procedure. The experimenter described the three test conditions as follows: (a) "The two pictures in a test pair are similar to each other, but you will have seen only one of them in the presentation sequence," (b) "The two pictures in a pair are not similar to each other; you will have seen one, but the other one also resembles a picture from the presentation sequence," (c) "The two pictures in a pair are not similar to each other; you will have seen one of the two; the other one does not resemble any picture from the presentation sequence."

After the orienting instructions and the initial demonstration, the 160 pictures of the study sequence were shown, one at a time. The pictures were projected on one of two rear-projection screens standing side by side. The rate of presentation was 2 seconds per picture. There was a short interval after the presentation of the 40th picture, and another one after the presentation of the 120th picture, during which the experimenter changed the slide trays in the projector. No mention was made of the fact that the recognition test would entail only

[2] To illustrate the design in concrete terms, consider four critical study items: A, B, C, and D. The three test conditions corresponding to this subset consist of pairs A–A′, B–C′, and D–X′. The study item C does not appear in the test but it is a critical item in the study list as it defines the identity of the distractor C′ in condition B–C′.

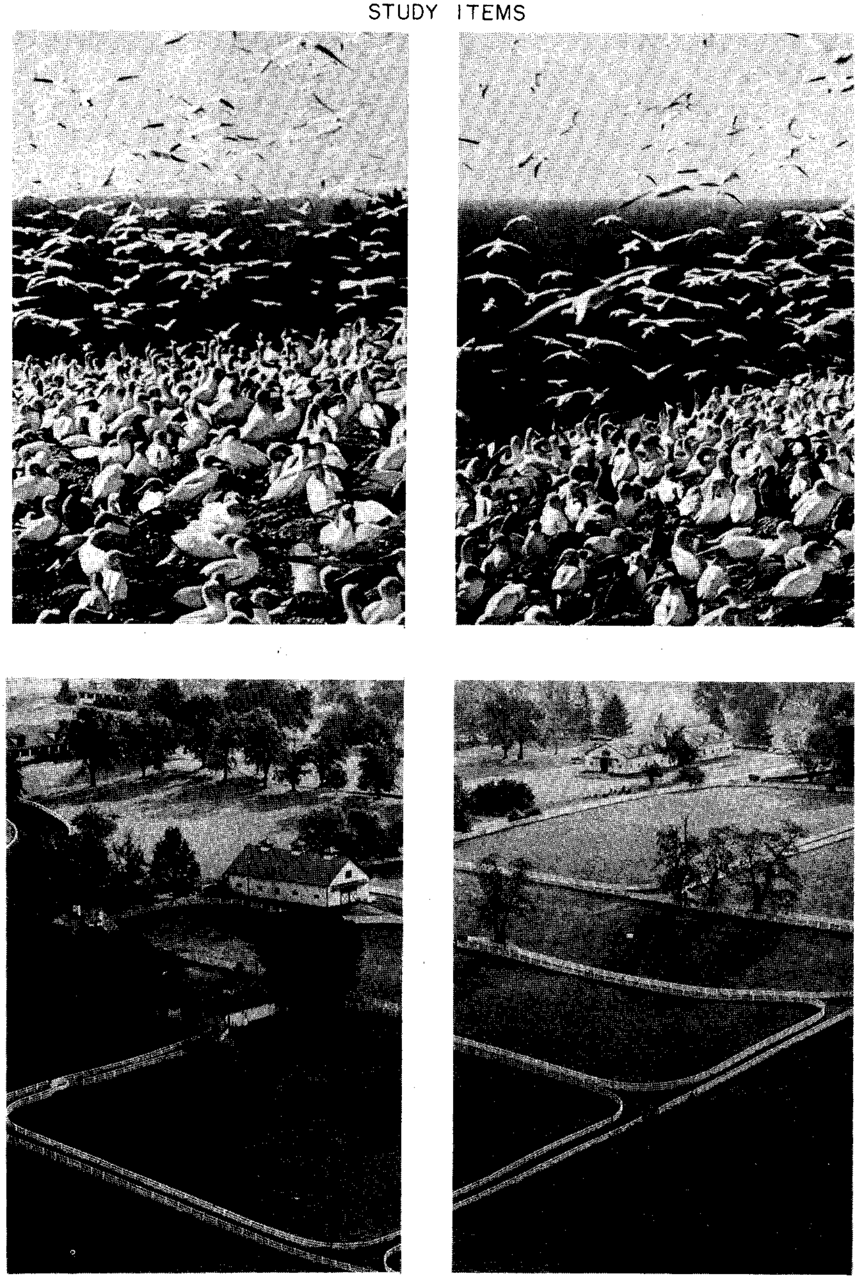
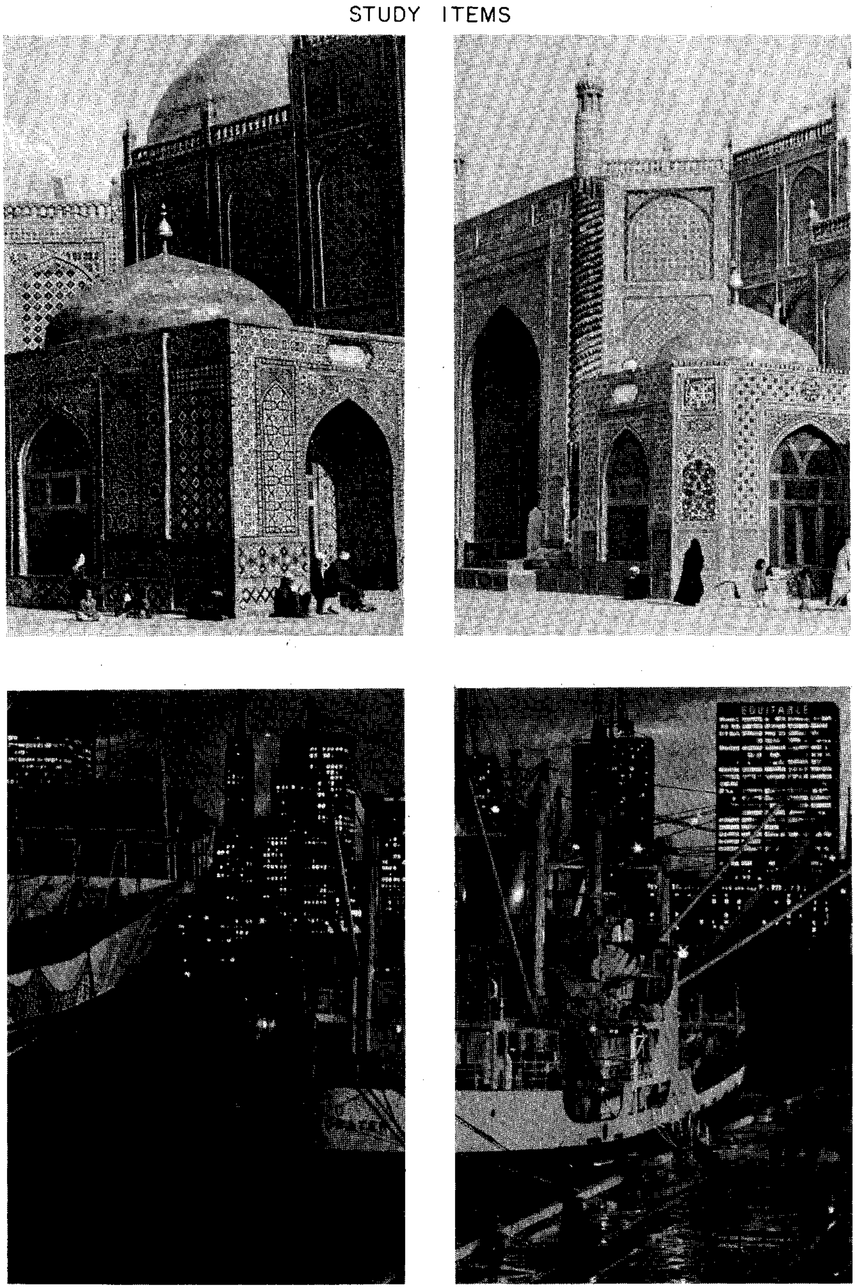


FIG. 1. Illustrative examples of picture pairs in the basic pool. One member of each pair would be shown in the study sequence, the same or the other member would appear in the recognition test.

the 48 pictures in the middle of the study sequence.

After the presentation of the whole sequence of 160 pictures, the subjects were given an interpolated task. It involved identification of famous people: Subjects had to write down what they knew about people whose names appeared on a test sheet. The interpolated task took approximately 7 minutes. During this interval the experimenter arranged the slides for the subsequent recognition test.



STUDY ITEMS

FIG. 1.—*Continued.*

Before the recognition test commenced, the experimenter gave another demonstration of the three different kinds of test pairs that were going to be used. Again there were six pairs of test items, involving pictures not seen by the subject, and accompanied by verbal descriptions of the same sort as used in the earlier demonstration.

In the recognition test, 36 pairs were presented in succession. One member of a pair was projected on one screen, and the other one on the other. The subject was al-

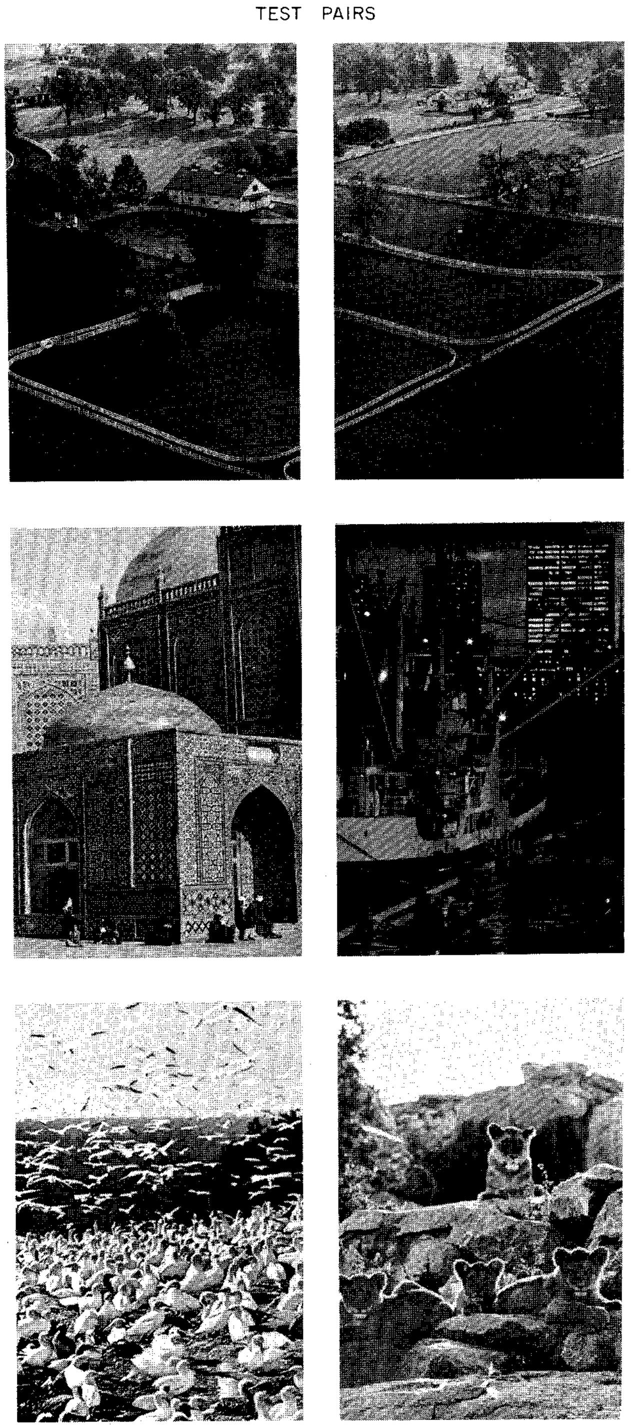


FIG. 2. Illustrative examples of *test* pairs of pictures in the three conditions in Experiment 1. The left member of each pair is the copy of one of the pictures seen in the study sequence (target), the right member is the distractor.

lowed 6 seconds to respond with "left" or "right" to indicate which of the two pictures he or she thought was the "old" one. The subject also gave a confidence judgment to accompany his or her decision. These confidence judgments were given on a 3-point scale, with 3 designating high confidence, and 1 representing guessing.

Results and Discussion

With each of the 48 subjects tested with 12 pairs in each of the three conditions there were 576 responses in each condition distributed between choices of targets (hits) and distractors (false positives) and among the six confidence judgment categories. The distributions of confidence judgments, together with hit rates and false positive rates, are shown in Table 3.

As can be seen in Table 3, hit rate was higher in condition A−X′ than in condition A−A′. Thus, when distractors are dissimilar not only to the targets within given test sets but also dissimilar to other items presented in the study sequence, people can discriminate between targets and distractors better than they can in a situation in which the distractor is similar to the target as well as to its episodic trace. This finding of an inverse relation between test-item similarity and recognition accuracy replicates many similar findings reported in the literature as indicated in the introduction. It represents the conventional test-item similarity effect and confirms the initial expectation.

Contrary to expectations, however, recognition performance also differed between conditions A−A′ and A−B′. The magnitude of the difference is not striking, but the difference itself seems to be reliable since it was obtained with each of six subgroups of eight subjects. These six subgroups differed from one another with respect to the study-sequence items (left or right halves of the two-page pictures) and with respect to the identity of the subset of the 24 two-page pictures from which the distractors in the A−X′ condition were drawn.

The higher hit rate in the perceptual similarity (A−A′) condition than the referred similarity (A−B′) condition means that subjects could better discriminate targets from their perceptually similar (A−A′) than their perceptually dissimilar (A−B′) distractors. Thus, here the test-item similarity effect is opposite to the one typically observed. The reasons for this reversal of the classical test-item similarity effect are not clear. The similarity relations between the distractors and the stored episodic information are identical in the two conditions. Even if we assumed that the perceptual similarity between the target and the distractor in a given test set plays no role in determining which of the two items the subject selects, and that the choices are completely determined by the similarity relations between test items on the one hand and the stored episodic information on the other, we would have to predict no dif-

TABLE 3
DISTRIBUTIONS OF CONFIDENCE JUDGMENTS AND HIT RATES AND FALSE POSITIVE RATES, EXPERIMENT 1

	Targets					Distractors				
	Confidence				Hit	Confidence				False
Condition	1	2	3	Mean	rate	1	2	3	Mean	positive rate
A−A′	124	117	186	2.15	.74	83	55	11	1.52	.26
A−B′	55	99	235	2.46	.68	49	69	69	2.11	.32
A−X′	67	146	287	2.44	.87	34	29	13	1.72	.13

ference for the hit rates in the two conditions. That, indeed, was the expectation for the outcome of this comparison before the experiment was done.

Given an unexpected and difficult-to-understand outcome of a single experiment, the most plausible conclusion is that the outcome is a fluke and could not be replicated. This conclusion was put to test in Experiment 2. Before that experiment is described, however, some other observations are in order about Experiment 1.

Distributions of confidence judgments are shown in Table 3 in full, because no generally accepted methods exist for summarizing and interpreting such data. Indeed, very few experiments have been reported in the literature in which confidence judgments are combined with forced-choice tests of recognition memory. Both Bower and Glass (1976) and Weaver and Stanny (1978), who did collect confidence judgments in forced-choice tests, reported a positive correlation between recognition accuracy and subjects' confidence in their judgments. Comparison of mean confidence ratings between conditions A−A′ and A−X′ replicates these findings inasmuch as the subjects were more confident of their choices in test pairs with dissimilar (A−X′) than similar (A−A′) distractors.

The same relation between test-item similarity and confidence was also found in the comparison between conditions A−A′ and A−B′. Here, too, confidence judgments were higher when test items were dissimilar (A−B′) than when they were similar (A−A′). But since the hit rate was higher in the A−A′ than in the A−B′ condition, the correlation between confidence judgments and recognition accuracy was negative with respect to these two conditions. Thus, although subjects made more errors in the A−B′ than in the A−A′ condition, they were more confident of being correct in the former condition than in the latter regardless of whether their choices were in fact correct or not.

Perceptual and Ecphoric Similarity

In preparation for Experiment 2, and as an afterthought to the design of Experiment 1, a further analysis of the results from conditions A−A′ and A−B′ was undertaken. It was prompted by reflections on what had been done and what had been found in the experiment.

Let us refer to the similarity between a test item and the stored relevant episodic information as *ecphoric similarity* of that item, and retain the term *perceptual similarity* to refer to the similarity between test items in a given set, such as a test pair. Thus, ecphoric similarity is a relation between an item physically present and information not physically present; perceptual similarity is a relation between two or more test items, both or all of which are physically present. These two kinds of similarity relations were varied in Experiment 1: Perceptual similarity was high in condition A−A′ and low in conditions A−B′ and A−X′; ecphoric similarity of distractors was high in conditions A−A′ and A−B′ and low in condition A−X′.

The purpose of the post hoc analysis was to study the relation between recognition performance and ecphoric similarity of distractors in conditions A−A′ and A−B′. Although in these conditions ecphoric similarity of distractors was high, it seemed reasonable to assume that some variability in the degree of ecphoric similarity of distractors existed among individual distractors used in the two conditions. This picture-to-picture variability in ecphoric

TABLE 4
Schematic Representation of the Post Hoc Design of Experiment 1

Perceptual similarity	Ecphoric similarity	
	High	Medium
High	A−A′	—
Medium	—	A−A″
Low	A−B′	A−B″

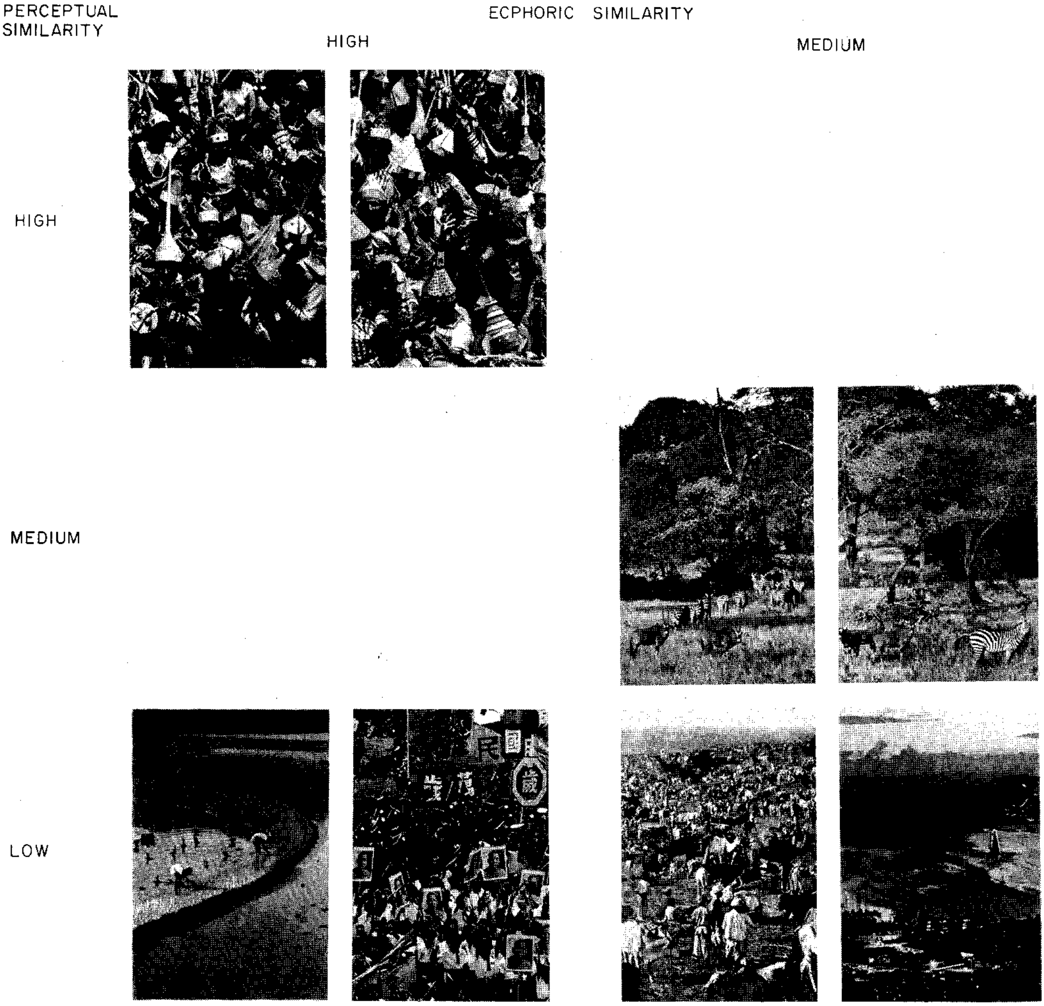


FIG. 3. Illustrative examples of *test* pairs of pictures in the four conditions of the post hoc design of Experiment 1, and the four conditions in Experiment 2. The left member of each pair is the copy of one of the pictures seen in the study sequence (target), the right member is the distractor.

similarity was measured and taken into account in the post hoc analysis.

The analysis entailed two steps. First, all 72 pairs of pictures in the basic pool, from which the critical study and test items were drawn, were rated for perceptual similarity. The rating was done by five judges, who had not participated as subjects in Experiment 1, on a 5-point scale. Second, on the basis of the mean similarity assigned to each pair, the *test pairs* of pictures in the A−A′ and A−B′ conditions were assigned to two equally large subsets. In one subset of test pairs the similarity of the distractors to their corresponding pair-mates was higher than the median mean rating, for the other half it was lower. This procedure resulted in four post hoc experimental conditions that are schematically depicted in Table 4. Illustrative test pairs corresponding to the design in Table 4 are shown in Figure 3.

In Table 4, high degrees of ecphoric similarity of distractor items are indicated by single primes (A′, B′), and medium degrees by double primes (A″, B″). The A–X′ condition of the original design would also fit into the post hoc design. The perceptual similarity of the two test items is the same as in the A–B′ and A–B″ conditions, and ecphoric similarity of the distractor is definitely lower than that in the A–A″ and A–B″ conditions. Thus, both perceptual and ecphoric similarity of the A–X′ condition would be designated as "low.") The perceptual similarity of the two test items in a given pair, in the new design, is still low for conditions A–B′ and A–B″. But the perceptual similarity of A–A′ test pairs by definition is higher than the perceptual similarity of A–A″ pairs, thus making necessary the distinction between "high" and "medium" degrees of perceptual similarity in Table 4 and Figure 3.

To summarize the post hoc design: Test pairs used in conditions A–A′ and A–B′ in the original design were classified into two categories of ecphoric similarity, labeled high and medium, on the basis of the rated similarity of distractor items to their corresponding, previously studied pairmates. This made it possible to look at recognition accuracy and confidence judgments as a function of not only perceptual similarity of targets and distractors within two-alternative test sets but also the ecphoric similarity between distractor items and the information stored from the study sequence.

The distributions of confidence judgments, together with hit and false positive rates, were analyzed according to the post hoc design and are tabulated in Table 5.

Compare first the A–A″ and A–B″ conditions. The hit rates, and consequently the false positive rates, are identical, and only the confidence judgments are slightly higher in the A–B″ condition. Thus, there is no evidence of the direct relation between test-item similarity and recognition accuracy when ecphoric similarity of distractors is not high. This means that this effect that was observed in the overall analysis of the data summarized in Table 3 must have been attributed to test pairs in which the ecphoric similarity of the distractor was high.

Comparison of conditions A–A′ and A–B′ in Table 5, indeed, shows that the initial effect is now greatly exaggerated. The hit rate for the high perceptually similar pairs was .71, whereas for the test pairs of low similarity it was only .58. The latter figure is only slightly higher than the hit rate of .50 expected by chance.

The inverse relation between the hit rate and the mean confidence judgments for the two conditions is also exaggerated when test pairs with only medium ecphoric similarity of distractors are eliminated from the analysis. Subjects made many more errors in the choice of the target in the A–B′ condition in the post hoc design, but they were much more confident that these erroneous choices were correct (mean confidence of

TABLE 5

DISTRIBUTIONS OF CONFIDENCE JUDGMENTS AND HIT AND FALSE POSITIVE RATES, ACCORDING TO THE POST HOC DESIGN OF EXPERIMENT 1

	Targets					Distractors				
	Confidence					Confidence				
Condition	1	2	3	Mean	Hit rate	1	2	3	Mean	False positive rate
A–A′	70	54	80	2.05	.71	45	32	7	1.55	.29
A–B′	26	47	93	2.40	.58	21	46	55	2.28	.42
A–A″	54	63	106	2.23	.77	38	23	4	1.48	.23
A–B″	29	52	142	2.51	.77	28	23	14	1.78	.23

2.28) than they were in selecting the distractors as targets in condition A–A′ (mean confidence of 1.55).

In summary of the results of the post hoc analysis, it looks as if the direct relation between test-item similarity and recognition accuracy holds only under conditions where the similarity between distractor items and the information stored is very high.

The results of the post hoc analysis do not eliminate the possibility that the whole pattern of results obtained in Experiment 1 was a historical accident. Experiment 2 was, therefore, undertaken to see whether the pattern of results summarized in Table 5 could be replicated in another experiment and under conditions where the two types of similarity manipulated in the post hoc design of Experiment 1 were systematically built into the experiment from the beginning. We will defer any further discussion of the results of Experiment 1 until after the question of the replicability of the results is settled.

Experiment 2

The design of Experiment 2 was patterned after the post hoc design of Experiment 1 as illustrated in Table 4. In terms of the symbols defined earlier, the design comprised four test conditions, designated as A–A′, A–A″, A–B′, and A–B″. The purpose of Experiment 2 was to test the replicability of the pattern of data observed in the post hoc design of Experiment 1. Would there be again a direct relation between test-item similarity and recognition accuracy for test pairs in which the ecphoric similarity of distractors is high, and would this relation be absent when the ecphoric similarity of distractors is less than high?

Condition A–X′ was not included in the design of Experiment 2. On the basis of the results of many previous experiments, replicated in Experiment 1, there was no doubt that recognition accuracy would be higher in condition A–X′ than in condition A–A′.

Method

The pictures used in Experiment 2 were drawn from a basic pool of 48 two-page pictures. Half of these had been used in Experiment 1, the other half were new. Each of the 48 pairs in the basic pool were rated for perceptual similarity by five judges on a 5-point scale on which 5 represented high and 1 represented low similarity. Agreement among the judges was good. When agreement was defined as no more than 2 points discrepancy, the T measure of interrater agreement (Tinsley & Weiss, 1975) for the 48 critical pairs was .867. (Within the adopted definition, the T value of 0 represents chance and 1.00 perfect agreement.) The similarity ratings of the four pairs shown in Figure 1 were 4.6, 3.4, 2.4, and 1.8, beginning with the top pair in Fig. 1 and ending with the one at the bottom. On the basis of the mean ratings, 24 pairs in the basic pool were assigned to the high-seimilarity subset, and 24 to the medium-similarity subset. The pictures in the high-similarity subset were used to make up A–A′ and A–B′ types of test pairs; pictures in the medium similarity subset were used to construct A–A′′ and A–B′′ types of test pairs. Thus, the similarity ratings provided by the subjects formed the basis for defining two degrees of ecphoric similarity of distractors to the stored episodic information. As in the post hoc design of Experiment 1, the three degrees of perceptual similarity were defined in terms of test items' belongingness to the same or different pairs: The target and distractor in each high- or medium-perceptual similarity pair (A–A′ and A–A″) represented two halves of one and the same two-page picture; the target and the distractor in each low-perceptual similarity pair (A–B′ and A–B″) came from different two-page pictures. The sample test pairs shown in Figure 3 concretely illustrate the four test conditions of Experiment 2.

Within each of the two subsets of 24 pairs of pictures—high and medium similarity—individual pictures served equally frequently as targets and as distractors; individual pictures in these subsets also appeared equally frequently in high- or medium-perceptual-similarity pairs, on the one hand, and the low-perceptual-similarity pairs on the other.

The procedure was very much the same as in Experiment 1. The subjects saw a series of 180 individual pictures, each presented for 2 seconds. The critical 48 pictures appeared in the middle of the sequence, with 66 buffer pictures preceding and 66 following them. After seeing the sequence of 180 study pictures, the subjects were given the recognition test consisting of 32 pairs, with 8 pairs representing each of the four test conditions. In each pair, subjects chose one of the items as the one that they had seen previously, and rated their confidence of being correct on a 3-point scale, with 3 representing high confidence and 1 representing guessing. Unlike Experiment 1, there was no interpolated task between the study sequence and the recognition test.

Subjects were 18 undergraduate students of both sexes at the University of Toronto who signed up for the experiment for modest remuneration. They were tested individually. As in Experiment 1 they were given thorough instructions about the task and the types of test pairs that they were going to encounter in the test.

Results

The distributions of confidence judgments, together with hit and false positive rates are shown in Table 6. The overall pattern of data is quite similar to that observed in the post hoc design of Experiment 1 (Table 5). Again, there is little difference between the hit rates in the A−A″ and A−B″ conditions. The small numerical superiority in the hit rate for test pairs of low perceptual similarity (condition A−B″) over the condition of medium perceptual similarity (condition A−A″) is not statistically significant. Replicating the results of Experiment 1, this finding shows that there is no evidence of the reversal of the typical inverse relation between test-item similarity and recognition accuracy when ecphoric similarity of distractors is not very high.

Comparison of the hit rates in conditions A−A′ and A−B′ also produces a replication of the corresponding finding in Experiment 1: With test pairs whose distractors were characterized by high degrees of ecphoric similarity to the stored information, subjects were capable of discriminating targets from perceptually similar distractors better than from perceptually dissimilar distractors. It seems, therefore, that the somewhat surprising reversal of the typical test-item similarity effect in Experiment 1 was not a fluke. It seems to represent a readily replicable empirical fact.

The distributions of confidence judgments in the four experimental conditions in Experiment 2 also closely resemble the corresponding distributions in the post hoc design of Experiment 1. The product-moment correlation between the 24 entries in Table 6 and the corresponding entries in Table 5 is +.97. As in Experiment 1, in Experiment 2, too, subjects exhibited greater confidence in their recognition judgments with perceptually dissimilar test alternatives (condition A−B′) than with perceptually similar alternatives (condition A−A′). Thus, the negative correlation between recognition accuracy and confidence between these two experimental conditions (A−A′ and A−B′) that was observed in Experiment 1 was also obtained in Experiment 2.

General Discussion

The two experiments described here have shown that recognition accuracy in a two-alternative forced-choice task is higher with similar than dissimilar distractors under certain conditions. This result represents a reversal of the typical inverse relation between recognition accuracy and similarity of test items.

This reversal of the typical test-item similarity effect occurs under conditions where

TABLE 6
DISTRIBUTIONS OF CONFIDENCE JUDGMENTS AND HIT RATES AND FALSE POSITIVE RATES, EXPERIMENT 2

	Targets					Distractors				
	Confidence					Confidence				
Condition	1	2	3	Mean	Hit rate	1	2	3	Mean	False positive rate
A–A′	37	31	33	1.96	.70	24	13	6	1.58	.30
A–B′	14	28	39	2.31	56	12	27	24	2.19	.44
A–A″	32	24	57	2.22	.78	17	10	4	1.58	.22
A–B″	18	29	73	2.46	.83	12	10	2	1.58	.17

the distractor items in test pairs are very similar to the items seen earlier in the study sequence but which do not appear as target items in the recognition test. Both the post hoc designs of Experiment 1 and Experiment 2 showed that similarity of the distractors to the information stored from the study episode must be very high for the reversal of the typical test-item similarity effect to occur. This requirement was met in conditions A–A′ and A–B′. Under these conditions, the magnitude of the novel effect of a direct relation between test-item similarity and hit rate can be quite large. In Experiment 2, for instance, the hit rate for perceptually similar pairs was .70, whereas for the perceptually dissimilar pairs it was .56. When these figures are corrected for guessing, using the standard high-threshold method, they become .40 and .12, respectively, representing a better than 3:1 ratio in favor of perceptually similar distractors.

The effect was eliminated with lower degrees of similarity between distractors and stored information, represented by conditions A–A″ and A–B″ in the post hoc design of Experiment 1 and Experiment 2. In Experiment 1 there was no difference in recognition accuracy between conditions A–A″ and A–B″, and in Experiment 2 the hit rate was numerically, although not reliably, higher for test pairs with dissimilar distractors.

The present results in no way question the facts about test-item similarity and recognition memory as they have been reported in the literature so far. These facts are real. The typical inverse relation between test-item similarity and recognition accuracy was replicated in Experiment 1 (condition A–A′ versus condition A–X′). The main point of the present experiments is the demonstration that exceptions do exist to the facts as they have been hitherto known.

To understand the exception described herein, it is necessary to distinguish between two kinds of similarity relations in multiple-choice recognition tests. One has to do with the *perceptual similarity* of the physically present test items within a given set, in these experiments the target and distractor of a pair. This similarity was labeled as "high," "medium," or "low," on the basis of (a) the belongingness of the two test pictures to the same or different "two-page pictures," and (b) similarity ratings assigned to the two halves of the same two-page picture. The fact that recognition accuracy systematically covaried with similarity thus defined can be regarded as a validation of these measurement operations.

The other similarity relation was labeled *ecphoric similarity*. "Ecphory" is a term used by Richard Semon (1909; see also Schacter, Eich, & Tulving, 1978) to refer to the actualization of a latent engram by an "ecphoric stimulus." In keeping with Semon's idea, I am using the term "ecphoric similarity" to refer to the similarity (or compatibility) between an "ecphoric stimulus'' (a retrieval cue, a recognition test item) and the stored information. Thus, while perceptual similarity is involved in comparisons of objects when they are physically present, ecphoric similarity involves comparisons of objects when one of them is not physically present (Shepard & Podgorny, 1978).

In the experiments described here, two degrees of ecphoric similarity of distractors—high and medium—were distinguished in terms of judges' ratings of perceptual similarity of two halves of corresponding two-page pictures. The underlying assumption is that ecphoric similarity between two objects is correlated with perceptual similarity between the same objects. Again, the fact that recognition accuracy systematically covaried with ecphoric similarity lends some validity to this assumption.

Given the two kinds of similarity relations, we can now express the main conclusion from the two experiments as follows: Performance in a multiple-choice picture-recognition test depends on the interaction between perceptual similarity and ecphoric similarity of the test items. Both similarity relations must be specified in the description of the results of relevant experiments in which similarity effects are of interest.

In previous experiments in which the typical inverse relation between test-item similarity and recognition accuracy has been demonstrated, the two kinds of similarity relations have been confounded: High degrees of perceptual similarity have implied high degrees of ecphoric similarity, and low degrees of perceptual similarity have been accompanied by low degrees of ecphoric similarity. As argued and demonstrated in this paper, the two kinds of similarity relations can be at least partly dissociated, and their separate effects on recognition judgments studied.

A finding of secondary interest from the two experiments concerned confidence judgments. Usually there is a high positive correlation between recognition accuracy and confidence judgments (e.g., Bower & Glass, 1976; Tulving & Thomson, 1971; Underwood & Freund, 1968; Weaver & Stanny, 1978). But both in the post hoc designs of Experiment 1 and in Experiment 2, in conditions characterized by high degrees of ecphoric similarity (A–A′ and A–B′), subjects' accuracy in discriminating targets from distractors was higher with perceptually similar pairs, whereas their confidence was greater with perceptually dissimilar pairs. This negative correlation between accuracy and confidence across two experimental conditions, too, constitutes an exception to previous results of wide generality. The reversal of the typical finding, of course, presumably only reflects the atypical finding with respect to recognition accuracy. The confidence judgments given by the subjects in the present experiments followed the pattern of all previous experiments in that they were higher with perceptually dissimilar test pairs than with perceptually similar ones. Nevertheless, the finding does demonstrate that recognition accuracy and confidence judgments can be negatively correlated.

How do we explain the reversal of the classical test-item similarity effect? Obviously much more empirical evidence is needed about the generality and boundary conditions of the present findings before informed opinions can be offered. By way of pure speculation, however, some comments can be made even now.

One possible reason for the superiority of recognition performance in the A–A′ condition over the A–B′ condition may have to do with the number of objects to which test items refer. In the A–A′ condition, both test items—the target and the distractor—refer to one and the same memory trace, whereas in the A–B′ condition the two test items must be matched to different traces. One might argue, therefore, that the two similar test items provide more efficient access to their corresponding trace than do two different test items to their two corresponding traces, and that these differences in trace access are reflected in recognition performance.

A related line of reasoning might hold

that the common reference trace makes it possible for the subject to disregard features shared by the two test items. Instead, the subject would concentrate on the features that distinguish a target and a distractor, comparing or matching these distinguishing features to those of the trace, and choosing the test item that provides the better match. When the test items are dissimilar, and refer to different traces, such selective attention to particular features is not possible, with a consequent more difficult comparison of two larger sets of features that match or do not match.

This type of speculation about the processes underlying the findings of the two experiments is not entirely unreasonable. We know that psychophysical judgments, for instance, magnitude estimations, are easier along a single dimension within a given sensory modality than they are between different dimensions in a modality, or different dimensions in different modalities. Similarly, we know that judgments of similarity between multidimensional perceptual or conceptual objects are easier in situations in which two different objects are compared to a common third reference object than they are in situations in which two objects are each compared with different reference objects. Tversky (1977) has provided a tightly reasoned theoretical account of findings of this sort; the same analysis could presumably be extended to judgments of similarity in situations in which one of the objects is not physically present (Shepard & Podgorny, 1978).

The major problem for the speculations of the sort just mentioned lies in the handling of the interactive effects of the two kinds of similarity, as evidenced by the pattern of data in Tables 5 and 6. In condition A–A″, too, there is only a single reference trace, and the two test items are sufficiently similar for the subject to be able to ignore shared features and concentrate on the distinguishing ones. But recognition performance in this condition is no higher than it is in condition A–B″ in which test items refer to different stored traces.

Because of these difficulties, a somewhat more promising idea might be that the basic finding of the experiments represents a consequence of a strategic decision on the part of the subjects. Highly similar test items may induce subjects to engage in deeper or more elaborate processing of retrieval information (Craik & Jacoby, 1979), or examine the relevant evidence more thoroughly (Bower, 1972, p. 98). An additional assumption here would have to be that it is only very high degrees of perceptual similarity of test items that bring about such strategic elaborative retrieval processes. With lower degrees of perceptual similarity, the probability that the distractor is (incorrectly) selected may be mainly determined by its ecphoric similarity.

Finally, what about the data that prompted the research described in this paper, the data summarized in Table 1? Is the hint of superior recognition with semantically related test items a faint analogue of the reversal of the classical test-item similarity effect observed in the picture-recognition experiments described here, or are these data comparable to those obtained in the A–A″ and A–B″ conditions? At the present, it is difficult to tell. We need more evidence on similarity relations in recognition memory. But, given the findings reported here and their logical implications for the relevance of two kinds of similarity relations, the original results seem considerably less baffling now.

References

Anisfeld, M., & Knapp, M. Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology,* 1968, **77**, 171–179.

Bahrick, H. R., Clark, S., & Bahrick, P. Generalization gradients as indicants of learning and retention of a recognition task. *Journal of Experimental Psychology,* 1967, **75**, 464–471.

Bower, G. H. Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory.* Washington, D.C.: Winston, 1972.

Bower, G. H., & Glass, A. L. Structural units and the redintegrative power of picture fragments. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, **2**, 456–466.

Craik, F. I. M., & Jacoby, L. L. Elaboration and distinctiveness in episodic memory. In L.-G. Nilsson (Ed.), *Perspectives on memory research.* Hillsdale, N.J.: Erlbaum, 1979.

Glass, A. L., Holyoak, K. J., & Santa, J. L. *Cognition.* Reading, Mass.: Addison–Wesley, 1979.

Jenkins, J. J. 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.* Hillsdale, N.J.: Erlbaum, 1979.

Jörg, S., & Hörmann, H. The influence of general and specific verbal labels on the recognition of labeled and unlabeled parts of pictures. *Journal of Verbal Learning and Verbal Behavior,* 1978, **17,** 445–454.

Kintsch, W. *Learning, memory, and conceptual processes.* New York: Wiley, 1970.

Klein, L. S., & Arbuckle, T. Y. Response latency and task difficulty in recognition memory. *Journal of Verbal Learning and Verbal Behavior,* 1970, **9,** 467–472.

McNulty, J. A. An analysis of recall and recognition processes in verbal learning. *Journal of Verbal Learning and Verbal Behavior,* 1965, **4,** 430–436.

Nagae, S. Nature of discriminating and categorizing functions of verbal labels on recognition memory for shape. *Journal of Experimental Psychology: Human Learning and Memory,* 1980, **6,** 421–429.

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.

Schacter, D. L., Eich, J. E., & Tulving, E. Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior,* 1978, **17,** 721–743.

Semon, R. *Die mnemischen Empfindungen.* Leipzig: Wilhelm Engelmann, 1909.

Shepard, R. N., & Chang, J.-J. Forced-choice tests of recognition memory under steady-state conditions. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 93–101.

Shepard, R. N., & Podgorny, P. Cognitive processes that resemble perceptual processes. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes.* Hillsdale, N.J.: Erlbaum, 1978. Vol. 5.

Tinsley, H. E. A., & Weiss, D. J. Interrater reliability and agreement of subjective judgments. *Journal of Counselling Psychology,* 1975, **22,** 358–376.

Tulving, E. Relation between encoding specificity and levels of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing and human memory.* Hillsdale, N.J.: Erlbaum, 1979.

Tulving, E., & Thomson, D. M. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology,* 1971, **87,** 116–124.

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

Tversky, A. Features of similarity. *Psychological Review,* 1977, **84,** 327–352.

Underwood, B. J. False recognition produced by implicit verbal response. *Journal of Experimental Psychology,* 1965, **70,** 122–129.

Underwood, B. J., & Freund, J. S. Errors in recognition learning and retention. *Journal of Experimental Psychology,* 1968, **78,** 55–63.

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

Weaver, G. E., & Stanny, C. J. Short-term retention of pictorial stimuli as assessed by a probe recognition technique. *Journal of Experimental Psychology: Human Learning and Memory,* 1978, **4,** 55–65.

Wickelgren, W. A. *Learning and memory.* Englewood Cliffs, N.J.: Prentice–Hall, 1977.

Wiseman, S., & Tulving, E. Encoding specificity: Relation between recall superiority and recognition failure. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, **2,** 349–361.

Woodworth, R. S. *Experimental psychology.* New York: Holt, 1938.

Wyant, S., Banks, W. P., Berger, D., & Wright, P. W. Verbal and pictorial similarity in recognition of pictures. *Perception & Psychophysics,* 1972, **12,** 151–153.

(Received January 5, 1981)