

On the Very Long-Term Retention of Knowledge Acquired Through Formal Education: Twelve Years of Cognitive Psychology

Martin A. Conway
Lancaster University
Lancaster, Lancashire, England

Gillian Cohen and Nicola Stanhope
The Open University
Milton Keynes, England

Former students ($N = 373$) of a course in cognitive psychology (CP), conducted between 1978 and 1989, completed memory tests to assess retention of CP. Memory for proper names of researchers, concepts, and conceptual relations varied with retention interval (RI), and memory performance declined over the first 36 months of retention and then stabilized at above-chance levels for the remainder of the retention period. Memory for general facts from the course and research methods did not, however, vary with RI and remained at the same above-chance level across all RIs sampled. The recall and recognition of proper names showed a more rapid decline than the recall and recognition of concepts. These findings suggest that knowledge structures formed at acquisition mediate the very long-term retention of CP. Also, Ss with higher grades retained more knowledge than Ss with lower grades. Finally, a dissociation between memory performance and confidence ratings indicates that at the longer RIs, Ss were unaware that course material was accurately remembered.

During the process of formal education many topics are learned in great detail, some even to expert levels, but subsequently many of these knowledge domains are utilized only rarely or remain wholly unused. What is the fate of such knowledge? Is it rapidly forgotten? Is it selectively forgotten with, for instance, details being lost first and more general aspects being retained longer? Does initial depth of learning determine the period of retention? And do such factors as motivation or interest in a knowledge domain significantly influence retention? Finally, is retention affected by the process of aging? For example, after a long retention period, does a person's age at retrieval influence memory performance? The cross-sectional study reported in this article addresses these questions by investigating the retention of cognitive psychology acquired by university students who took a course in cognitive psychology in one of the years between 1978 and 1989 (inclusive). However, before considering the study in detail, we first review the seminal work of Bahrck (e.g., 1979, 1983, 1984) that has already begun to provide answers to some of the aforementioned questions and that provided the impetus for the present research.

Very Long-Term Retention of Knowledge

In their research, Bahrck and his colleagues investigated the very long-term retention of various types of knowledge acquired outside laboratory settings. For example, Bahrck,

Bahrck, and Wittlinger (1975) examined the retention of the names and faces of high school classmates over a 50-year retention interval (RI). Bahrck (1983) reported a study of the very long-term retention of the spatial layout of a city; Bahrck (1984) examined the retention of Spanish, originally learned at high school, also over a 50-year RI, and Bahrck and Phelps (1987) investigated the effects of different types of learning schedule on the retention of Spanish-English word pairs. The studies by Bahrck et al. (1975) and Bahrck (1984) are most relevant to the present research, so we review these in some detail.

Bahrck et al. (1975) investigated the abilities of 392 high school graduates to recognize the names and faces of classmates, to match names to faces and faces to names, to free recall names, and to generate names from pictures of classmates. Subjects were allocated to nine time groups, according to time elapsed since graduation: For the shortest time group, the RI was 3.26 months, and for the longest group, it was 570.73 months. Thus, the total RI sampled spanned approximately 47 years. The principal findings were that retention, as assessed by free recall and picture cuing, showed a steady decline with increasing RI and an accelerated final decline at the longest RI. In contrast, the other measures of retention, name and face recognition and name-face matching, were 90% correct over the first 15 years of retention and then showed a marked decline over the remaining period of retention. This decline appeared somewhat later for face recognition, occurring after 35 years of retention. Interestingly, multiple regression analyses generally found no systematic relations between the measures of retention and various predictor variables, such as class size, recency of review of yearbook, and so on (see Bahrck et al., 1975, Table 5), although the recognition and matching tests were associated with class size.

Bahrck et al. (1975) argued that these long-lasting, high levels of retention reflect the effects of massive overlearning on memory and, possibly, are reinforced by long periods of naturally-occurring, spaced practice (e.g., Landauer & Bjork,

This research was supported by Economic and Social Research Council of Great Britain Grant R000231158.

All of us contributed equally to the study. We thank Philip Levy for considerable help with the analyses reported in the article. We also thank Harry P. Bahrck and two anonymous reviewers for their helpful comments on the article.

Correspondence concerning this article should be addressed to Martin A. Conway, Department of Psychology, Lancaster University, Lancaster LA1 4YF England.

1978) in memorizing names and faces. The differences between recall and recognition mirror differences between recall and recognition found in laboratory studies of memory (e.g., Jones, 1978; Kintsch, 1978; Mandler, 1980; Tulving & Thomson, 1973). Finally, the marked decline in retention at the 47-year interval occurs when subjects are in their late sixties and therefore could reflect an age-related memory impairment, rather than an effect of RI per se.

Bahrck (1984) reported a similar study, although in this case the knowledge domain sampled was that of Spanish learned in high school. The study spanned a 50-year RI, and measures of initial learning were available in the form of grades for Spanish courses and number of courses taken. Once again, strikingly high levels of retention were observed, and subjects who had achieved high grades and who had studied more courses reached levels of 80% on some of the retention tests after a 50-year RI. Differences between recall and recognition paralleling those found in the earlier study were observed, and recognition tests showed higher levels of retention than recall tests. Measures of rehearsal indicated that subjects had, in general, not rehearsed their knowledge of Spanish during the retention period. Hence, retention was predicted by initial depth of learning, with higher levels of training being systematically related to higher levels of retention.

However, the most interesting finding from the Bahrck (1984) study was the discontinuous nature of the retention curves that contained three distinct segments. It was found that retention declined exponentially for the first 6 years of retention and then stabilized for periods of up to 35 years before showing a final decline (see Bahrck, 1984, Figure 4). Bahrck proposes that the interval from 0 to 6 years represents a period of rapidly accelerating forgetting. However, the interval from 6 to 35 years represents a period of stability in retention that Bahrck refers to as a *permastore*. The interval from 35 years onwards shows a further and final decline in retention.

Bahrck's (1984) interpretation of the permastore portion of the retention curve is particularly interesting. He argued that learning is a discontinuous process in which knowledge must reach some criterion depth of learning over time, before being stored in such a way as to be comparatively resistant to forgetting. Other evidence supplied by Bahrck (1984, p. 24) suggests that neither frequency of exposure to a knowledge domain nor amount of interference from related knowledge domains are sufficient to account for either acquisition or forgetting. Rather, the cumulative experience of many spaced-learning episodes leads to a transition of knowledge already in memory into a more permanent and stable form.

The concept of a *permastore* has, however, been criticized by Neisser (1984), who put forward an alternative theoretical account. Neisser argued that subjects retain some schematic representation of the knowledge domain or parts of the domain, rather than storing specific knowledge in a quasi-permanent form. These schematic knowledge structures facilitate the reconstruction of specific knowledge and, in so doing, give the appearance that detailed knowledge has been literally retained. For instance, schemata might be used to generate above-chance guesses on recognition tests and so provide

evidence of retention of specific knowledge that has not in fact been retained. Moreover, by this account, the exponential decline in the early years of retention can be explained by assuming that knowledge that is not well integrated with existing schemata is rapidly lost until only the schematic structures themselves remain. As Neisser pointed out, this early rapid decline in retention will therefore depend on how well a knowledge domain lends itself to schematic representation. Consonant with this view, Bahrck (1983) found comparatively rapid forgetting of street names for a city in which people had briefly lived. Presumably, street names constitute a knowledge domain that is difficult to structure and represent abstractly in the form of a schema.

The purpose of the present study is to extend Bahrck's (1979) cross-sectional methodology to a new knowledge domain and explicitly contrast memory for conceptual knowledge with memory for more specific knowledge. If Neisser's (1984) schema theory interpretation is correct, then different types of knowledge should yield different retention functions. For instance, retention of conceptual knowledge from a highly structured knowledge domain, such as theories and models within cognitive psychology, should show less rapid forgetting than more detailed, less schema-central, domain knowledge (e.g., knowledge of names of researchers within cognitive psychology). Furthermore, to the extent that there are overall structural differences between the nature of the knowledge featured in the present study and that sampled by Bahrck, then additional differences in patterns of retention should be apparent. Nevertheless, we expect to obtain support for Bahrck's observations of differences between recall and recognition and for the influential role of original depth of learning.

Retention of Cognitive Psychology

The knowledge domain sampled is that of cognitive psychology, which was taught as a 1-year course in the final year of a part time degree. Samples of students from each of the 12 years from 1978 to 1989 (inclusive) took part in the study. The students were all Open University students, and full details of the course and subject sample are provided later. Here, we simply mention some unique aspects of this population.

Open University courses are highly structured with specially written texts for each part of a course, and these are presented in such a way as to highlight key concepts and research findings. Perhaps the key advantage of this subject sample over other populations (e.g., full time undergraduates) is that Open University students are dependent on their course material and rarely have the opportunity to draw on other sources of information concerning cognitive psychology (e.g., journals). Thus, the main sources of knowledge of cognitive psychology used by our subjects can be easily identified and, importantly, these do not change from year to year. Furthermore, all students complete the same regular coursework and final examination. Course grade is equally weighted for coursework and examination. Finally, all students are "mature students" (i.e., over the age of 21); typically, in any given year, about 25% of the students are older than 45 years, and consequently, some were 65 years and older at the time of

the present study. Thus, in contrast to Bahrick's studies in which age and RI were confounded, the age distribution of our subject sample allows us to examine the independent contributions of these factors to retention.

Given the highly structured nature of the course material, it was possible to construct retention tests that we could be certain featured knowledge that subjects had originally acquired. Separate recognition tests for proper names and for concepts from the course were constructed and paired with cued-recall tests for names and concepts. Note, however, that the recognition and recall tests sampled different names and concepts and were only similar in the sense that both types of tests sampled the domains of proper names and concepts. Schema theory, as outlined earlier, suggests that concepts will be embedded in an organized system of schemata, whereas names, which lack rich semantic associations (Cohen, 1990), may be less well integrated with schematic knowledge structures. Thus, according to Neisser's (1984) view, retention of names may show a more rapid decline than retention of concepts.

Further tests were devised to assess the retention of conceptual knowledge. Two fact verification tests were constructed in which subjects were required to make true-false judgments about theoretical and empirical statements. In one test, the statements were selected so as to refer to general, high-level, theoretical concepts or principles, and in the other test, the facts referred to specific findings or lower level details. Neisser's (1984) theory predicted that general knowledge should be better retained than specific knowledge, and indeed, Navah-Benjamin (1988) found that conceptual relations between specific concepts were lost more rapidly than conceptual relations between general concepts. We also devised a concept grouping test to examine that time course for the retention of conceptual relations.

Finally, we developed a multiple-choice recognition test to examine retention of knowledge of research methods. Research methods were formally taught and used extensively in practical work. The study of retention of research methods provided an opportunity to examine memory for a different type of knowledge: one that had been acquired in part by passive learning methods (e.g., reading and writing reports) and in part by active learning methods (designing and conducting experiments).

Method

Course

The course in cognitive psychology was divided into 6 topic areas: memory, perception, language, problem solving, research methods, and artificial intelligence. This latter topic area was not sampled in the present study. All areas received equal coverage, and coursework sampled each of the first 4 areas and included a number of experimental research projects sampling the 5th area. The original course ran between 1978 and 1985, and an updated version was introduced in 1986, which included coverage of more recent research. Thus, the memory tests we developed for the original course were slightly modified to take into account new material covered in the updated course.¹

Subjects

Three hundred and seventy-three former students of cognitive psychology took part in the study. Approximately half were volunteers responding to personal contact or to advertisements in a student newspaper. The remainder were traced through student records and contacted by mail (response rate 44%) or by telephone (response rate 90%). The subject sample constituted approximately 40% of the total population of former cognitive psychology students. Table 1 shows the number of subjects from each year of the course, age at retrieval, grade obtained, and subsequent contact with course-related material.

Design and Procedure

The main independent variable was RI, which comprised 12 separate groups. The RIs in months for each group are shown in Table 1. Note that RI is measured in months from date of completion of the course, marked by the date of the final examination that always occurred in the October of each year of the course, to time when retention was tested. Also, note that the spacing of RIs is not equal, and this is because testing was conducted in two waves some months apart. The main dependent measure was performance on the different memory tests, and confidence ratings (CRs) for each measure of memory performance were also collected. Various secondary measures were course grade, subsequent contact with course-related material, ratings of interest in the various course topics, and age at retrieval. These secondary measures were assessed separately in a questionnaire.

Questionnaire

All subjects completed the questionnaire before taking the memory tests. In the questionnaire, subjects stated (a) the year they took the course, (b) their age when taking the course, (c) whether they took the exam over (there were no positive answers to this question), and (d) the grade they achieved. There were five grades: A 1 was equivalent to a first-class degree, a 2 to an upper second-class degree, a 3 to a second-class degree, a 4 to a third-class degree, and a 5 equivalent to a pass grade. Only 1 of our subjects had a 5 grade, and the majority had 2, 3, and 4 grades (see Table 1). Subjects then rated how interesting they had found each of the course topic areas. Interest ratings (IRATs) were on a 3-point scale in which a 1 indicated *not interesting*, a 2 indicated *fairly interesting*, and a 3 indicated *very interesting*. (Note that interest ratings were not collected for research methods.)

Finally, subjects answered a series of questions designed to probe the extent of postcourse contact with psychology, in general, and cognitive psychology and research methods, in particular. The distribution of subjects' responses to these questions divided into two identifiable classes that were coded 1 and 2. A 1 indicated a considerable amount of postcourse contact, including study of further psychology courses that did not feature cognitive psychology, working in a profession (such as personnel, social work, education) that had some relation to psychology, or both. A 2 indicated little or no further contact with cognitive psychology. In this case, a *little further contact*

¹ The full set of course units for the first version of the course are available from The Open University Press, Buckingham, England, and are coded *D303 Cognitive Psychology*. The main units for the updated course are available in Cohen, Eysenck, and Le Voi (1986), Greene (1986), Kahney (1986), and Roth and Frisby (1986). Information concerning additional course material relating to the updated course can be obtained from us on request.

Table 1
 Subject Totals and Means by Retention Interval (RI) for Age, Grade, and Contact

RI in months	Subject totals		Age at retrieval		Grade		Contact	
	T	%	M	SD	M	SD	M	SD
3	33	8.8	39.6	8.1	2.5	.834	1.88	.331
15	37	9.9	39.8	8.9	2.5	.650	1.65	.484
27	35	9.4	45.5	10.1	2.7	1.01	1.57	.502
39	25	6.7	46.7	11.1	2.8	.913	1.60	.500
41	42	11.3	46.8	9.4	2.8	.881	1.69	.468
53	48	12.9	48.5	10.1	2.5	.849	1.60	.494
65	28	7.5	52.9	8.7	2.8	.803	1.68	.476
77	27	7.2	54.4	11.6	3.0	.898	1.78	.424
89	27	7.2	55.7	10.0	2.9	.759	1.63	.492
101	23	6.2	52.6	10.7	2.4	.988	1.57	.507
113	18	4.8	53.3	11.9	2.8	.707	1.89	.323
125	30	8.1	58.9	9.6	2.4	.817	1.63	.490

Note. T = total; M = mean; and SD = standard deviation. Contact refers to ratings of contact with course material with the exception of research methods (see text).

consisted of an occasional television program, magazine article, or conversation that related to cognitive psychology. Subjects who had directly pursued careers or further studies in cognitive psychology were not used, so overall, the amount of subsequent contact that subjects had with this very specialized material was very low.

Extent of postcourse contact with research methods, and in particular with statistics, was assessed separately by a similar question. We reasoned that a student, although having no direct postcourse contact with cognitive psychology, might nonetheless have some detailed contact with research methods and statistics. In fact, the pattern of mean contact ratings by year for research methods was similar to the pattern for contact with the topic areas.

To verify accuracy of responding to the course questionnaire, a subset of 20% of responses was checked against central student records for grade achieved and age of subject when taking the course. No discrepancies were found.

Tests

Five different tests of retention were used: recognition of proper names and concepts, fact verification of specific and general facts, cued recall of proper names and concepts, concept groupings, and a recognition test of memory for research methods. Separate versions of each of the first 4 tests were constructed in each of the four topic areas of perception, memory, language, and problem solving. All items used in the tests featured prominently in the course material.

All the tests were extensively piloted on former students, and the comments and performance of pilot subjects were used to modify the tests so that the difficulty level and time required to complete the tests were acceptable to the subjects. In addition to this, a further group of former Open University students who had not studied cognitive psychology and who were similar in age range to our subject sample were recruited to act as a control group. The control group completed all the tests and were encouraged to make plausible guesses on the tests that required discrimination between items and to guess in response to the cues on the cued-recall tests. Across all tests, control subject performance never reliably exceeded chance levels, and in the case of cued recall, the chance level was determined, on the basis of control subjects' responses, as being zero. The data from the control group are not reported further, and in the Results section, memory performance is compared with chance levels.

Test 1: Name and concept recognition. The recognition tests assessed recognition of proper names and concepts from the course

material. Foils for names were created by searching telephone directories for names. Foils for non-United Kingdom target names were generated by the experimenters, and these were plausible real surnames for the appropriate nationality. All foil names were selected so that they were the same length and contained the same initial letter as the target name. Concept foils were only approximately the same length as targets. These foils named concepts that had not been featured in the course and, as far as we could determine, were completely novel (e.g., *chronometric cycle*). Note, however, that concept foils (e.g., *eigen stimulus test*, *vector cells*, *quadratic activation effect*) were designed to be stylistically similar to the targets (e.g., *embedded figures test*, *texture cues*, *preattentive processing*). Plausibility of the foils was assessed in the pilot studies, and foils that were identified as implausible were replaced. Thus, a two-part recognition test was constructed; the first part contained 16 target concepts and 16 foil concepts, and the second part contained 16 target names and 16 foil names.

Subjects always completed the concept recognition test first, followed by the name recognition test. Within each test, the items were randomly ordered and listed in a single column. Next to each item were printed the words *OLD/NEW* followed by the confidence ratings (3, 2, and 1). Subjects were told that approximately half the items had appeared in the course, whereas the remainder originated from other sources. They were asked to circle *OLD* if they thought the item had appeared in the course material, and *NEW* if it had not. They were asked to indicate how confident they were that their answer was correct. A rating of 3 indicated *certain*; 2, *fairly sure*; and 1, *guessing*.

Test 2: Fact verification. Selection of stimuli for the specific and general fact verification tests followed a similar procedure. In this case, specific facts were facts that referred to details of specific theories and findings highlighted in the course (e.g., *SHRDLU can discuss a very limited world in considerable depth*). General facts referred to more global aspects of theory (e.g., *The behaviorists considered learning to be a passive process*). Twelve specific facts and 12 general facts were selected from each of the four topic areas of the course.

Each true fact was then paired with a false fact. We generated false facts by modifying or by replacing one or two concepts in a true fact to produce a plausible but false assertion (e.g., *The behaviorists considered learning to be an active process*, *PARRY can discuss a very limited world in considerable depth*). For each of the topic areas, 12 true and 12 false facts were selected, with the constraint that specific and general facts were represented equally. A further con-

straint was that each set of facts did not contain corresponding true and false facts (e.g., *The behaviorists considered learning to be an active process, The behaviorists considered learning to be a passive process*). The facts were listed in an unsystematic order, and printed next to each fact were the letters *T* and *F*, followed by the 3-point confidence rating scale. Subjects were instructed to circle *T* if they judged the statement to be true and *F* if they judged it to be false. Confidence in each judgment was registered by circling the appropriate number of the CR scale.

Test 3: Concept grouping. The concept-grouping task required subjects to sort 24 concepts into six groups. One concept in each group was provided as a heading, and subjects selected concepts from the remaining list of 18 that were listed in an unsystematic order on the test sheet. Subjects were not told that concepts divided equally into six groups of 4 concepts. The concept groups had been explicitly emphasized in the course material. The concept that acted as a group heading was selected on the basis that the concept accurately reflected that conceptual grouping. So, for example, in the area of memory, *Baddeley, working memory, mental arithmetic, and articulatory loop* formed a group, and *working memory* was listed as the group header. Subjects were instructed to select items from the list of remaining concepts and names and write these under the group heading and also to write a number, using the same 3-point rating scale as before, reflecting their confidence that the concept was in fact a member of that group. Subjects were instructed that no item should be placed in more than one group and that they should assign all items to groups even if this required guessing.

Test 4: Cued recall. The cued-recall test featured 12 true statements. Six of the statements referred to proper names and 6 to concepts. Items for the cued-recall test were selected on the basis of their prominence in the course material. The critical name or concept was deleted from the statement and replaced by the initial letter followed by a blank space. For example, *E— was an early German psychologist who studied the learning of nonsense syllables. In p— inhibition forgetting is caused by interference from prior learning.* The subjects' task was to fill in the blank. Again, subjects were instructed to complete all items and to guess if necessary. For each item, subjects provided a CR as before. Subjects were also informed that the size of the blank did not indicate the length of the missing word.

Test 5: Research methods. Finally, the research-methods test contained 20 recognition statements with two forced-choice alternatives concerning statistics and experimental design: For example, *An experiment that uses different subjects in each experimental condition is known as (A) a between-subjects design, (B) a within-subjects design; Data that can be ranked from best to worst is known as (A) ordinal data, (B) nominal data.* Subjects completed all items and rated their confidence for each response.

Because there were four topic areas (perception, memory, problem solving, and language), 4 versions were constructed for Tests 1–4, 1 version for each topic area. The 16 different versions were grouped into four sets of four tests each, so that each set contained one of each type of test, and each test sampled a different topic area. For example, Set 1 contained recognition (language), fact verification (perception), concept grouping (memory), and cued recall (problem solving). Each set also contained the same research-methods test (Test 5). Within an RI, subjects were cycled through test sets so that Subject 1 would receive Set 1; Subject 2, Set 2; and so on. In this way, roughly equal numbers of sets of tests were completed within the different RIs. For any one subject, then, each test sampled a different topic area, and it was reasoned that this method of presentation of tests would reduce priming from one test to another (e.g., a recognition test for names of memory researchers might prime cued recall of memory researchers' names but not cued recall of the names of language researchers).

All subjects were mailed a pack of tests. The pack contained an introduction to the study, the questionnaire, and the tests. The

introduction stated that the purpose of the study was to investigate memory for names and concepts from the course. Subjects were requested not to consult course material, and it was explained that such consultation would simply invalidate their tests. Also included were general instructions informing subjects that although the tests must be completed in the specified order, they were allowed to take as long as they needed but should complete a test without interruption once started and return to the next test at their convenience. The order of tests was fixed and was designed so that tests that subjects in the pilot study had identified as being "easy" (e.g., recognition) were presented before tests that had been identified as "hard" (e.g., cued recall). The purpose of this ordering was simply to minimize the initial load placed on our subjects and to encourage them to complete all the tests. Thus, subjects always completed the tests in Order 1–5, as listed earlier.

For all tests, subjects were instructed to guess whenever they came across items they found difficult to judge. They were reassured that we did not expect exceptional performance on the tests and were informed that the tests had deliberately been designed to be difficult. Subjects were told "Don't worry if most or all your answers are guesses, but please make sure you complete *all* the items and confidence ratings," and were again reminded of the importance of not consulting course material at any time.

Specific instructions were provided for each of the tests. Each set of instructions contained an example from the test and instructions regarding CRs. For the recognition tests, subjects were instructed to work through the items in the order in which they were listed and to complete an old–new judgment and corresponding CR for each item before proceeding to the next item. Subjects were cautioned not to look ahead at items further down the list. Similar instructions were provided for each test but slightly altered to take into account the nature of particular tests. All subjects were paid a small fee for their participation.

Finally, many of the former students who took part in the study spontaneously wrote to us offering comments on their experience in taking the tests and expressing their curiosity about the study and its outcome. These comments indicated that most subjects had not found the tests onerous but, rather, had found them comparatively simple to complete and inherently interesting.

Results

The results are divided into four parts. The first part reports various analyses of the retention data; the second part deals with the CRs; the third part reports analyses of relations between retention and various predictor variables across all RIs; and the final section reports joint analysis of retention measures, CRs, RI, and age at retrieval.

A few of the subjects' responses were illegible, and very occasionally, a subject omitted a response to one of the questions on the questionnaire. These missing cases totaled less than 0.5% of all responses and, for the purpose of the analyses, were replaced by use of an iterative algorithm from the GENSTAT package (see Alvey et al., 1983, Section 7.1).

Analysis of Measures of Retention

Bahrck (1979) described two ways in which to adjust mean retention scores so that the effects of such factors as initial level of learning and rehearsal are minimized. One method is to regress a range of predictor variables on retention scores within a year and adjust the mean retention score for each of the predictor variables on the basis of these within-years

regressions. This method was followed in the initial analyses of our retention data. Retention scores for each year of the course were treated as separate groups and entered into a cumulative hierarchical regression (see Cohen & Cohen, 1975, p. 98). The order of variables was as follows: First, grade was entered, followed by rating of contact, IRAT, age at retrieval, and topic area. The reasoning behind this regression model was that we expected grade, as an index of original depth of learning, to be most directly related to retention within a year (Bahrck, 1984). Contact as a measure of rehearsal also seemed likely to strongly determine retention, but obviously this is dependent on the types of postcourse contact that actually occurred. The IRATs were included next, as it was reasoned that these might reflect motivational factors leading subjects to attend—during learning—to one topic area, rather than to another. Age at retrieval was included to assess the effects of age, although as less than 10% of subjects in the total sample were over the age of 65 years at test, and this percentage was considerably smaller within some years, we did not expect this factor to influence retention greatly within years. Finally, topic was included as a variable on the assumption that some topics were more accessible than others. However, because of the structured nature of the course and the equal spacing of coursework across topic areas, topic was expected to be only weakly related to retention.

The regressions were performed for each of the tests of retention separately, and the mean scores for retention were adjusted on the basis of these within-years regressions (see Bahrck, 1979). The adjustments, however, made very little difference to the retention curves and therefore are not reported further.² Thus, throughout this section, we report the unadjusted means only.

The percentage of variance accounted for by the full within-years regression model averaged 26% of retention variance over regressions. We also examined individual predictors that—when added to the model—accounted for significant portions of the within-years retention variance. However, the five predictor variables did not consistently predict within-years retention across different RIs for any of the tests. The one exception to this was grade, which did reliably predict retention for the shortest two RIs (3 and 15 months) for the fact-verification, concept-grouping, and research-methods tests. These associations between grade and retention were small and averaged less than 5% of retention variance. In a later section, we report a more sensitive test of the relation between grade and retention, and it might simply be noted here that the failure to find reliable associations between grade and retention is surprising given that Bahrck (1984) found that grade predicted retention at all RIs.

The failure of the other predictor measures to reliably account for retention is less surprising. Ratings of contact were low (see Table 1) and indicated that subjects had few further contacts with any area of psychology once they had completed their course. The IRATs showed some relations with retention, but these were not systematic across RIs, and it may be that this measure simply does not accurately assess interest or motivation: Perhaps, after a lengthy retention period, subjects cannot accurately gauge their earlier interests. There were some instances in which topic was related to

retention within a year, but these were sporadic, rather than consistent, across the RIs. Systematic effects of age were also absent, but again, this is not surprising as within any given year only a few subjects were over the age of 60 years. Thus, age effects were probably not present because there were not enough old subjects within any RI. However, we will also return to the effects of age in a later analysis.

To analyze the effects of different types of test and RI, a number of analyses of variance (ANOVAs) were conducted.³ For the analysis of recognition of names and concepts, the raw data were converted to *d*-primes. This transformation was used because it was expected that discriminability would decrease with RI and also because subjects had been instructed to guess when in doubt. (It should be noted, however, that a similar analysis using the raw data produced identical findings.) The *d*-prime values were entered into a mixed-model ANOVA with the 12 levels of RI forming the between-subjects variable and the two levels of type of test (names vs. concepts) forming the within-subjects variable. A main effect of RI was observed, $F(11, 361) = 10.05$, $MS_e = 7.01$, $p < .0001$, as was a main effect of type of test, $F(1, 361) = 5.553$, $MS_e = 1.77$, $p < .02$. The main effect of type of test arose because more concepts were accurately recognized than names: The mean values of *d*-prime were 1.20 versus 1.09, respectively. Overall, the *d*-prime values ranged from 2.20, for the shortest RIs, to 0.85, for the longest RIs, demonstrating that subjects at different RIs were sensitive to old items from the course material and that false positives did not dominate their responding. The effect of RI and type of test is shown in Figure 1. Note that to provide some indication of absolute levels of performance, Figure 1 plots the mean percentage correctly recognized per RI, rather than *d*-prime, and shows the level of chance performance as a percentage. Note that overall mean percentage of correctly recognized concepts was 71% compared with 68.5%, for names; in both tests across all RIs, the mean percentage of correctly recognized *old* items never fell below 63%, and recognition performance was always reliably above chance. It can be seen from Figure 1 that retention rapidly declines over the first 36 to 48 months of retention and then levels out and stays at the same level throughout the remaining 8 years. It is also clear from Figure 1 that the decline in retention is more rapid for names than concepts. To explore this further, simple main effects analyses were conducted. These analyses found that the only two points in the retention curve in which recognition of names and concepts differed reliably were the 27-month and 39-month RIs. At these points, recognition of concepts was reliably more accurate than recognition of names: $F(1, 361) = 4.58$, $MS_e = 0.32$, $p < .033$, for the 27-month RI, and $F(1, 361) = 5.88$, $MS_e = 0.32$, $p < .016$, for the 39-month RI. These findings show that

² The means were also adjusted using pooled within-years betas (see Bahrck, 1979), but once again, no substantial differences in the retention curves were observed. Full details of all these analyses are available from us on request.

³ Note that exactly the same analyses were conducted using analysis of covariance (ANCOVA) with the variables from the regression model as covariates. Because the ANCOVAs produced the same findings as the ANOVAs only the latter are reported here.

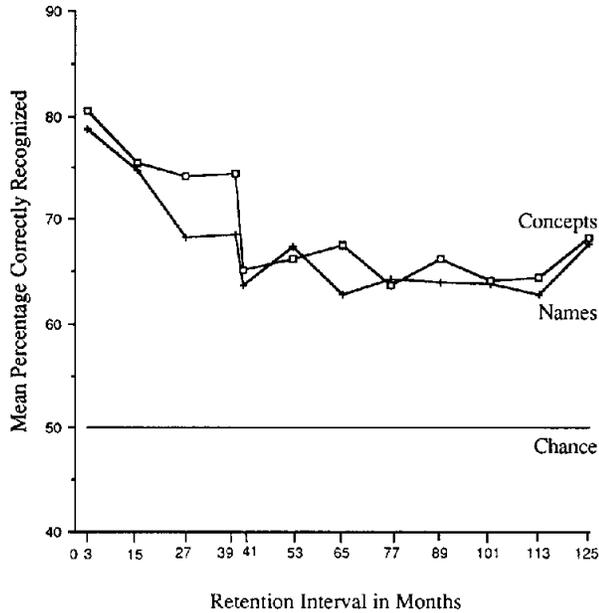


Figure 1. Mean percentages of correctly recognized names and concepts across retention intervals.

the decline in retention of concepts is less rapid than the decline in the retention of names.

An identical ANOVA performed on the percentage of correct responses to cued recall of names and concepts produced similar findings. (Note that for this and all subsequent analyses percentage correct was analyzed.) In this case, significant main effects of RI, $F(11, 361) = 5.07$, $MS_e = 5,356$, $p < .0001$, and the interaction of RI by type of test, $F(11, 361) = 3.54$, $MS_e = 802$, $p < .0001$, were observed. The main effect of type of test failed to reach significance, $F(1, 361) = 3.401$, $MS_e = 771$, $p = .06$, and the mean number of items correctly recalled for names was 30.2% compared with 32.3% for concepts. The effects of RI are shown in Figure 2 in which the mean percentage correct for each year is plotted against RI in months. Note that chance level for cued recall, established on the basis of control subjects' responses, was zero. Simple main effects were calculated comparing retention at each RI across types of tests. Cued recall of concepts was reliably higher at 15 months, $F(1, 361) = 11.33$, $MS_e = 226$, $p < .001$, and 27 months, $F(1, 361) = 17.80$, $MS_e = 226$, $p < .001$. Thus, the differential decline in memory for names versus concepts observed in the recognition tests was preserved in the cued-recall tests, and it should be noted that the two tests assessed memory for different names and concepts. A reversal of this effect was observed at the 125-month RI in which cued recall of names was reliably higher than cued recall of concepts, $F(1, 361) = 8.80$, $MS_e = 226$, $p < .003$. This difference in the retention functions was, however, uninterpretable. No other significant effects were observed in this analysis.

A rather different pattern of findings emerged from a similar ANOVA of general and specific fact verification. In this case, there was no reliable effect of RI ($F < 1.76$) or of type of test

($F < 1$). However, the interaction of RI with type of test was significant, $F(11, 361) = 3.32$, $MS_e = 532$, $p < .001$. The percentage correct for general and specific fact verification plotted against RI are shown in Figure 3, and chance level of performance is also plotted. It is apparent from Figure 3 that there is no systematic decline over time in memory for general facts, and memory performance fluctuates across RIs around a mean of 65%. For specific facts, there is a small but constant decline in retention over the whole RI, although this is somewhat masked by an unusually high level of memory performance at the 77-month RI. Using the method of orthogonal polynomials (Keppel, 1973, p. 113), we found the curve for specific facts to be fit by a significant linear trend, $F(1, 361) = 10.30$, $MS_e = 214$, $p < .002$. For general facts, however, no reliable trend was observed ($F < 1$). Memory performance for general facts did not produce any reliable differences across RI, and general and specific facts differed significantly at a number of points in the retention period: At 27, 39, and 77 months, memory for specific facts was superior to memory for general facts, and at 65 and 89 months, memory for general facts was superior to memory for specific facts.

The analyses of the grouping and research-methods tests used between-subjects designs, and RI was the only factor analyzed. A significant effect of RI on grouping was observed, $F(11, 361) = 4.20$, $MS_e = 1,717$, $p < .001$, and this is shown in Figure 4. It can be seen from Figure 4 that the effect of RI arose because of a rapid decline in grouping accuracy over the first 2 RIs, followed by a leveling off of performance that was then maintained at an above-chance level over the remaining RIs. For research methods too, RI was significant, $F(11, 361) = 2.88$, $MS_e = 380$, $p < .01$, and the retention curve is shown in Figure 5. In this case, however, no decline

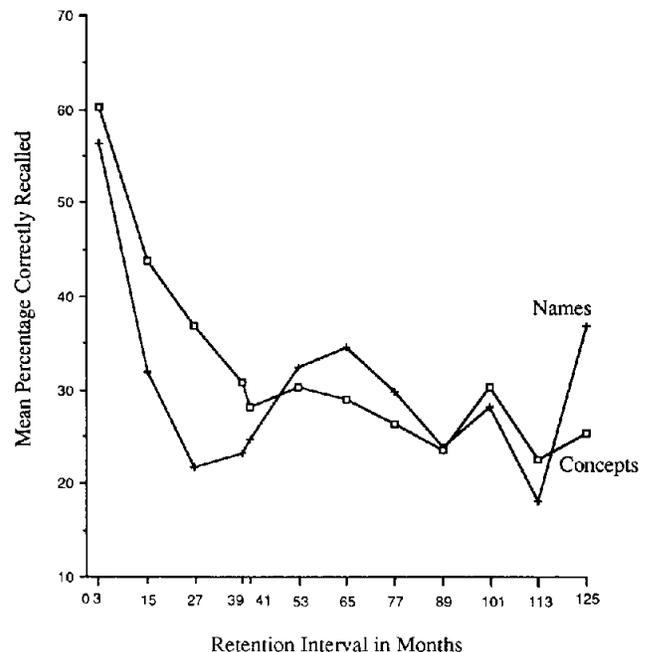


Figure 2. Mean percentages of correctly recalled names and concepts across retention intervals.

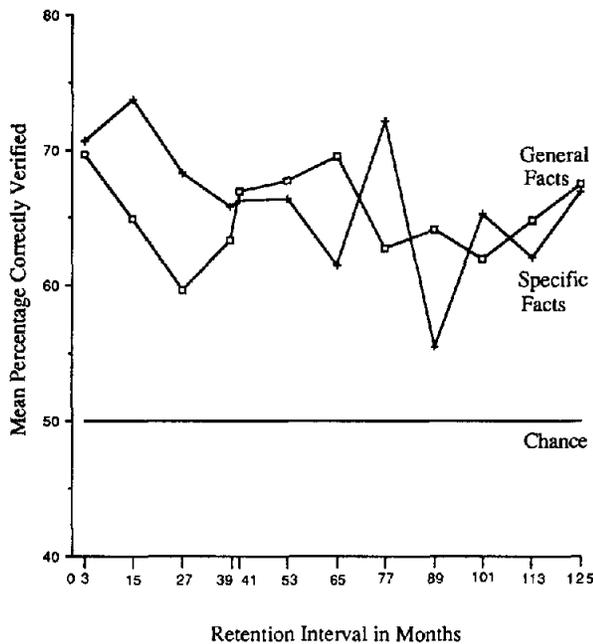


Figure 3. Mean percentages of correctly verified general and specific facts across retention intervals.

in memory performance across the RIs was present (see Figure 5). Indeed, comparisons among the means for each RI found that memory performance at the 125-month RI was significantly higher than retention at all other RIs, which in turn did not differ from each other.

In summary, then recognition of names and concepts led to different patterns of retention, and concepts were forgotten less quickly than names. This pattern of retention was also present in cued recall in which, again, concepts were forgotten less rapidly than names. Grouping of concepts also showed a marked decline in memory performance over the first 2 years of retention, whereas memory for specific facts showed a more gradual decline in retention over the whole retention period. No reliable or systematic decline in retention of general facts and research methods was observed. Memory for general facts remained at a fairly constant 65% correctly verified throughout the whole RI, and memory for research methods remained at 75% correct verification. It should also be noted that on all tests memory performance was always reliably above chance and, even at the longest RIs, never fell to chance levels.

Confidence Ratings

Exactly the same analyses were conducted on the ratings of confidence for each of the tests with the exception of the cued-recall tests. The CRs for cued recall were not analyzed as a number of subjects (18%) failed to recall any items and, thus, could not provide CRs. Cumulative regressions of grade, contact, IRAT, age, and topic on within-years CRs did not lead to marked adjustments to mean retention scores, and the adjusted scores are therefore not reported further. On average,

the regressions accounted for 22% of the variance in the CRs, although there was no systematic pattern for any individual predictor across tests and RIs. Once again, grade was the most frequent single predictor but, in contrast to the retention tests, was not associated with the most recent RIs. Thus, initial level of learning, as indexed by grade, does not appear to be a systematic predictor of confidence in memory performance.

Figure 6 shows the mean CR for each of the six texts across RIs. In the ANOVA of CRs for recognition of names and concepts, significant main effects of RI, $F(11, 361) = 5.80$, $MS_e = 1.598$, $p < .001$, and type of tests, $F(1, 361) = 168.60$, $MS_e = 9.5$, $p < .001$, were observed. It can be seen from Figure 6 that the effect of RI reflects the steep decline in CR for both names and concepts over the first 6 RIs and that CR stabilizes after the 77-month RI. This decline is striking and represents a shift in CR from "highly confident" to "guessing," for recognition of names, and from "highly confident" to "moderately confident," for concepts. These differences in confidence were reliable as shown in the main effect of type of test, and the mean CR for names was 1.91 compared with 2.14 for concepts. Moreover, in analyses of simple main effects, CR for concepts was significantly higher than CR for names at all RIs. Thus, the advantages of concepts over names present in the recognition tests are reflected and amplified in the CRs in which subjects are reliably more confident in recognizing concepts. However, it should be noted that CR does not accurately parallel the differential decline in retention of names and concepts observed in the recognition tests.

In the analysis of fact-verification CRs, main effects of RI, $F(11, 361) = 3.23$, $MS_e = 0.99$, $p < .001$, and type of test, $F(1, 361) = 4.90$, $MS_e = 0.39$, $p < .05$, were found. The effect of type of test was small, and mean CR was 2.20, for general

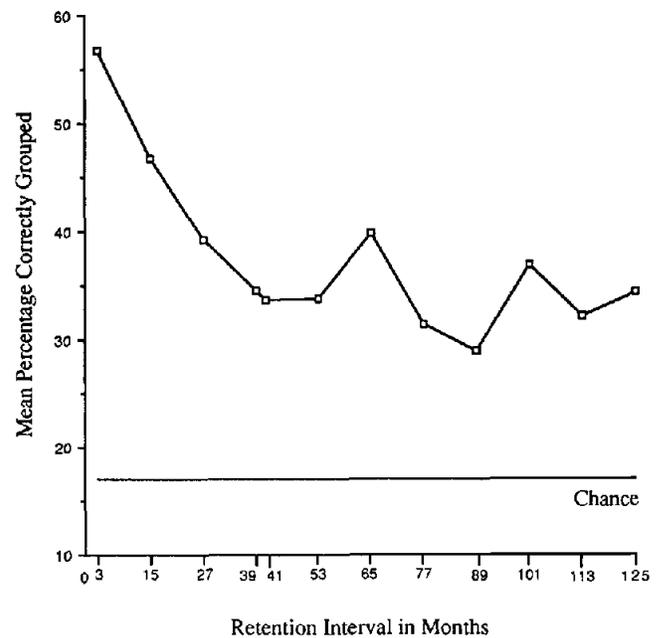


Figure 4. Mean percentages of correctly grouped concepts across retention intervals.

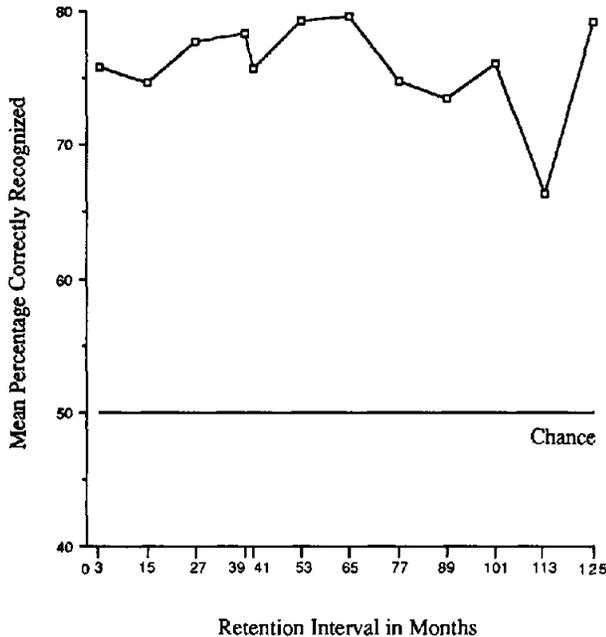


Figure 5. Mean percentages of correctly recognized items from research methods across retention intervals.

facts, compared with 2.14, for specific facts. There was, however, a significant interaction of RI with type of test, $F(11, 361) = 2.45$, $MS_e = 0.19$, $p < .01$, and it can be seen from Figure 6 that CR for general facts declines only gradually over the whole retention period, whereas CR for specific facts declines sharply. This agrees, to some extent, with the retention-test findings of a gradual decline in memory performance for specific facts compared with stable retention of general facts.

The grouping test showed the most marked decline in CR, $F(1, 361) = 3.24$, $MS_e = 1.20$, $p < .001$, falling from 2.20, for the most recent retention interval, to 1.70, for the most remote (see Figure 6). This shows a drop from responses of "highly confident" and "moderately confident" to responses of "moderately confident" and "guess." Finally, the research-methods test showed no reliable effect of RI, and mean CR for all points in the retention period was dominated by "highly confident" judgments.

In terms of main effects, the pattern of findings from the CRs corresponds to the general pattern of findings for memory performance, although some particular effects present in the memory tests, such as more rapid forgetting of names than concepts on recognition tests, are not present in the CRs. However, in terms of responses plotted across RIs, the decline in CR, when this occurs, is far more rapid than the corresponding declines in retention observed in the memory tests. Thus, when the tests of retention show above-chance levels of performance (i.e., at the longer RIs), the corresponding CRs indicate that, by and large, subjects were guessing. These dissociations between memory performance and CR are considered in detail in the General Discussion section.

Predicting Retention

One problem with the within-years regressions reported earlier was that by sampling students within any 1 year of the course, only comparatively small numbers of students could be assessed (30 per year on average), and this may have reduced the sensitivity of these regressions. In this section, we report cumulative hierarchical regressions across the whole sample of 373 students for each measure of retention. In the model used in these regressions, the first predictor to enter the regression was grade, followed by contact, IRAT, age, RI, and CR. Increments to the percentage of retention variance accounted for by the model were calculated as each variable was entered into the model, and Table 2 shows the additional variance added by each predictor variable. (Note that topic is not included in this model as other analyses had failed to find any systematic associations, and very few reliable associations, between topic and retention and between topic and all other variables.)

The reasons for entering grade, contact, IRAT, and age as the first 4 variables were the same as those stated earlier. The RI, treated as a continuous predictor variable expressed in months, was included in the model to further examine the relation between RI and retention. However, it seemed important to examine whether there were any effects of RI over and above the effects of initial level of learning (grade), rehearsal (contact), interest (IRAT), and age, and hence the position of RI in the regression model. Finally, CRs were entered last in the model, and the purpose of this was to examine the relation of confidence to memory performance after all other factors had been taken into account.

It can be seen from Table 2 that grade is a significant predictor of retention for all tests, averaging 5.6% of retention

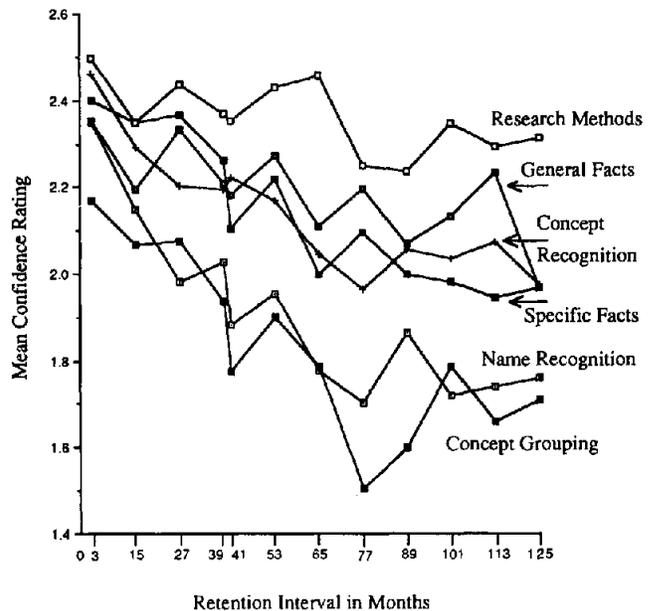


Figure 6. Mean confidence ratings for all tests across retention intervals.

Table 2
Percentage of Retention Variance Accounted for by Grade, Contact, Interest Rating (IRAT), Age, Retention Interval (RI), and Confidence Rating (CR) for Eight Tests of Retention

Test	Predictors						Total
	Grade	Contact	IRAT	Age	RI	CR	
Recognition of names	4.9*	0.4	0.9	0.6	6.3*	6.2*	19.3*
Recognition of concepts	4.3*	0	3.7*	2.4*	9.5*	1.5*	19.3*
Cued recall of names	2.2*	0.7	3.1*	0	1.2*	—	7.2*
Cued recall of concepts	5.6*	0.1	5.1*	3.9*	5.1*	—	19.8*
Verification of general facts	6.2*	0.2	0.7	0	0	1.6*	8.7*
Verification of specific facts	4.4*	0.2	1.6*	2.5*	0.6	4.9*	15.1*
Grouping	7.7*	0.4	1.7*	1.3*	4.2*	20.7*	36*
Research methods	9.6*	1	—	0	0.2	8.3*	19.1*

* $p < .05$.

variance across tests. This small but reliable association between grade and retention reflected negative correlations between grade and retention, in which students with higher grades (1 and 2) remembered more than students with lower grades (3 and 4). In additional analyses (see Cohen & Cohen, 1975, p. 202), the interaction of Grade and RI was not found to be a reliable predictor of retention ($t < 1.2$), for all tests, indicating that the effects of grade on retention were constant across RIs. Furthermore, when the model was changed so that grade entered the regression as the last variable, all the significant effects were preserved, although the average variance accounted for was reduced to 4.1%. Thus, in these more sensitive analyses using all subjects, higher grades are associated with higher levels of memory performance, indicating that initial learning does influence very long-term retention as Bahrack (1984) found.

Unlike grade, however, contact did not emerge as a significant predictor (see second column of Table 2), and this is because very few of our subjects had further contact with or experience of cognitive psychology. In contrast IRAT, which had been a highly inconsistent predictor in the within-years regression analyses, emerged as a reliable predictor for six of the retention tests. Again, however, this association—when present—was small and accounted for an average of only 2.2% of variance in retention. Nevertheless, this finding indicates that the level of interest taken by a student in a particular area can also influence very long-term retention of knowledge. Effects of age at retrieval are also present in Table 2, and these too are small and only present on tests that show RI effects (see Figures 1–5).

From the third column of Table 2, it can be seen that RI is generally a reliable predictor on those tests that showed retention effects in the ANOVAs, although the percentage of variance accounted for is low. This is unsurprising, as mean retention, as measured by the memory tests, only varies with RI for the first 2–4 RIs and therefore could not be expected to account for a large portion of the variance in retention. Nevertheless, these RI effects arise even after grade, contact, IRAT, and age have been taken into account and must surely reflect the rapid decline in retention over initial RIs.

Finally, CR accounts for somewhat larger portions of retention variance and is a reliable predictor on all the tests in which this analysis could be performed. The correlations

between CR and retention were always positive, indicating that higher CR was reliably associated with more accurate memory performance. In addition to this, the correlation matrices for all the variables in the regression model for each test of retention were examined to gauge the degree of association between grade and CR. Across all tests, the correlations were low and never exceeded $r = -0.194$ with an average correlation across tests of $r = -1.48$, and none of these correlations were significant. Thus, CR appeared to be unrelated to grade.

Relations Between Retention, Confidence, RI, and Age

To examine overall relations between type of retention test, CR, RI, and age at retrieval, one final analysis was conducted. All the measures of retention and CR, together with RI and age, were entered into a principal-components orthotransvarimax factor analysis. Table 3 shows the factor loadings of these variables in an oblique four-factor solution.

It can be seen from Table 3 that all the tests of retention, with the exception of the recognition tests, load heavily on Factor 1. The recognition and the grouping tests also load on other factors, so Factor 1 is primarily defined by the cued-recall tests, and we therefore refer to this factor as the *Recall Factor*. Factor 2 is made up of the CRs, although name recognition CR and grouping CR also load on Factor 4. We refer to this factor as the *Confidence Factor*. Factor 3 is defined solely by RI and age and is referred to here as the *Temporal Factor*. Finally, Factor 4 is defined by the recognition tests and is referred to here as the *Recognition Factor*. Intercorrelations between factors were low ($r < 0.17$), with the exception of the Recall and Confidence Factors, which showed a very modest intercorrelation of $r = 0.228$. The only variable to show any factor complexity was grouping (factor complexity measure greater than 2), and it can be seen from Table 3 that grouping clearly contains both a recall and recognition component. Overall, however, the factor structure is remarkably clear, and the majority of variables do not load on more than one factor.

Three aspects of Table 3 are particularly notable. First, none of the retention variables or CRs loaded on the Temporal Factor, indicating that overall, variance in these measures was—by this analysis at least—not reliably associated

Table 3
Loadings of Retention Tests, Confidence Ratings (CRs), Retention Interval in Months, and Age at Retrieval on Four Factors

Test or variable	Factor			
	1	2	3	4
Recognition of names	—	—	—	0.759
Recognition of concepts	—	—	—	0.762
Cued recall of names	0.775	—	—	—
Cued recall of concepts	0.702	—	—	—
Verification of general facts	0.533	—	—	—
Verification of specific facts	0.576	—	—	—
Grouping	0.541	—	—	0.564
Research methods	0.411	—	—	—
CR-recognition of names	—	0.579	—	0.455
CR-recognition of concepts	—	0.724	—	—
CR-verification of general facts	—	0.788	—	—
CR-verification of specific facts	—	0.681	—	—
CR-grouping	—	0.339	—	0.457
CR-research methods	—	0.744	—	—
Retention interval in months	—	—	0.752	—
Age at retrieval	—	—	0.864	—

Note. Only loadings of 0.3 or higher are shown.

with variance in RI and age. Second, variables making up the Recall Factor did not load on the Recognition Factor, indicating little variance overlap in memory performance on recall and recognition tests. The grouping test, however, clearly included both recall and recognition components. Third, CR measures did not, in general, load on factors other than confidence, and this suggests that variance in CR judgments was not primarily related to variance in memory performance.

General Discussion

The findings of the present study share one major similarity with those of Bahrck and his colleagues (Bahrck, 1984; Bahrck et al., 1975). Retention functions were observed that showed a rapid decline in memory over the first few years of retention, which then leveled out and remained at an above-chance level up to the longest RI sampled (125 months). In other respects, however, the present findings extend or differ from earlier research. For instance, the initial decline in the retention of concepts was less rapid than the corresponding decline in the retention of names, and this novel finding relates directly to the purpose of higher education, which is, surely, to provide students with an enduring understanding of a subject, rather than an enduring memory for arbitrary details, such as proper names. In contrast to previous research, some of the tests in the present study showed no decrements in memory performance across the retention period, and this suggests that knowledge structures formed at acquisition may represent different types of knowledge in different ways. Finally, although initial level of learning was found to be reliably associated with measures of retention, these associations were considerably smaller than those observed in previous studies. In this discussion, we consider these differences in detail and develop an account of the nature of retained knowledge.⁴

Differences Between Names and Concepts

Consider, first, the differences in recognition memory between names and concepts. Recognition of names declined rapidly over the early RIs, reaching the asymptote for name recognition at the 27-month interval. In contrast, recognition of concepts declined less rapidly, reaching the asymptote for concept recognition at the 39-month interval, the point at which both curves meet and remain at the same level throughout the remainder of the 125-month RI (see Figure 1). A very similar pattern was present in cued recall in which recall of names declined even more rapidly, over a 15-month period, whereas recall of concepts, again, declined less rapidly over a 39-month period (see Figure 2).

⁴ The aim of the present study is to probe retained knowledge in a variety of different ways, and we hope, in this way, to provide a more extensive examination of patterns of very long-term retention than that achieved by previous work. However, one inherent problem with both the present study and previous studies in this area relates to the validity of intertest comparisons. For instance, different types of tests (e.g., cued-recall vs. recognition) may vary in their task sensitivity, and therefore, any differences in retention between tests might be ascribed to type of test, rather than to the nature of retained knowledge. Thus, although the data in Figures 1–5 indicate that different tests yield different levels of performance and different retention functions, it is difficult to evaluate these differences unambiguously. The chance levels of performance differ, and even if one equated performance across tests in terms of the extent to which students' scores exceed chance levels, differences in test difficulty still cannot be ruled out. We attempted to obviate this problem by using pairs of tests (e.g., recognition of names vs. concepts, cued recall of names vs. concepts, and fact verification of specific vs. general facts), and we confine our main comments to these intratest contrasts, although some comments featuring intertest comparisons are made, particularly with reference to the factor analysis reported earlier (see Table 3).

These differences in retention between names and concepts lend some support to the schema view of very long-term retention (i.e., Neisser, 1984). Schemata, by definition, do not represent detailed knowledge but, rather, represent knowledge abstracted from sets of experiences (Alba & Hasher, 1983; Schank & Abelson, 1977; e.g., numerous trips to restaurants or encounters with various sets of concepts in different contexts, such as essays, laboratory reports, different parts of the course material, and so on). Thus, schematic knowledge structures would be highly unlikely to extensively represent arbitrary details of events, such as the color of the waiter's tie or the names of researchers associated with particular aspects of psychology. Indeed, recent findings suggest that the arbitrary nature of proper names may, generally, preclude their direct representation in schematic knowledge structures. For example, Cohen (1990) argued that proper names lack semantic associates and are not part of the semantic knowledge network. It follows from this view that proper names are not represented in memory in terms of abstract conceptual knowledge. Thus, if, as Neisser claimed, schematic knowledge is retained, whereas specific details are lost, then this would explain the differential retention of concepts and names found in the present study. The knowledge structures or schemata resulting from the processing of course material would centrally represent conceptual knowledge, rather than names, and so the different retention functions for names and concepts, in both recognition and recall, may reflect the more rapid loss of schema-peripheral knowledge (names) than schema-central knowledge (concepts).

Fact Verification and Concept Grouping

Performance on the fact-verification tests was generally high across all RIs, further demonstrating that conceptual knowledge was comparatively well retained. Verification of specific facts showed a constant and reliable decline in retention over the whole 125-month RI, although this was a small decrement of less than 3% between earliest and most remote RIs. Moreover, an uninterpretable rise in memory performance for specific facts was present at the 77-month RI (see Figure 3). Nevertheless, this pattern of retention for specific facts differed from the pattern for general facts, which showed no systematic decrement over the whole retention period. Thus, as with recognition and cued recall, the fact-verification findings can be taken to support a schema theory that proposes that schema-central knowledge (general facts) is better retained than schema-peripheral knowledge (specific facts).

The grouping test was a more direct measure of the retention of schematic knowledge and largely focused on the grouping of concepts, although some proper names were included. The retention curve resulting from the grouping test showed a 27-month decline in retention as opposed to a 39-month decline for the recognition and cued recall of concepts (see Figure 4). This steeper decline in the grouping test perhaps reflects the inclusion of some name material. Yet proper names comprised less than 20% of all the items used in the grouping tests, and moreover, the same pattern of retention was present in groups that did and did not include proper

names. Thus, the decline in retention for this test cannot be explained by the rapid forgetting of proper names. Indeed, this is the one test that should show very little decline in memory performance if, as we have argued, knowledge is maintained by schemata that represent abstract relations between concepts. We return to this point shortly.

Component Processes in the Retention Tests

Consideration of the component processes involved in the different tests suggests further aspects of the knowledge structures underlying memory performance. For instance, one major difference between the fact-verification and grouping tests, observed in the factor analysis, was that the grouping test loaded equally on the recall and recognition factors, whereas the fact-verification tests loaded on the recall factor only (see Table 3). This difference is surprising as it might be argued that fact verification and concept grouping require similar processes: the retrieval of relational conceptual knowledge, generation of relational conceptual knowledge, or both. However, it seems possible that accurate fact verification may require less relational conceptual knowledge than accurate concept grouping.

For example, to verify that *X* is a proposition of Theory *Y*, it may simply be sufficient to recall that *X* and *Y* were presented jointly in the course material. In contrast, to correctly group *X*, *W*, and *Z* under *Y*, relations of each item to *Y* must be separately accessed and, quite possibly, so must interitem relations. In terms of memory processes, this means that concept grouping may require more memory processes than fact verification, thus providing more points at which grouping might fail. Consider the case in which a subject scanning the list of items recognizes only one item as related to a group heading (e.g., the subject recognizes that *Wason* should be grouped with *four-card problem*). To select other items from the list, additional knowledge must be retrieved. Possibly, this is an incremental process, and as the individual recognizes a pair of related concepts, an effective retrieval context is constructed that facilitates the recall of group-related items and the eventual correct classification of concepts. In the case of the aforementioned example, the subject might use the compound-cue (Ratcliff & McKoon, 1988) *Wason's four-card problem* to recall that *Wason's* subjects had great difficulty in using falsification strategies when solving the four-card problem, and only at this point recognize *falsification* as an item that can be grouped with *four-card problem* and *Wason*, hence the loading of the grouping test on both recognition and recall factors.

It might be argued that fact verification also contains a large recognition component and so should similarly have loaded on both the recall and recognition factors. To see why this may not have been the case in the present study, the nature of the three tests—recognition, grouping, and fact-verification—must be considered. The recognition and grouping tests both require subjects to discriminate between items: recognition between old and new items, and grouping between different conceptual groupings of items. But recognition, unlike grouping, need not entail a recall process. Similarly, fact

verification need not require discrimination, for a subject merely has to recall that two items either did or did not co-occur together in the course to make an accurate judgment. Thus, such tasks as fact verification and recognition may be simple in that they require the operation of only one memory process, whereas concept grouping is comparatively complex because it involves the operation of two or more processes, as our factor analysis indicated.

Retention of Research Methods

Finally, memory performance for concepts in research methods showed no decline over the retention period and remained at a reliable level of 30% above chance throughout the 125-month RI. Four points appear relevant to this finding. First, the course material for research methods was highly structured and circumscribed, essentially covering experimental design and statistics. As Neisser (1984) suggested, this high degree of structure may predispose subjects to form schematic representations of a knowledge domain, and these schemata may mediate effective long-term retention. Second, students carried out a number of experimental projects, and so research methods were undoubtedly covered more frequently than other areas. Third, as students were required to apply their knowledge of research methods in their coursework experimental projects, it is possible that some aspects of this knowledge became proceduralized, and perhaps this proceduralization of knowledge conferred some immunity to forgetting. Fourth, this was the one area of the course that had some overlap with other courses taken in previous years, and it is possible that this prior exposure may have influenced the high levels of long-term retention for research methods. Because of this last point, we do not consider research methods in further detail here. It is, however, worth noting that students were required to utilize their knowledge of research methods in practical settings. This was not the case for other knowledge acquired in the course, and it might tentatively be suggested that learning exercises that require the practical application of knowledge (i.e., active learning) lead to levels of retention that do not show a rapid decline with time.

Very Long-Term Retention and Schema Theory

We argued earlier that schema theory can be used to interpret some of the patterns of retention observed in the present study, but what sort of schema theory? Neisser's (1984) proposal was that schemata act to reconstruct specific items and relations between items and so produce high levels of memory performance. The suggestion appears to be that after some initial decline in specific knowledge (schema-peripheral knowledge), which occurs over the first few months or years of retention, a residue of schema-central knowledge remains and is preserved in long-term memory. This schema-central knowledge, which may have been formed in response to extended periods of naturally occurring, spaced practice (see Bahrick, 1984, and Bahrick & Phelps, 1987), comprises

abstract or decontextualized knowledge (Barsalou, 1982) and rules. (Indeed, schemata must primarily represent decontextualized knowledge, for if they represented detailed accounts of experienced events, then they might be more accurately characterized as autobiographical memories; see Conway, 1990). In the case of schemata for cognitive psychology, schema-central knowledge may take the form of propositions asserting theoretical relationships linking groups of concepts and pointing to sets of associated research findings. According to Neisser's view, this abstract schema-central knowledge can be used to reconstruct specific knowledge and so give the impression that specific knowledge has been retained in long-term memory. We call this the *strongly reconstructive view* of long-term retention (after Brewer, 1986).

What predictions does the strongly reconstructive view make for our tests? Retention of general facts should show little or no decline as these facts are schema-central and represent the core schema knowledge that is resistant to forgetting. Specific facts should show some early decline simply because many of these will be schema-peripheral. However, many (although not all) specific facts should be reconstructible from schema-central knowledge, and so after an initial decline, the retention function should stabilize at an above-chance level. Thus, our findings for memory for facts fit the strongly reconstructive view reasonably well.

The strong version of reconstruction also predicts that memory for names should decline rapidly and that this decline should continue until chance levels are reached. This is because proper names can neither be abstracted in the way that concepts can nor be reconstructed inferentially. Memory for names did indeed decline, as the strongly reconstructive view predicts, but asymptoted reliably above chance and remained at that level throughout the whole retention period (Figures 1 and 2). Thus, the findings for retention of names show that some significant proportion of proper names is retained over very long RIs without any notable decrement. The strongly reconstructive view cannot account for this.

Finally, knowledge of conceptual relations should be well retained and should show an essentially flat function across the retention period, indicating little, if any, forgetting. This is because schemata, by definition, represent abstract conceptual relations as core schema-central knowledge from which specific knowledge can be reconstructed or inferred. The rapid decline in knowledge of conceptual relations observed in the conceptual grouping task does not support this prediction of the reconstructive model, although the above-chance asymptote for retention of conceptual relations present in the concept-grouping task (see Figure 4) indicates that at least some conceptual relations are preserved over long RIs.

Our findings, then, provide only mixed support for the strongly reconstructive view of the long-term retention of knowledge as proposed by Neisser (1984). Particularly problematic for the reconstructive view are the retention of names and the declines in conceptual knowledge. Our students did show superior retention of general knowledge, but this did not appear to be in the form of abstract schemata because memory for conceptual relations was poor. It seems likely, therefore, that students retained a core of general knowledge

supplemented by fragmentary knowledge of specific facts, some names and co-occurrences of related items. Thus, the findings suggest that the original knowledge structures formed by the students were, primarily, highly specific and detailed rather than abstract and relational. After a 125-month RI, during which virtually no rehearsal took place, a significant portion of this specific knowledge is still retained.

Effects of Initial Learning, Confidence, and Age

One further difference between the present findings and those of Bahrck (1984) was that although initial depth of learning was systematically related to very long-term retention, as Bahrck found, the correlations were small and accounted for an average of only 5.5% of variance in the retention measures. The index of initial learning was the grade a person obtained for the course. It might be argued that this was not a sensitive measure of learning, hence the failure to find larger associations between grade and retention. It is difficult to see how this argument can be sustained when one considers how grade was assessed. Grade was based on an assessment of learning shortly after completion of acquisition (the final examination). Marking schemes were standardized, and markers were provided with model answers; in addition to this, the coursework covered all the topic areas assessed in this study equally. It is difficult to see how a more sensitive measure could be devised.

However, the main reason for this difference between the present study and that of Bahrck (1984) may relate to the more comprehensive measure of learning used by Bahrck. The former students in Bahrck's study had taken a number of Spanish courses at different levels of expertise, and so levels could be used as an index of initial learning. Apart from levels constituting a more comprehensive measure of original acquisition, it is the case that students who took the highest level courses both learned and remembered more. Unfortunately, in the present study, students took only one course in cognitive psychology, and with very similar course profiles for previous courses, it was not possible to develop a measure of acquisition analogous to that of *levels* used in Bahrck's study. Nevertheless, the present findings support the view that original learning is a consistent determinant of long-term retention with the qualification that, in the case of a single course, the relation of learning to retention is small.

Remarkably, grade was not reliably associated with judgments of confidence in memory performance, and subjects who had obtained a third-class grade were just as certain or uncertain about their memory performance as subjects who obtained a first-class mark. This was the case at all RIs. Reliable but small associations between confidence and memory performance were, however, present across all tests (Table 2), and higher ratings of confidence were associated with higher levels of memory performance. Thus, not surprisingly, subjects often knew they were correct in the recall and recognition of the various items, and this was independent of the grade a student obtained in the course.

Yet, CRs did not account for large portions of the variance in the measures of retention, and it should be noted that across all tests, CRs accounted for an average of only 7.2% of

memory performance variance. Moreover, the distribution of CRs across the whole retention period showed rapid and systematic declines on all tests (Figure 6). This linearly decreasing level of confidence with increasing retention interval indicates that the CRs cannot have been solely, or even primarily, based on memory performance because the patterns of retention for the memory measures did not exhibit linear decreases with time. Indeed, CRs (on some tests) rapidly fell to the guessing level over the whole retention period, whereas memory performance remained at above-chance levels. CRs formed a separate factor in the factor analysis and were not associated with the tests of retention.

One possible explanation here is that CRs were largely determined by a metamemory belief, rather than by memory performance per se. Subjects may have believed that the probability of correct memory performance decreased with increasing retention period, and as a consequence of this belief, the CRs of subjects at the longer intervals may have been guided (implicitly, explicitly, or both) by the belief that given that they took the course a number years ago, then they were unlikely to be highly accurate in their responses on the memory tests. Indeed, simple regressions of RI in months on CR for each of the six tests of retention for which CRs were available (see Table 3) found significant associations on all tests between RI and CR, further suggesting that a metamemory belief may have mediated CRs. That this belief is quite wrong can be clearly seen by comparing Figures 1-5 with Figure 6.

Interestingly, this dissociation between memory performance and CR with increasing RI suggests that subjects' performance on the memory tests (especially name-recognition and concept-grouping) at the longer RIs may have been guided, on at least some occasions, by implicit, rather than explicit, memory (see Schacter, 1987, for a review of implicit and explicit memory). Indeed, this must have been the case, because at the longer RIs, subjects' CRs indicated that they were guessing (i.e., had no phenomenal experience of remembering), but their guesses were well above chance. Intriguingly, this suggests that acquired knowledge may influence behavior even when an individual is unaware of having retained that knowledge. One implication of this is that memory tests that require conscious, effortful, conceptually driven retrieval processes (see Roediger & Blaxton, 1987) may underestimate the amount of retained knowledge. In contrast, tests that are data driven (e.g., perceptual-identification) may reveal the retention of knowledge that an individual is unable to access consciously.

Lastly, reliable effects of age at retrieval were also found, and older subjects performed at lower levels than younger subjects. These effects were, however, small and only present on tests that showed declines in retention. That age effects are small and not consistent across RIs is due to the nature of the subject sample. Relatively few subjects were over 65 years, the age at which memory impairment typically begins to be manifest, and older subjects were unevenly distributed across RIs. It is usual to find that age-related deficits are confined to more difficult tasks and do not appear in relatively easier tasks (see Cohen, 1988). Hence, it is not surprising to find, in this study, that age-related deficits in performance are con-

fined to the tests that show a decline with increasing RI and do not appear in the tests in which retention proved to be time-resistant. However, the existence of age effects in some tests does suggest that the final acceleration on forgetting in Bahrick's (1984) retention curves that corresponded with subjects being in their sixties may have been caused by an age-related memory impairment (see Cohen, Stanhope, & Conway, in press).

In summary, the present study found that in the very long-term retention of cognitive psychology, proper names were forgotten more rapidly than concepts. Knowledge of relations between concepts was also forgotten rapidly, and knowledge of specific facts showed a small but reliable decline with increasing RI. In contrast, general factual knowledge and knowledge of research methods showed no decline with increasing RI. This pattern of findings does not fit well with a strongly reconstructive view of very long-term retention but, rather, suggests that specific and detailed knowledge may be retained in memory over very long retention periods. We conclude that, in the present study, knowledge was retrieved from memory, rather than reconstructed by schemata.

References

- Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*, 203-231.
- Alvey, N. G., et al. (1983). *GENSTAT: A general statistical programme*. Oxford, England: The Numerical Algorithms Group.
- Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*, 296-308.
- Bahrick, H. P. (1983). The cognitive map of a city: Fifty years of learning and memory. *Psychology of Learning and Motivation*, *17*, 125-163.
- Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, *113*, 1-29.
- Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*, 54-75.
- Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 344-349.
- Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition*, *10*, 82-93.
- Brewer, W. F. (1986). What is autobiographical memory? In D. C. Rubin (Ed.), *Autobiographical memory*. Cambridge, England: Cambridge University Press.
- Cohen, G. (1988). Memory and aging: Toward an explanation. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 77-83). New York: Wiley.
- Cohen, G. (1990). Why is it difficult to put names to faces? *British Journal of Psychology*, *81*, 287-298.
- Cohen, G., Eysenck, M. W., & Le Voi, M. E. (1986). *Memory: A cognitive approach*. Buckingham, England: Open University Press.
- Cohen, G., Stanhope, N., & Conway, M. A. (in press). Age differences in the very long-term retention of knowledge. *British Journal of Developmental Psychology*.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlational analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Conway, M. A. (1990). *Autobiographical memory: An introduction*. Buckingham, England: Open University Press.
- Greene, J. (1986). *Language understanding: A cognitive approach*. Buckingham, England: Open University Press.
- Jones, G. V. (1978). Recognition failure and dual mechanisms in recall. *Psychological Review*, *85*, 464-469.
- Kahney, H. (1986). *Problem solving: A cognitive approach*. Buckingham, England: Open University Press.
- Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.
- Kintsch, W. (1978). More on recognition failure of recallable words: Implications for generation-recognition models? *Psychological Review*, *85*, 470-473.
- Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). San Diego, CA: Academic Press.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252-271.
- Naveh-Benjamin, M. (1988). Retention of cognitive structures learned in university courses. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 383-388). New York: Wiley.
- Neisser, U. (1984). Interpreting Harry Bahrick's discovery: What confers immunity against forgetting? *Journal of Experimental Psychology: General*, *113*, 32-35.
- Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming. *Psychological Review*, *95*, 385-408.
- Roediger, H. L., & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. Gorfein & R. R. Hoffman (Eds.), *Memory and cognitive processes: The Ebbinghaus Centennial Conference* (pp. 349-379). Hillsdale, NJ: Erlbaum.
- Roth, I., & Frisby, J. (1986). *Perception and representation: A cognitive approach*. Buckingham, England: Open University Press.
- Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 501-518.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.
- Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352-373.

Received January 16, 1991

Revision received May 24, 1991

Accepted June 13, 1991 ■