Long-Term Retention of Sequentially Organized Information, Knowledge, and Skills: An Empirical Review

by

K. Anders Ericsson and Robert J. Crutcher

Department of Psychology

and

Institute of Cognitive Science

University of Colorado
Campus Box 345
Boulder, CO 80309-0345

ICS Technical Report #88-14

October, 1988

Abstract

This report reviews empirical studies of long-term retention and attempts to assess generalizable characteristics of performance as a function of the length of the retention intervals. First the results from laboratory studies of retention for lists of nonsense syllables and meaningful texts are discussed and a framework for methodological issues and critical procedural variables are proposed.

Using this framework long-term retention of three types of performance studies testing memory acquired under real-life conditions are reviewed. First long-term retention of procedures, such as cardiopulmonary resuscitation, are examined. Then studies of long-term retention of knowledge, such as knowledge acquired in college-level courses, are described. Finally, studies of long-term retention of acquired skills, such as typing, are reviewed.

In a concluding section an attempt is made to explore how the retention of these different types of performance can be studied and described in a single conceptual framework.

In the military and civilian sectors of our society there is an ever increasing demand for personnel trained to perform according to new procedures or to operate complex equipment. Considerable research has been directed towards making this training efficient. Often the training is followed by daily encounters with the new procedures or the new knowledge, and there is no problem with retention of the newly acquired information.

In many circumstances, however, opportunities for performance after training are either delayed or infrequent. Primary and secondary schools and colleges educate students with the expectation that knowledge is acquired on a long-term basis and remains available indefinitely. Another major area, emergency procedures have a low probability of occurrence; but when they do occur, accurate recall of knowledge and successful execution of the procedures are crucial. A related area concerns experts, such as doctors and pilots, who have to master a broad range of knowledge, parts of which may be needed only rarely.

The goals of designing instructional and training procedures for skills and knowledge that must be maintained during longer periods of disuse might be quite different than the goal of simply attaining a standard of mastery rapidly and efficiently during training. It is critical to have accurate information about how skills and knowledge decay during periods of disuse. In this paper we will attempt to review available data on the retention of information, especially for durations longer than a couple of weeks.

Outline of the review

It is very difficult to review studies that measure the retention of knowledge and skills. The number of studies using long-term retention intervals is relatively small, in part a function of the difficulties of getting subjects to return for retention tests without losing too many of them to systematic selection. If retention performance is only available for systematically selected sub-groups, inferences to populations of normal subjects are made difficult or even impossible. There are large differences in the types of information that are acquired for long-term retention and in the amount and type of training between retention situations and these are not the only differences. On the other hand the available knowledge about retention after long intervals is quite limited, and an attempt to systematize it is sorely needed.

We will start by providing a general conceptual framework for classifying different variables that may influence retention after delay. We will then identify the most controlled case for studying retention, which appears to be the old laboratory research on memory. In that context we will briefly review some of the factors affecting retention and the methodological issues related to measuring these factors. Using some of the generalizations from this research as background, we will then review retention measurements in more realistic contexts.

General Conceptual Framework

The pure study of retention would consist of three test phases and two principal phases as illustrated in Figure 1.

Insert Figure 1 about here

The initial test in Figure 1 simply assumes that the subject does not really possess the target skill or knowledge prior to study and training. In some studies performance after initial training can be substituted; in others, one can argue a priori that performance must be at a zero level. The acquisition phase contains all learning and training activity as well as nontraining activity when the acquisition requires several sessions spread over a longer interval of time. At some point the training is terminated, either after a fixed amount of practice or when a specific standard of performance is attained. In the latter case an additional, final test is not necessary. Otherwise, it is important to obtain some measure of the final performance at the end of training. It is not realistic to test all aspects of complex skills and large bodies of knowledge, so the final test covers a selected subset.

Ideally the retention phase is completely free from cognitive activity relating to the acquired information. In real life it is likely that situations similar, related, or analogous to the training conditions are encountered and that the acquired information can either help or interfere with the acquisition of new and different knowledge or skills. The knowledge that this information could be or would be retested might influence subjects' spontaneous rehearsal activities.

Finally, at the end of the retention phase a retention test is administered. Such a retention test may often be the first part of a retraining phase, where the performance level acquired earlier is reacquired.

Some of the sources for estimating retention after disuse will only loosely fit the general scheme described above. For example, studies of retention of knowledge acquired in schools often lack information about initial performance before the relevant course started. Similarly, studies of disuse of skills often poorly document the acquisition phase. For some skills the acquisition phase includes years of active use of the skill, which makes comparisons with skills acquired during limited periods difficult.

It is necessary to distinguish at least three different types of acquired information. The first type will be referred to as complete information. In this case the final test evaluates the subjects' mastery of all the information to be acquired. Often this information is acquired perfectly to a specified criterion. The second type is knowledge of a domain. For this type the amount of information to be acquired is large, and the acquired information is estimated by testing the subject on a subset of that information. The inferential aspect of the measurement of retention distinguishes this type from the first.

The third type concerns skills. For our purposes skill represents efficient performance in a large number of situations in which the speed of responding is crucial. In the acquisition of skills subjects are normally not taught knowledge about the task or strategies for dealing with the task but they must generate and discover this information themselves.

Measuring retention under laboratory conditions.

In preparation for discussing problems with studying retention in realistic situations, we will review studies that have attempted to measure retention under laboratory conditions. A description of the theoretical considerations underlying the study of retention under optimal conditions will clarify some of the problems of measuring retention in realistic situations. We will also review some confounding factors that have been demonstrated in laboratory studies and that need to be considered in any study of retention. We will begin with studies that measure retention of unrelated materials and then turn to studies of meaningful materials, such as texts.

Retention of "meaningless" materials

At the time Ebbinghaus was conducting his pioneering research on memory, it was assumed that all information retained in memory could be represented as a large number of associations between elements of different strengths. The fundamental question of forgetting could then be resolved by studying the loss of strength in these associations as a function of time. Under the plausible assumption that this decay was invariant over associations with the same strength, it was thought possible to describe the decay as a mathematical relationship between strength and time.

Ebbinghaus (1885/1964) attempted to estimate the rate of decay by creating new associations of a given strength in memory. He created a pool of meaningless units, or nonsense syllables and memorized sequences of them until he could recall all the items in the series. He assumed that a new sequence of associations of the same strength were created in this way. After some delay Ebbinghaus (1885/1964) then tried to relearn the same sequence to the same criterion of mastery. The amount of savings or retention can be measured by finding the difference between the number of trials necessary for the original learning and the number necessary for relearning, and then calculating the proportion of this difference to the original learning. The average proportion of savings as a function of time between the original learning and relearning is given in Figure 2.

Insert Figure 2 about here

As can be seen from the arrangement of points along a line the reduction in retention follows an approximately logarithmic function of retention duration. The relation in Figure 2 appears as a linear function as the delay interval is given in log units of time. Theoretically this mathematical function expressing the rate of forgetting associative connections should be replicable. Retention functions from several other studies (Finkenbinder, 1913; Luh, 1922; Radosavljevich, 1907) are also illustrated in Figure 2. These studies used procedures quite similar to Ebbinghaus' to examine groups of subjects rather than a single subject. The overall similarity of the retention curves is striking. However, the differences were sufficiently large to stimulate critical review of differences in detailed procedures and of individual differences in subjects.

One of the most important conclusions from this research appears to be that retention levels at longer intervals can be fairly well predicted from retention levels observed at shorter intervals. The significant costs associated with measurements of retention at very long time intervals do not seem to be warranted. Consequently, most of the subsequent research that has analyzed the factors influencing retention have used much shorter intervals.

Several factors influencing retention or rate of forgetting were demonstrated in Ebbinghaus' (1885/1964) original study. By continuing to practice a list after reaching the original criterion of 100% correct recall, Ebbinghaus (1885) found that savings the next day (24 hours later) was roughly proportional to the amount of additional time used for memorization beyond the criterion. Several studies have replicated the effects of additional memorizing or overlearning (Krueger, 1929; Luh, 1922), but the relation between the amount of overlearning and savings was not reliably replicated (Woodworth & Schlossberg, 1954).

Ebbinghaus (1885/1964) also demonstrated that memorized sequences of different lengths yielded different amounts of savings. The accepted account for this effect is that with longer lists, a larger proportion of the associations are overlearned while the remaining items that have not yet been learned are still being acquired. Subsequent studies using lists of paired associates have validated this account. In these studies any paired associate was eliminated from the study as soon as it was correctly recalled. Under these conditions savings is not a function of the length of a list, that is, of the number of paired associates (Hovland, 1951).

Finally, Ebbinghaus (1885/1964) demonstrated that when a list is relearned to the original criterion, the subsequent rate of forgetting is reduced. Later research has extended this finding by showing that study distributed over time often yields better retention than when the same amount of study time is expended in a single interval (Hovland, 1951).

Generalizability of different measures of retention

Following Ebbinghaus' (1885/1964) pioneering study of memory, investigators started to explore measures of retention other than the number of repetitions required for relearning.

There are several disadvantages to the relearning method. For short retention intervals with little forgetting, the relearning method is not a very sensitive indicator of forgetting (Luh, 1922). When the amount of information to be memorized is large, substantial time and effort are required to obtain data about relearning. The relearning method is applicable only to material that has been completely acquired by the end of the learning and study period. Finally, a critical problem of the relearning method is that it confounds the speed of learning with estimates of retention. If the speed of learning increases in the interim, retention will be overestimated. Considerable increases in the speed of learning nonsense material have been observed in many studies and their results are reviewed in Ericsson. (1985).

There are several alternatives to the relearning method for assessing retention. One approach is to examine the number of correctly anticipated items on the first test-trial. Similar information about retention prior to relearning can be obtained by having the subject recall as much as possible about the list learned earlier. A different test of the subject's memory for the learned list is to ask the subject to recognize the items of the learned list among a number of other items. In early studies, which obtained all of these measures of retention, the overall shape of the forgetting function was remarkably similar although the level of retention indicated by the different methods was quite different (Luh, 1922). Normally the recognition test gives the highest estimates of retention, the recall method gives the lowest estimates, and the relearning method gives intermediate estimates (Woodworth, 1938).

Recognition, the most sensitive method for measuring retention, has several major drawbacks. The most significant of these is the difficulty of selecting foils or items which should be correctly rejected as not part of the list. The more similar the foils are to the items on the list, the harder the recognition test. Bahrick and Bahrick (1964) have shown that recognition tests can be constructed that produce lower retention scores than recall tests do. In general, recognition tests are a sensitive indicator of some features of retention although it is difficult to infer from these tests the amount and type of information that is retained in memory.

The amount of information a subject can recall is the most common measure of retention and the measure that most closely corresponds to common sense views of the amount of information retained. The main problem with recall is that it tends to underestimate the amount of information retained. The subjects have to generate their own retrieval cues, a task that is both effortful and time-consuming. Subjects may therefore differ in their ability and motivation to engage in such retrieval, and these differences can dramatically influence the amount recalled. Even when subjects are motivated, it is often difficult to determine how long subjects should engage in recall. For large amounts of information acquired some time before the test, significantly more information can be recalled even after the first hour of recall that is offered during the test (Williams & Hollan, 1981).

If asked to recall the same information on repeated occasions without any opportunity for further study of presented information, subjects will later retrieve presented information that they did not recall previously. The experimenter can reduce many of these problems by giving the subject cues to permit recall of a specific piece of information or of a constrained set of information.

Summary and discussion of memorization of meaningless or unrelated material

The research on nonsense syllables and unrelated material has shown remarkably consistent exponential decay as a function of delay. The functional form of the decay is similar across different measures of retention, but the measures indicate different levels of retention. Other factors influencing the amount retained, such as the distribution of practice and the amount of overlearning, do not appear to change the rate of exponential decay but simply the level of retention. Let us now consider how these results can be extended to cases of meaningful learning and memorization, in which subjects' prior knowledge mediates the encoding and retrieval of presented information.

Subsequent research has conclusively demonstrated that Ebbinghaus' (1885/1964) nonsense syllables are differentially difficult to learn. Hence, acquiring any syllable in the list cannot be seen as a new association. Difficulty of learning can be accurately predicted by measures of how easily subjects come up with word associations in response to a given syllable. This index is referred to as a degree of meaningfulness. Actual monosyllabic words would have an extremely high degree of meaningfulness. Some retention studies have used the same acquisition and retention procedures with monosyllabic words as have been used with nonsense syllables.

Although highly meaningful items are learned faster than items low in meaning, retention, at least for short delay intervals, can be equated by special learning procedures (Postman, 1976). In these learning procedures, acquisition is equated for all items by varying the number of learning trials each item receives, with fewer trials for more meaningful items and more trials for less meaningful items. It may be that meaningful materials encoded in terms of subjects' preexisting knowledge do not show the same exponential decay as the memory for nonsense syllables previously reviewed. This possibility is intriguing and will be discussed in the next section of this paper. From Postman's (1976) account we would expect the initial level of retention to be higher for meaningful encoding with higher "strength" possibly exhibiting a slower rate of exponential decay.

In research published shortly after Ebbinghaus' (1885/1964) monograph, Mueller (Mueller & Pilzecker, 1900; Mueller & Schumann, 1894) demonstrated that memorization of lists of nonsense syllables involved much more than strengthening associations between items. Using standard experimental techniques, Mueller produced evidence for the grouping of items and the encoding of the spatial location of items. In a recent review, Ericsson (1985) showed that many early studies that found significant improvement in memorization of lists with practice imply the emergence of efficient, meaningful encoding and retrieval structures for encoding the serial order of list items. Chase and Ericsson's skilled memory theory (Chase & Ericsson, 1982; Ericsson, 1985) accounts for improvement with practice as well as for observed interference effects from previous and subsequent memorization of other lists.

Retention for unrelated material has been found to be remarkably influenced by events that precede the learning episode as well as by events that follow it. In a classic paper, Underwood (1957) demonstrated that a large number of studies showed a consistent reduction in recall of a memorized list as a function of the number of previously memorized lists. This phenomenon, proactive interference, is shown in Figure 3.

Insert Figure 3 about here

The same reduction in recall as a function of the number of previously studied lists was found by Murdock (1974) when he studied three subjects who memorized a different list on ten successive days and then tried to recall the corresponding lists 22 hours after initial memorization. Hence, lists memorized earlier dramatically interfere with the retention of a list memorized later.

Learning that occurs after the memorization of a given list will also interfere with retention of the material learned first. This effect is called retroactive interference. The amount of interference increases as the material learned subsequently increases in similarity to the original material. The magnitude of these observed interference effects has led some theorists to argue that forgetting can be completely accounted for by such interference occurring over time. A detailed review of the extensive research on interference effects is given by Postman (1971).

In sum, there is no qualitative difference between meaningless and meaningful material. The differences can be better described as the degree of meaning or rather as the availability of meaning and meaningful associations.

The lists and paired associates previously discussed can be differentiated from texts and other meaningfully organized material in that the former lacks inherent organization and is often described as unrelated. The cognitive structures or mental cues, that is, retrieval structures, used by subjects to encode the order of the presented material are used over and over and hence lead to considerable proactive and retroactive interference. Let us now consider laboratory studies of the retention of texts and other material with inherent organization and structure.

Laboratory studies of retention of meaningfully organized material

In this section we will first discuss retention as a function of delay for meaningfully organized materials. We are particularly interested in evaluating the generalizability of the exponential decay function identified for the unrelated materials. Then we will discuss various efforts to characterize the organization of information, especially in texts; and we will review studies of differential forgetting for items with different organizational attributes. Finally, we will discuss the effects of interference and overlearning for meaningfully organized material.

In the laboratory, meaningful material has generally meant short texts. The problem of assessing retention for text is complex because in addition to differences among recognition and recall measures, establishing a measure of retention is difficult. For example, a test of retention of text can target either verbatim retention (i.e., memory for the actual wording) or retention of gist (i.e., memory for ideas).

As Welborn and English (1937) pointed out some time ago,

> It seems fairly clear, then, that just as verbatim learning of unconnected but meaningful material is easier than the learning of nonsense material and in turn as connected material is easier to learn verbatim than unconnected material, so the "ideas" are easier to learn than the exact words. (p. 2)

The retention studies in Figure 4 support this conclusion.

Insert Figure 4 about here

The Briggs and Reed (1943) study, which measured retention of paragraph content, shows very little decline over delay, a result we might expect because the study measured retention of content or "ideas." The gradual decline is further explained by the method of testing. Students were allowed as much time as necessary to read and re-read the passage, and the study used a true/false test. Furthermore, while reading the text, the subjects answered the same questions that appeared later on the delayed retention test. Dietz and Jones (1931) asked subjects to read a factual prose passage but restricted them to a single reading and employed a 5-item multiple choice test. Moreover, the test consisted of 100 to 125 items (the Briggs and Reed tests had 50 items), and subjects were instructed not to guess. Not surprisingly, the retention function here declines more rapidly, though the general forms of the curves are quite similar. McGeogh and McKinney (1934a) measured verbatim retention of poetry over various delays from 15 to 120 days. As we would expect, the rate of decay is more rapid than for the prose passages but not quite as rapid as for nonsense material.

An important issue is whether the retention functions for meaningful versus less meaningful materials differ in form or simply in rate of decline. Most researchers have emphasized the relative similarity of exponentially declining retention curves for meaningful and meaningless materials. However, they have noted differences in the rates of decline, with the functions for meaningful materials usually declining less rapidly (Davis, 1935; Hovland, in Stevens, 1951; Woodworth & Schlossberg, 1965). One way to examine this issue is to log transform the retention times for the various retention functions to see whether, like the Ebbinghaus curve, they can be fit with a straight line. Figure 5 shows the log transformations of the retention functions in Figure 4.

Insert Figure 5 about here

For the most part, the data seem to fit a log-linear function, just as the Ebbinghaus data do. Importantly though, as we move from gist retention of text to verbatim retention of text to verbatim retention of nonsense syllables, the increasingly negative slopes indicate a potential relationship between organization (or meaning) and rate of retention.

Before we discuss in more detail research efforts to relate organizational variables to rate of retention, we must discuss a methodological problem with all studies purporting to compare retention rates for different types of information or materials.

Studies of the effect of organizational variables on retention must carefully distinguish between an acquisition phenomenon and a retention phenomenon. In order to determine whether a variable specifically affects retention independently of acquisition, researchers often try to equate the level of acquisition for different materials or types of information. As Underwood (1949) suggests, however, this is difficult to do. For example, assume that two lists, two texts, or two aspects of a text are learned to the same criterion but at different speeds (see Figure 6).

Insert Figure 6 about here

Measures of retention for both types of material seem equivalent, but the test itself constitutes an additional learning trial, which moves the subjects one step further along the acquisition curve. Given the steeper slope of the curve for the more meaningful material, the level of acquisition is now higher for the meaningful material. One way to compensate for this problem is to slightly lower the criterion for learning the more meaningful material.

An alternative is not to try to equate retention for the two variables but to consider whether the retention rates for the two interact with the retention interval. This unfortunately confounds the variable of interest with degree of acquisition; still, the interaction can indicate a difference in the retention rates. For example, if material that is acquired to a higher level of retention decays at a faster rate than material acquired to a lower level, the interaction would seem to indicate a difference in the retention rates.

Even if acquisition for different types of material can be equated on the basis of some gross performance criterion such as number of items correct, equivalence of acquisition is still difficult to assert. It is certainly possible that two individuals might exhibit equivalent performance on retention of a text yet have acquired quite differently organized knowledge structures. This suggests that a more detailed analysis of the organizational attributes of the text as well as of the knowledge base is important in accounting for either acquisition or retention of text.

Organization

Research with text or prose is no longer concerned with the individual word unit; text is composed of individual words, but these words are organized into sentences, and sentences in turn are organized into paragraphs, and so on. Furthermore, the concepts or ideas represented by the words are organized by the text structure as well as by domain knowledge. Thus one of the continuing aims of text research has been to describe the structure or organization of text and to relate this organization to the processing, acquisition, and retention of text.

The retrieval structure for text is much richer than that for less meaningful information because there are multiple ways in which the information is organized. Information is organized in terms of text structures such as theme or plot as well as in terms of content. Structural retrieval cues can be used to access textual information just as they can to access a digit from a number matrix. The content of the stored information can also serve as a retrieval cue. Moreover, as we will describe, because there seem to be many coexisting organizational descriptions of a text, the number of available retrieval cues is much greater for text than for unrelated information.

Text retention studies have often attempted to show differential retention for different categories or types of information. Often these categories of information are arranged in some sort of hierarchy. Retention can then be predicted on the basis of position in the hierarchy.

One such approach has been to describe texts in terms of propositions and the pattern of connection among these propositions (Kintsch, 1974; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). In this scheme, a propositional hierarchy is established by first selecting the topical proposition or propositions and then constructing successively lower levels of the hierarchy on the basis of referential coherence or of overlap in the arguments of propositions. A coherence graph is constructed such that the main propositions occur at the top of the graph and the subordinate propositions fan outward in successive levels below the main level. (This model has been substantially revised in Kintsch, 1983, but we need consider only the older part of the model.)[1] This <u>levels effect</u> of the model predicts that the probability of recalling a proposition is a function of its level in the hierarchy (Kintsch & Keenan, 1973; Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; Kintsch & van Dijk, 1978). Figure 7 shows retention as a function of level.

Insert Figure 7 about here

In addition to hierarchical structuring at the local level, in which the propositions are called micropropositions, the model posits a more general or macrostructural level, which characterizes the gist of the text. This macrostructural level is also composed of propositions that are created from micropropositions by deletion and generalization processes. Kintsch and van Dijk's model predicts that the rate of decay for the micropropositions will be much faster than that for macropropositions: in fact, the forgetting rate for micropropositions in their 1978 experiment, in which they tested text retention over 2 delay intervals (1 month and 3 months), was four times greater than for macropropositions.

Few studies have compared retention for microstructure and macrostructures over longer delay intervals. An exception is a study by Singer (1982) that compared memory for natural and laboratory reading of newspaper articles over delays of 1, 14, and 42 days. As in the Kintsch and van Dijk study, macropropositions were retained better than micropropositions at all retention intervals (see Figure 8).

Insert Figure 8 about here

However, the rate of loss of micropropositions was not faster than that for macropropositions. In fact, a log transformation of the retention functions shows that there is very little difference in the retention rates (see Figure 9).

Insert Figure 9 about here

Notice, however, that both retention functions seem to be fit by a straight line when the retention times are log-transformed. This suggests that the form of the retention functions is very similar.

In short, Kintsch and van Dijk's results suggest differential rates of retention for microstructure information and macrostructure information whereas Singer's results suggest that the rates of retention are not essentially dissimilar. Both studies suggest, however, that retention for the different types of information decays exponentially and that macrostructures are retained better than microstructures at all delay intervals.

Two other related approaches to describing and attempting to relate text structure to retention are story grammars (e.g., Thorndyke, 1977) and thematic structuring schemes (Dooling & Lachman, 1971). Thorndyke has shown that the story structure itself (e.g., setting, theme, plot) provides a hierarchical structure that can aid retention of propositions. Thorndyke proposes a story grammar consisting of a set of productions for parsing a story into its narrative structures. The successive application of these productions results in a hierarchically organized representation of the story. Propositions occurring in the top levels of the hierarchy are retained better than those in lower levels (see Figure 10), just as in Kintsch and van Dijk's model.

Insert Figure 10 about here

Unfortunately, because Thorndyke did not test retention over extended delay intervals, there is no evidence either way for the differential decay hypothesis.

Dooling and Christiaansen (1977) and Christiaansen (1980) have attempted to describe prose in terms of different levels of abstractness and to test the hypothesis that different types of text information have different retention rates. Unlike Kintsch and van Dijk (1978), they used a forced-choice recognition procedure to assess retention instead of a recall procedure. They argue that the forced-choice recognition procedure prevents differences in output strategies over time from influencing retention performance. They tested four types of information: sentence format, sentence gist, paragraph theme, and main character, which were presumed to vary along the dimension of abstractness from least abstract (sentence format) to most abstract (main character).

Their results also provide evidence for a hierarchy of retention, with the more abstract, higher level information better retained at each of the retention intervals (see Figures 11 and 12).

Insert Figures 11 and 12 about here

There was no evidence for differential retention rates, however. Figure 12 is especially helpful in conveying the results, as retention time is plotted on a log scale. Once again, each of the retention functions is a straight line. However, here there seems to be no evidence for differential decay of information.

What can be concluded from these long-term retention studies of text? It seems clear that higher level information is usually retained better than lower level information in a text. It also seems from the few studies we have been able to locate that the retention functions for all types of text information conform to the exponentially declining retention function for less meaningful material, though the rate of decline may not be as fast as for more meaningful material. However, it is not clear from the little evidence we have whether the rate of decay consistently varies as a function of different organizational characteristics.

The approaches we have just discussed for describing the structure of text are only a few of the many that have been proposed to account for differential retention of text information. Unfortunately, many of the other structural descriptions have not been tested over delayed retention intervals, so they provide no evidence concerning differential rates of retention. Nevertheless, many of these other schemes for describing the organization and structure of text could be incorporated into long-term retention studies, and we would like to describe two of them in more detail.

One approach is described in two studies by Graesser and his colleagues (Graesser, 1978; Graesser, Robertson, Lovelace, & Swinehart, 1980) in which the representation of texts describing common procedures was investigated. Subjects read a text description of a common procedure, such as catching a fish, but did not carry out the procedure themselves. Graesser suggests two structural hypotheses to capture the text structure: (1) a hierarchical structure hypothesis in which a sequence of actions in a procedure is a product of a hierarchically structured arrangement of plans; and (2) a relational structure hypothesis in which the degree to which an action is related to other actions is measured. It would have been ideal to derive a canonical tree structure graph that represented the actions along with the intermediate constituents of the plan; but because there was no objective criterion for constructing such a canonical representation, statements were scaled along the two organizational dimensions by means of a question-probe technique. That is, for each of the action statements in the passage subjects answered a why-question (e.g., why do you dip the rag in soapy water?). The answers to the why-questions were taken to refer to superordinate plans. Thus actions could be classified according to the distinct reasons generated by why-questions.

Relational density was estimated by the distribution of reasons. Statements were assumed to have high relational density if they shared many reasons in common with other statements and low relational density if they shared few reasons in common with other statements.

Graesser found that both hierarchical structure and relational density predicted retention of action statements and that there was no interaction between the two. Thus both organizational structures seem to predict retention.

In another series of experiments (Trabasso & Sperry, 1985; Trabasso & van den Broek, 1985) the relationship between structure and memory was investigated in a different way. Here the emphasis was not on hierarchical organization but on causal and logical relatedness. Trabasso maintains that the importance of a statement in a narrative is a function of the reader's apprehension of the conceptual dependencies of the statement. These dependencies take into account antecedents, consequences, and implications. Importance, and therefore memory, depend on the relations between a statement and other statements. These relations are identified in two ways. First, the relationship of the statement to other statements is identified: Is the statement part of a "causal chain"--that is, of a series of successive causes and consequences that extends from the story's opening to its conclusion? Second, the relation of a statement to other statements is measured by the number of direct operative links a statement has to other statements. This is similar to what Graesser (1978) called "relational density".

To determine the causal and logical relations between statements, Trabasso uses a procedure different from Graesser's why-questions. Trabasso's procedure specifically provides a logical, analytic, and a priori basis for identifying causes and consequences in narrative text and identifies the number of connections as well as the causal chain. Causal relationships between events are established by a criterion of "necessity in the circumstances" (Mackie, 1980, cited in Trabasso & Sperry, 1985). Necessity is determined by counterfactual reasoning: an Event A is necessary to Event B if, in a specific set of circumstances, the nonoccurrence of A means B would not have occurred. The results of these studies show that both structural variables (connections and causal chains alone) account for significant, high proportions of the variance in retention, with no significant interaction between the two variables.

Graesser's and Trabasso's studies along with the few long-term retention studies we have discussed show that meaningfully organized information in text is differentially retained after an immediate reading as well as after further delay. Most texts can be viewed as a hierarchical expansion on some theme or topic. During the process of comprehension, subjects are able to extract that organization, which in turn is highly predictive of their retention as measured by recall. The information presented in the texts differs in how well it corresponds to the subjects' prior knowledge. In some cases part of the presented information can be derived through causal inferences from partial memory for the antecedents of an action or a procedure. As might be expected, tightly related information produces better retention during immediate test and delay.

Given that the information presented in the text is stored and encoded according to the hierarchical organization of the content of the text, retrieval at recall will use that structure as retrieval cues for accessing the information. If information is stored as relations between specific objects or propositions in memory, text that is subsequently presented may not produce the interference observed in the retention of nonsense syllables.

Interference

Some early studies suggested that the influence of proactive and retroactive interference does not operate for meaningful material as it does for meaningless material (Ausubel, Robbins, & Blake, 1957; Hall, 1955; McGeogh & McKinney, 1934a; McGeogh & McKinney, 1934b). McGeogh and McKinney (1934a, 1934b), for example, compared retroactive interference of different types of material on retention of prose and poetry. In the poetry study, they found that the interpolated learning of nonsense syllables had about the same general effect on retention as did the interpolation of poetry when recall occurred after 15 minutes and after 7 days. Retroactive interference at 7 days increased when the 15-minute test was omitted. In both cases, however, the amount of interference was not nearly as great as the retroactive interference for nonsense material. Subjects recalled 80% or more of the lines after 15 minutes.

In the prose study, the pattern of results was fairly similar although the interpolated learning of nonsense syllables was followed by a smaller decrement in the recall of the prose than in the recall of the poetry. McGeogh and McKinney found evidence for retroactive interference, but it was small.

Hall (1955), in another study, tested the effects of retroactive interference of interpolated prose material varying in similarity with the learned material. In this study, all groups read 30 sentences, which described the characteristics and customs of a hypothetical primitive tribe and which were presented one at a time. Each group was presented with a different type of interpolated material. For the control group, this material was a spatial relations test. The first experimental group read an additional 30 sentences, all of which described the characteristics and customs of another primitive tribe. These sentences were similar in form to the first 30 but as dissimilar in content as possible. The second experimental group also read 30 additional sentences, 20 of which were dissimilar to the original sentences and 10 of which were identical to the sentences they had previously learned except for the substitution of one new word. Subjects were not tested on the interpolated material. All groups were tested on the original 30 sentences with a completion test. For the ten sentences in which a new word had been substituted, this word was the word missing in the completion test. Twenty-one days later the groups were retested.

The results seem to parallel McGeogh and McKinney's (1934a, 1934b) results. Only small amounts of interference were generated by the interpolated material, and no significant differences in the amount of interference were produced by the three kinds of interpolated material.

Ausubel and his colleagues (1957) conducted a study in which three groups read a 1700-word passage on Buddhism and took a multiple choice test on the material. Twenty-four hours later, one group studied a 2100-word comparative essay on Buddhism and Christianity, another group restudied the original passage, and a third group studied a 1700-word passage on Christianity. A week later, each of the groups was retested with the same test. There were no significant differences between the groups. In another study (Ausubel, Stager, & Gaite, 1968) Ausubel and his colleagues found that a passage judged to be "similar but conflicting" actually facilitated memory. These two studies led Ausubel and his colleagues to conclude that retroactive interference does not apply to meaningful learned material as it does to nonsense material. In fact, under certain conditions interpolated material can facilitate retention of previously learned material.

Critics have pointed out problems with the earlier studies that can account for the lack of interference effects. For example, Myrow and Anderson (1972) claim that the following conditions are necessary for retroactive interference: "(a) corresponding stimuli in the original learning and interpolated learning must be similar while the responses vary; and (b) the retention measure must differentiate items on the basis of the similarity of stimuli and responses" (p. 303). Ausubel's experiments most likely failed to detect retroactive interference because the passages and tests contained about equal numbers of response-different and response-same items, which balanced one another and thus created a net effect of no retroactive interference (Myrow & Anderson, 1972). Myrow and Anderson also note that recall or cued recall measures are more susceptible to retroactive interference than recognition tests are. This may explain the difference between Ausubel's results and the results reported by McGeogh and McKinney and by Hall. Ausubel's study used a recognition test procedure whereas the McGeogh and McKinney and the Hall studies used a recall or cued recall procedure which produced at least some degree of retroactive interference.

More recent investigations have shown that retroactive interference operates just as lawfully for meaningful connected discourse as for nonsense syllables or paired associates, provided that the research carefully specifies the similarity between the original and interpolated material and the way this similarity is to be tested (Bower & Hilgard, 1981). Bower and Hilgard note that "one must attend carefully to what are the atomic assertions that relate concepts in the initial text and how the specific predictions about these concepts are altered in the interpolated learning" (p. 163).

Bower (1978, cited in Bower & Hilgard, 1981), for example, demonstrated clear negative and retroactive interference for sentences. Subjects first learned sentences of the form $S_1 - V_1O_1$ (subject, verb, object) and then sentences of either the form $S_1 - V_2O_2$ or $S_2 - V_1O_1$. In both cases there was associative loss between S_1 and V_1O_1 when measured by cued recall in either direction. Substituting a paraphrase for the V_1O_1 portion $(P(V_1O_1))$ and associating it with another subject $(S_2\text{-}P(V_1O_1))$ still produced interference, suggesting that the interference was at the semantic or conceptual level. It is important to note that some manipulations can produce interference at one level and facilitation at another level. For example, learning S_1 - V_1O_1 and later S_1 - $P(V_1O_1)$ facilitates recall of the gist if an S_1 cue is given because at this level the associations to S_1 are the same. However, verbatim recall for the first association (V_1O_1) will probably suffer if the cue S_1 is given because of the competing and, at this level, different association $P(V_1O_1)$. Similarly, experiments with texts have simultaneously facilitated recall of information at one level (e.g., the conceptual macrostructure or schema level) but interfered with recall of information at another level (Thorndyke & Hayes-Roth, 1979).

The more recent studies we have cited for interference effects on text retention have not employed retention intervals after long delays. We have not yet located any such studies. The earlier studies suggest that interference effects would be even stronger over longer delay intervals.

The seemingly contradictory results of interference studies on text retention apparently arise from the problems of assessing the similarity between the original and interpolated materials. Two items that are quite similar in their verbatim expression, for example,--"Peter sprinted to the store" versus "Peter ran to the store"--can interfere with each other if they are associated with the same cue and if the retention test forces the subject to differentiate between them. However, the same two items can be redundant if the retention test measures an aspect that is true for both of them, for example, "Did Peter go to the store slowly or quickly?" One implication of this correspondence is that a set of words or concepts become resistant to interference if the items are organized or integrated so that they are not similar to one another within the organized structure. One example would be a list of words versus a sentence. The sentence provides a syntactic and semantic structure that organizes the retrieval of the words.

Smith, Adams, and Schorr (1978) conducted a particularly interesting study related to this idea. In this study, subjects learned a set of facts about a person. The familiar interference or "fan effect" predicts the more facts one learns about a person, the slower retrieval will be because of associative interference. But as the results of this study show, if the facts about the person are integrated by relevant world knowledge, interference is reduced. The interesting questions for long-term retention studies are, What is the relationship of organization and interference to retention over different delay intervals? Can different organizations of a set of words or concepts result in knowledge structures less susceptible to interference?

Overlearning

Studies of the effects of overlearning on retention of prose or connected discourse have not yielded clear-cut conclusions. Either the retention intervals are too short or the measures of overlearning are questionable. Gilbert (1957), for example, tried to measure the effects of overlearning on retention of prose. Using a modification of Gillette's (1936) drop-out procedure for learning of paired associates, he tried to control for the degree of original learning of each fact unit in a passage in such a way that individual facts could be dropped from the passage as they were learned to the criterion. He employed three degrees of overlearning: 0%, 100%, and 200%, where 0% overlearning means that all questions had been answered correctly once. The three conditions yielded correspondingly different degrees of retention, but there were no significant changes in the slopes of the curves. Unfortunately, the retention intervals were quite short (15 minutes, 24 hours, and 48 hours).

Craig, Sternthal, & Olshan (1972) looked at overlearning of ads presented on slides. In this study overlearning was defined in terms of presentation trials. One hundred percent learning was established by first establishing the number of presentations necessary for subjects to achieve a 100% correct criterion for 12 print ads, each ad exposed 7 times for 5 seconds per time. Two hundred percent learning (100% overlearning) was 12 ads exposed for 5 seconds each and presented 14 times; and 300% learning (200% overlearning) was the same number of ads exposed 21 times. The retention intervals were 1,7, and 28 days, with different subjects used for each retention interval. There was an immediate retention test in which subjects wrote down as many brand names as they could recall.

The interesting result of this study was an interaction of degree of overlearning and retention interval (see Figure 13), with 200% learning producing better retention after 1 day or 28 days but 300% learning producing better retention than 100% learning at 7 days. At no point is 100% learning better than either of the overlearning conditions with either the shorter or longer delays. These results suggest that a moderate degree of overlearning can sometimes result in better retention.

Insert Figure 13 about here

However, the measures of overlearning creates a problem with interpreting these results. Simply extending the number of trials does not ensure the same degree of overlearning for all subjects. As suggested earlier, items may vary in difficulty, so that not all items are equally learned or overlearned. Furthermore, differences in motivational or attentional factors may vary considerably from subject to subject. The additional presentations of some ads may have received little processing or attention from some of the subjects.

The problems of measuring or determining the degree of overlearning, especially with meaningful material, may be considerable. We have already mentioned that some items may be studied more than others unless a drop-out procedure (Gillette, 1936) is used. However, problems may also arise from differences in the organizational strategies subjects use. When organizational devices are used in learning meaningful information or even in learning less meaningful information, the organization of the material may interact with the degree of overlearning. For example, in studies we have carried out with waiters memorizing dinner orders (Ericsson & Polson, in press), there are differences in the organization mnemonics subjects use to try to memorize orders. Some of these devices result in much faster acquisition of memory skill for orders. For example, strategies for clustering menu items by category seem to result in better performance than strategies for using cues such as facial expression to associate each menu item with the person ordering it. When measured by only a performance criterion, subjects using these two strategies may show equivalent levels of performance. However, there may be differences because of the organization employed.

This difference may also occur in subjects' memory for prose. One person may have an extensive background knowledge in a given domain and the other person may have very little. If both acquire information to a given performance criterion and the number of presentations or the number of correct responses is used as an index of degree of learning, there may still be differences in retention. An alternative or additional measure of overlearning is to measure the speed of recall. This measure reveals differences in the degree of learning not revealed by the other measures.

Extending the study of retention towards retention in real-life domains

Up to this point we have focussed on studies of retention for unrelated lists and meaningful text. In the following sections we will extend our analyses to retention observed in other domains. The primary goal is to evaluate the results of retention studies in these domains against the general principles previously described, and thus to extend the scope of these principles to provide a first-order description of human retention in general.

In the introduction we distinguished three classes of behavior for which retention has been studied with rather different methodological approaches. The first class corresponds to traditional memory studies, where retention can be measured through the reproduction of a fixed sequence of discrete, often verbal, behavior. The second class concerns retention of knowledge in a domain, where the amount of knowledge is so large that only a selective subset of it can be tested. The third class corresponds roughly to acquired performance and skill, where the acquired speed and efficiency of responding is an essential part of behavior.

Since the beginning of scientific psychology many investigators have observed much higher degrees of retention for motor skills than for other types of memory. The possibility that these differences were caused by the verbal nature of most traditional memory tasks was rejected by McGeogh and

Melton (1929). They compared retention of a finger maze, which is a motor task, to the retention of a list of nonsense syllables. Structurally these two tasks are very similar in that the finger maze requires storage of a number of discrete behavioral actions, where the relation between one action to the next is completely arbitrary. The measures of savings (time and errors) showed no essential difference in retention.

One of the most consistent differences in retention for different types of tasks is found for tasks requiring a continuous response rather than a series of discrete responses. Continuous responses required by such tasks as pursuit and tracking are usually retained much longer than the discrete responses in the series of discrete responses required in procedural tasks (Naylor & Briggs, 1961; Schendel & Hagman, 1980; Schendel, Shields, & Katz, 1978). Continuous movements are generally retained quite well over long-term retention intervals whereas procedural skills decay much more rapidly without regular practice. For example, the continuous aspects of flying skills are retained quite well, but the procedural aspects deteriorate over retention intervals of a few weeks or months (Schendel, Shields, & Katz, 1978).

A number of explanations of this difference have been proposed, including the verbal-cognitive nature of procedural tasks, the relative difficulty of establishing what constitutes a trial for a continuous task (so that overlearning is more likely), and differences in the way errors are scored--performance deviations may be more detectable in motor tasks involving a sequence of discrete steps. Some have suggested that motor tasks are more highly overlearned than procedural tasks.

We tend to agree with those who have emphasized that the problem in remembering most procedural tasks is in remembering not how to do the task

(Hurlock & Montague, 1982) but in recalling the correct step from a large group of possible steps (e.g., which lever to push and when) and in recalling the correct sequence of steps. Thus within our framework, we wish to distinguish between procedural skills or tasks and other skills in the following way. Studies of procedural skills are more closely related to studies of knowledge retention or of retention for meaningful material because remembering knowledge is critical to selecting or remembering the correct step at the correct time in a sequence. Thus, initial task performance and subsequent retention performance is quite similar to memory tasks in which a unit, for example, a nonsense syllable or a word, must be distinguished from a set of possible units. This means that the degree of retention of procedural tasks may will depend on the arrangement and relationship of the individual task steps. Skill, on the other hand, always implies a time component, and retention performance is tied to the proficiency or speed with which a person can execute the task. This distinction implies that if a procedural task is so highly overlearned that the speed of performing the correct response sequence, not the selection of the correct response, is the retention bottleneck, then the retention characteristics of the task should be more similar to those of motor skills. However, in most of the studies of procedural tasks that we will consider, the tasks have not been learned to a high level of proficiency, and the performance criterion is the successful selection of a sequence of responses. Thus we expect the retention characteristics of these tasks to be more similar to those of knowledge retention in general. In a section on retention of skill we will closely examine whether the general characteristics of retention can be extended to retention of skill as well.

In three sections we will discuss the characteristics of retention after delay for memory for procedures, memory for knowledge of domains, and memory for acquired skills

Memory for procedures and other fixed sequences of actions

The retention performance for procedures is both similar to and different from the retention performances previously discussed within our conceptual framework. First, the type of performance measured at retention involves all the acquired information. In this sense, procedural tasks resemble laboratory retention tasks (i.e., retention of lists or texts) rather than retention of knowledge in a domain. Some procedural tasks more closely resemble laboratory retention tasks of meaningless or unrelated information; others resemble laboratory tasks for retention of meaningful information. Many procedural tasks are acquired in natural settings such as the classroom, the military, or industry where there is less control over the conditions of acquisition. For example, for many procedural tasks learned by soldiers in basic training, there is little control over how often the task will be subsequently practiced.

Retention of CPR

The ability of a wide range of professionals to administer cardiopulmonary resuscitation (CPR) rapidly and effectively is literally vital to society. CPR is a clear example of a procedural task. Two interesting studies of retention of CPR training indicate drastic decrements in successful administration of CPR after 6 months and 12 months of disuse (Shields, Goldberg, & Dressel, 1979; Weaver, Ramirez, Dorfman, & Raizner, 1979).

In a comprehensive study Shields et al. (1979) analyzed the retention of CPR over retention intervals of 6 to 12 months. The retention at delay could be related by regression analysis to actual time since training. The statistically significant slopes of decay for the different procedures are illustrated in Figure 14 for percentage of complete, correctly recalled procedures and in Figure 15 for the percent of correctly performed steps.

Insert Figures 14 and 15 about here

It is tempting to view these results as evidence of a linear decay of retention of these procedures. A closer examination of the Shields et al. study (1979) shows that retention performance was analyzed only for delay intervals of 5 to 12 months. The clearest evidence in favor of exponential decay over linear decay would emerge in a simultaneous analysis of shorter and longer retention intervals. More conclusive evidence against a linear decay of performance is Shields et al.'s (1979) finding that "several of the tasks had significant quadratic components" (p. 17). The existence of quadratic components are at least consistent with an exponential decay hypothesis. Furthermore, many of the estimated intercepts, especially for the completely correct recall, deviate from 100%, another finding that is consistent with the exponential decay hypothesis. If these data are available for reanalysis, the hypothesis of exponential decay can be evaluated statistically.

Two additional studies provide data on retention of CPR. McKenna and Glendon (1985) tested 124 occupational first aiders on their ability to carry out CPR at varying times following training (up to 3 years). Their results indicated rapid decay in CPR skills over time, with fewer than 20 percent of the subjects obtaining a score of 75 percent on performance after only a six month delay.

McKenna and Glendon concluded that the decay function was linear (see Figure 16). The total score (17 points) was a composite of three scores: performance (8 points), technique (5 points), and diagnosis (4 points). The performance score measured the crucial part of CPR performance (i.e., the ratio of 15 compressions to 2 breaths). This score, along with the total score, is represented in Figure 16, and, consistent with the previous studies, indicates that certain components of the CPR procedure are retained much less well than others. The two remaining scores, technique, which represented general skills (e.g., adequate lung inflations), and diagnosis, which represented accurate diagnosis of absence of breath and carotid pulse prior to commencing CPR, following precordial thump, and 1 minute after CPR, are reported separately in Figure 17. Though the authors interpret their results as supportive of a linear decay function, their interpretation is questionable.

Insert Figures 16 and 17 about here

First, their claim for a linear decay function is based on the fit of a regression line to the total score decay function. However, this total score is a composite of the other three scores, only one of which (performance), does not seem to be an exponential decay function (see Figure 18). We can see from plotting the four sets of scores on a log scale, the technique and diagnosis scores seem best fit to an exponential decay function. Even the performance score shows exponential decay from 2 to 18 months and only the final point, at 36 months, deviates from an exponential decay function. If one compares the performance score function with the total score function, it is clear that the departure from an exponential decay function is due almost solely to the drop in the performance score.

Insert Figure 18 about here

It is worth discussing the performance score in somewhat more detail. McKenna and Glendon describe this score as follows:

> This score (out of 8) was for the crucial parts of CPR performance, i.e., the first four ratios of 15 compressions to two breaths. Thirteen to 17 compressions were awarded one mark if they were correctly positioned and of the correct strength. A delay of up to seven seconds between each set of compressions was accepted.

From this description it is difficult to determine how the performance score was assigned, though it sounds as if an individual who performed within an acceptable range (e.g., the 13 to 17 compressions) would receive a point for each of the four cycles of compressions and breaths, whereas someone outside the range would receive no points. Such a scoring procedure could quite easily produce a threshold effect, such that within a specific delay interval performance would decline but remain within the acceptable range; however, after a longer delay, though performance continued to decline at the same rate, it may have now been outside the acceptable range (e.g., 18 compressions instead of 17), so that no points were received for each of the four cycles. This interpretation seems quite plausible given the exponential decay of performance over the first three delay intervals of 2, 6, and 18 months, and the sudden drop in retention over the fourth, very much longer, 36 month delay interval.

It should also be noted that although McKenna and Glendon asked for information regarding practice during the retention intervals and found no significant correlation between practice and any of the dependent variables, they did not specify how this practice variable was measured. In fact, they themselves admit that a number of variables, including practice, were not adequately controlled for in their study.

A study by Latman and Wooley (1980), which tested emergency care attendants, emergency medical technicians, and paramedics on retention of knowledge and skill proficiency for various first-aid tasks, including CPR, suggests how important it is to control for practice over the delay interval. Latman and Wooley found knowledge retention showed very little decay over a 24 month period, approximately 10% for each of the groups, whereas skill proficiency loss was much greater. Given the large amount of training (24 to 400) hours and that the written tests were almost completely mutltiple choice items, the high retention level for knowledge of CPR seems consistent with other studies of knowledge retention (see the section on knowledge retention). However, consistent with all of the CPR studies we have discussed so far, skill proficiency showed significant decay. Most importantly, Latman and Wooley's results show that the amount of practice inidividuals engage in during the delay interval is significantly related to the level of skill retention (see Figure 19).

Insert Figure 19 about here

Organization and Structure

We have alluded to the similarity between procedural tasks and the retention of verbal material. Procedural tasks consist of a series of discrete steps that must be performed in a given sequence. The difficulty in retention is to remember which step to perform at a given time. To the degree that the steps of the task are unrelated to one another or to the task as a whole, remembering the individual steps ought to be quite similar to trying to remember a list of nonsense syllables in order. Thus the decay function ought to be fairly rapid.

However, to the degree that the individual steps of the task are related to one another or can be organized into large structures or related to the overall task, the retention function ought to be more similar to that found for meaningful materials such as texts or lists of organizable words. At one extreme would be procedural tasks in which none of the steps has any relationship to the other steps, and the steps would have to be learned by rote. At the opposite extreme would be procedural tasks in which the top level structure would completely constrain the selection and order of task steps. Most procedural tasks are probably somewhere in between, the task steps exhibiting some degree of schematic organization and some of the task steps being more related than others.

A number of studies (e.g., Goldberg, Drillings, & Dressel, 1981; Hagman, 1980; Hagman, & Rose, 1983; Osborn, Campbell, & Harris, 1979; Shields, Goldberg, & Dressel, 1979) have suggested that overall performance on many procedural tasks learned by military personnel is often low because of only one or two steps in the procedure. A common feature of these poorly retained steps is that they are memory dependent: that is, they are usually unrelated to the task as a whole, not related to or cued by the previous sequence of steps, or not suggested by the equipment or task.

Shields and her colleagues (1979) studied a number of different skills and their decay rates. They looked at overall task performance, the average percent of steps performed correctly, as well as the type of performance steps missed in each task. A high Gutman coefficient of reproduceability for each of the tasks showed that subjects consistently missed the same steps and that these steps were consistently the most difficult steps. These steps were highly memory dependent and were not suggested by the equipment or by previous steps. For these reasons safety procedures were frequently missed; for example, a soldier forgot to clear his weapon.

Shields' study shows that it may be possible to specify the components that make a procedural task more likely not to be retained. Her results demonstrate that specific, memory-dependent steps account for much of the retention loss. The question is, how to determine which steps are memory-dependent. We need a much more detailed understanding of the organization or structure of the task before we can predict which steps are likely to decay.

Shields et al. (1979) also found that subtasks were often not performed and that the number of steps in a task was a reliable predictor of overall task performance: tasks with fewer steps decayed less rapidly. In addition, the earlier soldiers had been trained in a task, the better they retained it. Some of the tasks had been taught earlier on in basic training and possibly received more practice. Shields and her colleagues were able to produce a regression equation that related differences in the rate of decay for overall task performance to three variables:

$$Y = -.25X1 = 1.5X2 - 2.5X3 + C$$

where X1 is the number of steps in the task, X2 is the order of original training (1, 2, 3, or 4), and X3 is the presence or absence of subtasks (0 or 1). The best predictor, accounting for 25% of the variance, was the number of steps. Because the order of training was completely confounded with certain tasks, this parameter is completely uninterpretable and serves only the function of a covariate. However, the number of steps as a predictor of retention seems consistent with laboratory results that imply that longer lists of unrelated items are more difficult to remember.

Shields et al. (1979) point to rather systematic omissions in steps across subjects and memorized procedures and argue that "task steps that are either unrelated to hardware and/or the previous sequence of steps" (p. 4) are most likely to be omitted and hence should receive more training. For example, in the CPR task one such step was the 15 compressions to 2 breaths ratio. They represent the CPR procedure as a sequence of 14 steps (see Table 1), each of which was evaluated at training and at the subsequent retention test.

Insert Table 1 about here

The systematic omission of unrelated steps and steps lacking perceptual cues from the situation and hardware suggests that cognitive representations may be required for reliable retrieval of such steps. The results of the CPR study by Weaver et al. (1979) strongly support this view. In addition to evaluating the retention of the actual procedure, Weaver et al. (1979) also tested knowledge related to the reasons for the different actions. They found that the best predictor of CPR retention was related to this understanding.

Overlearning

Consistent with other studies, retention studies of procedural skills suggest that an important variable related to retention is the degree of original learning (Hurlock & Montague, 1982). Laboratory experiments have shown that increased repetitions can improve original learning and subsequent retention, and therefore increasing the amount of training for a complex task should also increase the degree of original learning. In a study on learning a 52-step procedure for testing alternating electrical output, Hagman (1980) varied the number of repetitions (either 1, 2, 3, or 4) of the task during training. Both accuracy and speed of performance measures were improved by each of the first three repetitions, but there was no additional gain in performance after the fourth repetition.

A number of studies suggest that overlearning, or as it is commonly referred to in the skills literature, mastery training, enhances procedural skill retention and slows decay (Goldberg, et al., 1981; Hagman, 1983; Schendel, et al., 1978; Schendel & Hagman, 1980). As with studies discussed in the knowledge retention section, however, these studies measure overlearning in terms of number of repetitions, repetitions beyond proficiency, or number of complete correct performances (i.e., all steps correct).

In a study by Goldberg et al. (1981), crewmen were trained to boresight and zero the main gun of an M60A1 tank. One group was trained to a single correct performance (standard training); the other to three correct performances (mastery training). The mastery training group did significantly better than the standard training group after 1- and 5-week retention intervals, although most of this difference occurred in performance on the first retention trial. This finding suggests that with minimal practice or testing opportunities, the performance of groups with standard training should equal that of groups with mastery training. Furthermore, only 15% of groups with mastery training were able to perform the task correctly on the first retention trial.

A study on mastery training by Schendel and Hagman (1980) examined the effect of mastery training received at the time of original training versus an identical amount of overlearning that occurred 4 weeks after the original training. At 8 weeks, both groups showed fairly equivalent retention for the task, assembly and disassembly of an M60 machine gun. One implication of this study is that mastery training might be more cost-efficient and useful than standard training, especially for tasks that are not too fatiguing.

Although these studies do not argue overwhelmingly for a mastery learning approach, there are other factors to consider in evaluating mastery training. For example, most of these studies have considered the effect of mastery training on only a few repetitions beyond the proficiency level.

More extensive amounts of overlearning have been studied with skilled pianists memorizing music (Rubin-Rabson, 1941). Around 11 learning trials were required for the first completely correct performance of the piano pieces. In addition, the pianists received 50%, 100% or 200% overlearning on the corresponding pieces, which amounts to a total of 17, 21 and 31 learning trials respectively. Relearning after two weeks showed an average savings of about 3.5 trials (35%) with no reliable difference attributable to overlearning. The same pieces were relearned again after an additional 7 months with a savings of around 40%. Again the differences between overlearning conditions are consistent with an effect of overlearning, but the differences are so small (always less than a single learning trial) that they do not even approach significance.

Taken together, these studies on overlearning indicate rather small if any improvement in retention after delay. This lack of effect of overlearning may be even more consistent if one discards studies using a very liberal criterion for complete acquisition, namely the first time the subject performs the sequence correctly. It is likely in such cases that the procedure is not completely acquired and that additional learning trials constitute continued learning rather than overlearning.

Long-term Retention of Knowledge in Real-life Contexts

In the studies of long term retention considered so far, the conditions of the acquisition phase have been fairly well controlled so that there is a definite acquisition phase, afterwards a period of disuse or of no further acquisition and then a retention test. Furthermore, these studies have for the most part measured all of the acquired information or complete information.

In learning a text, for example, there is generally a free recall test or a test that measures retention of all or nearly all of the ideas in the text. Testing to assess retention of knowledge in a domain is inevitably much different. It is not practical to test the learner's retention for the entire body of material learned in a course. Instead, the test must sample the knowledge base, and retention of the entire knowledge base must be inferred from the sample.

In trying to assess retention of knowledge learned in real-life contexts, a number of new problems present themselves. Control over the conditions of acquisition is more difficult because most of these studies occur in classroom settings and it is difficult to control how the information is acquired or to control motivational or attentional factors. It is also more difficult to control rehearsal activities after the acquisition phase because presumably the information acquired is more relevant to people's work or lives than is a text learned in a laboratory; and there are bound to be some people who engage in more rehearsal than others.

To control for these variables it is often necessary to use statistical methods (e.g. Bahrick, Bahrick, & Wittlinger, 1975; Bahrick, 1979, 1984). However, the validity of statistical control depends on accurate assessment of the variables. For example, subjects must be able to accurately estimate how often they have rehearsed the material since learning it.

The testing of retention of information from a single text is also clouded by obvious differences in subjects' background reading skills and familiarity with the topic. These differences can be somewhat controlled, however, by testing reading skill and selecting texts on topics unfamiliar to subjects. Naturally, there will still be differences in how subjects organize their understanding of the text based on previous knowledge.

These problems are compounded in studies of the retention of knowledge acquired in realistic contexts. These studies must acknowledge varying degrees of background knowledge that subjects bring to the acquisition phase.

Before we examine the long-term retention of knowledge through delayed tests, we need to discuss in more detail how studies have attempted to deal with several methodological issues. Unlike the retention of procedures described earlier, the acquisition and retention of knowledge can be assessed only through a selected subset of the corresponding knowledge. Some investigators have tried to avoid the problems of comparability by using the same test for both acquisition and retention. As we will show below, such retests, especially multiple retests, yield a rather different level of retention from the level indicated by parallel tests. In some studies very long term retention records of original acquisition performance are not available, and this performance has to be inferred.

Problems in inferring retention of knowledge from selected test items

In most laboratory settings, we can construct retention tests that test all or nearly all the information acquired. However, it is impractical to test a knowledge base such as knowledge of Spanish or psychology for all of the information learned, so the retention test must sample the knowledge base. Sampling introduces the possibility that two retention tests for the same material may yield different results. Most long-term retention studies have not provided much information about the retention tests given. However, some studies, for example, Watson's (1938) and Bahrick's (1979, 1983, 1984; Bahrick et al., 1975) studies, have provided detailed information about the retention items.

Another potential problem in tests that sample the knowledge base is that the sample may in some way not be representative and the questions may therefore be too easy or too difficult. In Watson's study (1938) different subjects received different but parallel versions of the same test. This approach to some extent avoids the danger that one specific form of a test may in some way be biased.

Bahrick's studies have apparently used only a single version of the test. Given the claim for permastore, it would be important to do an item analysis to be sure that what looks like a sudden drop in retention is not simply a small group of questions that are difficult to answer. Bahrick has claimed that his retention functions differ from the more common, exponentially declining retention functions found in short-term episodic memory experiments where a diminishing number of responses are lost in successive, equal intervals of time. He argues that the frequency distribution of responses is discontinuous. The retention function falls off exponentially for 3 to 6 years and then flattens for 25 years or longer, during which the frequency of responses with life spans within these limits is almost zero. Finally, after several decades there is an additional positive decline in the retention function.

It is in the third stage that an item analysis would be useful. Perhaps the decline is simply due to the difficulty of some test items--to answers that are not obvious or to questions that are unclear. The items showing little or no decay may be those that can be guessed at or inferred from general knowledge. Using different test forms to test the same information may help establish whether in fact specific types of information decay more readily than others.

An alternative account of Bahrick's lack of decline for very long intervals concerns possible cohort differences. Such differences due to historical change in the environment and in educational goals and practices are well documented in the study of intelligence. Baltes and Kliegl (1985) describe the systematic change in Schaie's (1983) intelligence profile of adults with birthdates ranging from 1890 to 1930. The observed changes in the intelligence components are consistent with a compensatory effect; that is, older subjects with the longer retention interval have higher values on the language performance components of IQ, such as word fluency.

Assessing the original level of acquisition

A number of studies (e.g., Bassett, 1928; Johnson, 1930; Kennedy, 1932) suggest the same conclusion we have found before, that the level of acquisition is related to the level of retention. Therefore, it is important to accurately assess the level of acquisition in a domain. The extreme difficulty of doing so in studies of knowledge acquired in realistic settings forces experimenters to resort to statistical control of different acquisition levels. As a result, studies of long-term retention of course material have often been popular. As many experimenters have pointed out (e.g., Watson, 1938), there are three advantages to measuring retention of course material. First, assessing the level of acquisition is simplified because testing is a normal part of most courses. There is also a slightly better chance of controlling or equating acquisition conditions for subjects. Finally, students' acquisition can be measured at the beginning of their stay at an institution, and they are still likely to be there after a delay of a few years or so. Using subjects from the same class at least ensures that the presentation of the material is equivalent, although it does not necessarily follow that acquisition is equivalent.

Many long-term studies of retention of course material are difficult to compare because of differences in acquisition variables. Many studies, for example, fail to report immediate retention scores (e.g., Frutchey, 1937); or lacking actual scores, try to infer what they were (e.g., Eikenberry, 1923; Myers, 1917). However, the basis for these inferences, including standardized test scores or grades are far from desirable. Some studies (e.g., Wert, 1937) provide retention scores as a percent of the amount originally acquired but do not report the actual amount of knowledge originally acquired.

Some studies (e.g., Johnson, 1930) include subjects from different classes but fail to ensure that subjects for different retention intervals achieved equivalent levels of acquisition or to statistically control for such differences. Finally, some studies (e.g., Greene, 1931; Johnson, 1930) have used data from atypical subjects, for example test scores of advanced students.

A 5-year retention study of psychology by Watson (1938) is one of the few studies that has carefully controlled for these factors and others as well. Watson tested a large number of subjects, all of whom had taken introductory psychology in their freshman year at various times in the past. Watson ensured that the retention scores for all the various delay groups were not significantly different from one another. He even rescored the tests to ensure the scores did not differ because the original graders had used different grading standards. The study also included a control group that had not taken any course in psychology. Such a control group is necessary for determining to what degree a test assesses retention as opposed to general knowledge or inferencing skills. (Bahrick (1984) also employed a control group that had had no previous instruction in Spanish.)

Watson's careful assessments to ensure the comparability of the different subject groups is admirable but not always possible. Bahrick's studies (1979, 1983, 1984; Bahrick et al., 1975), have involved extremely long retention intervals of up to 50 years and his subjects have acquired the information of interest under a variety of conditions in different settings. To equate these groups Bahrick has used statistical procedures that attempt to assess the original level of knowledge as well as the degree of rehearsal during delay. He then uses multiple regression to adjust the retention function. This adjustment process corrects for the inequalities among the groups. However, both the level of original knowledge and the degree of rehearsal during the retention interval must be estimated with acceptable reliability (Bahrick, 1979, 1984).

Bahrick assessed the original level of his subjects' acquisition of Spanish from their responses to a questionnaire concerning the number of years of Spanish they had taken and the grades and level of instruction they had received. He also obtained information about the type and amount of rehearsal his subjects had given their knowledge. This approach to establishing the original level of knowledge acquired does not provide actual retention data. Two individuals may both have taken Spanish and received an A in the course but they may actually have acquired different amounts of knowledge. Although often impractical in studies of such long-term retention, an initial retention test like Watson's (1938) provides a better measure of the retention level at acquisition.

Assessment of retention with retests or with the same test

We have noted that repeated testing, especially with the same test, does not give a pure assessment of retention because the learner will reacquire or reaccess the information during test, thus altering the memory trace for subsequent retention tests. Nevertheless, many studies have used the same test to measure long-term retention at successive intervals. Frutchey (1937), for example, did not clarify whether he tested the same or different subjects at 1-, 2-, and 3-year intervals or whether he used the same test on all occasions. Bassett (1928) used the same test to measure retention at successive intervals of 4, 8, 12, and 16 months. The relative flatness of the curve (see Figure 20) is most likely due to confounding the acquisition and retention curves. Each of the retests is an opportunity for relearning or reaccessing information thereby reducing the decline of retention. Anderson and Jordan's study (1928), which measured long-term retention of Latin words and phrases, showed the same pattern of results as the Bassett study did (see Figure 21). Repeated testing with the same test at 1 day, 1 week, 3 weeks, and 8 weeks resulted in a flattened retention function.

Insert Figures 20 and 21 about here

These studies, then, cannot be considered pure retention studies because retention and acquisition have been confounded. During the retention interval there should be no reacquisition or reaccessing of the material that was learned. If the interval is punctuated by repeated testing, it should be considered an extended acquisition period, and the interval between the next to last test and the last retention test should be treated as the actual retention interval.

Another problem with repeated testing using the same test is that it confounds the retention of the material being tested with retention of the test itself. Subjects may remember the unusual wording of a question or something unrelated to the content of the domain and may not reaccess previously learned information at all. The actual measure of retention is thereby inflated. For example, Landauer and Ainslie (1975) found an extremely high level of retention for material retested at 6 months and then at 12 months as compared to material not retested until 12 months later. However, the actual effect of the retesting on retention cannot be determined because retention for the test items and retention for the material learned were confounded.

Measures of Retention

Relearning measures of retention have to our knowledge seldom been used in long-term retention studies of course material. Undoubtedly the amount of material involved makes it infeasible to have students relearn an entire course. Relearning measures therefore seem more useful in measuring retention for material when retention for complete information is possible. Recognition and recall measures, however, are quite common in studies of retention for course material.

Numerous researchers have pointed out that the curves for retention of course material may be quite different depending on which measure of retention is used (Bahrick, 1979, 1983, 1984; Davis, 1935; McGeoch, 1942; Watson, 1938). For example, Figure 22 displays the separate retention curves for recognition and recall in Watson's long-term retention study of psychology. All the retention scores have been transformed and are expressed as a percentage of immediate retention, which is set at 100% for each test in the study.

Insert Figure 22 about here

Immediate recall, ranging from 50 to 60 percent on the three tests, dropped rapidly over the first 6 months and then more slowly until it reached about 10 percent for all three tests after 30 months. Immediate recognition, on the other hand, began at about 70 percent for all three tests and dropped off much more gradually. Moreover, the recognition scores after 54 months were more divergent than the recall scores, ranging from 20 percent for Test 2 to 44 percent for Test 3. All three tests were actually part of the same test but covered different material: Test 1 covered the second semester of psychology, Test 3 covered the first semester, and Test 2 covered a special reading by Allport that was not part of the regular classroom instruction. The higher retention level for Test 3 all along the retention curve may be due to the repetition or reaccessing of material from the first semester during the second semester of study. There is also a large difference between the curves for recognition and recall measures in Bahrick's (1984) long-term retention study of Spanish (see Figures 23 and 24). Despite these differences between recognition and recall measures, many long-term studies of retention for course material (e.g., Eikenberry, 1923; Greene, 1931; Johnson, 1930; Layton, 1932; Worcester, 1928). either fail to specify the exact retention measures or else indiscriminantly combine various measures into a single retention measure

----

Insert Figures 23 and 24 about here

Careful inspection of Figure 20 may account for some of the differences in the retention curves for these studies. Bassett (1929) repeated tests consisting of 40 true-false items, 20 multiple-choice items, 10 completions, and a choice of 3 out of 4 essay questions requiring short, four-sentence answers. The total possible score was 80; so clearly the test was predominantly true-false and multiple choice questions (about 4:3 of T/F and multiple choice to all other items) with no correction for guessing. The composition of this test combined with Bassett's retesting of the same subjects over the many different delay intervals explains the extreme flatness of the curve. In Greene's zoology test, which consisted of 113 completion items, 125 true-false items, and 104 multiple-choice items the predominance of recognition over recall items is still about 2:1. For Greene's psychology course there were 81 completions, 23 multiple-choice items, and 34 labeling items, and thus a greater number of recall items than recognition items. The retention curve for Johnson's botany study is somewhat lower yet, although in his test, which consisted of 172 completion items and 126 true-false items, recall items again predominated. Obviously, without the tests themselves, there is no way of knowing what differences there were in the difficulty of items or what other factors might account for the lower level of retention for the curves involving greater recall. Nevertheless, these data, taken together with Bahrick's and Watson's results, clearly suggest that recall follows a sharper decline and remains lower over long retention intervals.

Similarly, in studying long-term retention of Spanish, Bahrick found that recognition performance was maintained at a higher level than recall performance. Furthermore, the direction of search (i.e., from English to Spanish or vice versa) did not make much difference in the retention functions for recognition. In the cued recall task, on the other hand, the

Spanish-to-English direction resulted in better recall retention at all retention intervals.

Bahrick also noted important differences in performance on the recognition subtests by the control group that had received no instruction. On the vocabulary subtests, the performance of the control groups significantly exceeded chance success, whereas their performance on the idiom recognition subtest was well below chance. This difference is explained in part by the opportunity in the vocabulary test to employ guessing, knowledge of other romance languages, as well as information about Spanish that had been acquired incidentally. However, performance on the idiom recognition test was depressed because some of the foils contained English cognates for portions of the Spanish idioms.

The analysis of this study shows why caution and care must be used in comparing recognition results as well as other retention measures. The importance of no-instruction control groups cannot be overemphasized. The selection of foils should be made available so that researchers can judge the relative difficulty of the objective questions used in different studies.

Structural and Component Variables

Our analysis of retention studies for meaningful materials such as text suggests the increased importance of organization and structure in attempts to account for retention. Clearly retention is related to the organization of a text, although the details of the relationship are not yet clear. Organization may be even more important in the retention of large bodies of knowledge than in the retention of limited amounts. A number of older as well as more recent long-term retention studies have suggested that differences in retention relate to different kinds of information. Some studies have attempted to specify how retention rates for different kinds of information relate to the way the information fits into an organized structure or body of knowledge.

A number of studies of course material have shown differences in retention for different types of information. For example, Bassett (1928) investigated which test questions were forgotten most often. She found that history knowledge involving the names of persons and places is more likely to be lost; for example, the names of prominent men are confused with one another. She also found that 7th-grade children apparently could not comprehend or retain the abstract and technical phraseology of documents and government. Bassett further concluded that history knowledge that is concrete and personal tends to be well retained, as is information about heroes and episodes that have received considerable emphasis in teaching. Unfortunately, Bassett does not say how these categories were determined.

A few studies have tried to characterize differences in retention in terms of different types of knowledge, the definition of types being orthogonal to specific subject domains. Glassnap, Poggio, and Ory (1978) studied retention of material in an educational measurement course. Though interested primarily in mastery versus nonmastery outcomes, these researchers compared retention effects for different types of knowledge classified according to Bloom's taxonomy (1956). Retention curves for the various types of knowledge are shown in Figure 25. The higher-level items category is a conglomerate of the three highest levels of Bloom's taxonomy. Although Bloom's taxonomic system may not be the one to use in analyzing how different types of knowledge decay, some such analysis of knowledge may be quite useful for this purpose.

Insert Figure 25 about here

A few studies (Anderson & Jordan, 1928; Bahrick, 1984; Kennedy, 1932) have examined retention of language or vocabulary and tried to assess how different types of information decay. Kennedy's study looked at retention for Latin syntax and concluded simply that those principles of syntax best known initially were also those best remembered at succeeding test periods. These principles included the accusative case, place to which, ablative of accompaniment, and the predicate nominative.

Anderson and Jordan (1928) also studied retention of Latin but considered retention of Latin phrases and words. Latin phrases or words were presented with their English equivalents. Latin words were classified into groups according to the similarity in sound between the Latin words and their English equivalents.

In the first group, the words sounded almost identical to their English equivalents; for example, provincia and its equivalent province. In the second group, called "association words," the Latin words were dissimilar in sound to their English equivalents but similar to closely related English derivatives. These derivatives were placed to the right of the English equivalents for students to study if they wished. An example of an association word is lingua. Its English equivalent is language, and its similar derivative is lingo. The third group, referred to as the "non-association words," consisted of words that bore no similarity in sound to their English equivalents, nor could they be related to a similar-sounding derivative. The retention results for the different word groups and phrases are shown in Figure 21.

All of the curves are rather flat, undoubtedly the result of the repeated testing procedure, in which students were given the Latin words or phrases and supplied the English equivalents. Nevertheless, the relative heights of the retention functions suggest how different the retention for vocabulary items may be, depending on how similar in sound an item is to its English equivalent.

Bahrick's (1984) study of memory for Spanish measured several aspects of this knowledge, including recognition and recall for grammar, vocabulary, idioms, and word order. Figures 23 and 24 show these retention functions. Recall curves for vocabulary recall and grammar fall off more rapidly than those for other types of information. Retention of word order also declines quite rapidly after about 10 years. Idioms, along with recognition of vocabulary and grammar, show a much slower decline in retention. In fact, recognition of vocabulary remains fairly constant from 3 years to 25 or 30 years. Bahrick suggests that the reading comprehension retention function does not decline because it is maintained on the basis of recognition, vocabulary, and grammar. In another study (1983) he investigated memory for cognitive maps of a city and used subjects who either were or had been students at a university in the city. He found that retention losses were greatest for recall of street names, less for the recall of city buildings, and least for the recall of campus landmarks.

A group of studies on retention of zoology and chemistry suggest how a knowledge domain might be characterized in a general enough way to make predictions (see Figure 26). Greene's retention study of zoology does not identify which aspects of zoological knowledge are retained and which decay most rapidly. Wert (1937), on the other hand, broke down zoological knowledge into the categories shown in Table 2. Though it is somewhat difficult to determine the precise nature of the individual tests, clearly the information labelled "other facts" was retained best of all the types of information. According to Wert, these other facts were principles and other generalizations of zoological significance. The lowest retention was for naming animal structures, the most specific information asked for. Applying principles to new situations and interpreting new experiments showed not a loss but a gain in retention.

Insert Table 2 and Figure 26 about here

Interestingly enough, another study of zoology shows a very similar retention pattern. Tyler's (1933) retention scores are reported as the percentage retained of gain. In Table 3, these scores have been converted to percentage retained of material learned. Unfortunately, Tyler provides only a single 15-month retention score. Nevertheless, the pattern of scores is similar to that in Wert's study. Naming animal structures is at the bottom, application of principles and interpreting new experiments show a gain in retention, and the other scores fall in between.

Insert Table 3 about here

Although we do not wish to make too much of this comparison because the categories are not described in detail, it is interesting how similar they in fact are. Moreover, two additional studies of retention for chemistry also seem to support this pattern. In Table 4 the results of Frutchey's (1937) study of chemistry show the highest retention for application of principles and the lowest retention for chemical terminology, in which students were required to identify the chemical term from a definition or description. This task seems quite similar to the task of naming the animal structures. Finally, in still another study of long-term retention in physical science courses, Freud and Cheronis (1940) found that "only those facts which are tied together by ideas remain; the others are lost" (p. 292). They also concluded that "the subject matter which is best retained consists of principles, theories, their applications, and related facts. That which shows poorest retention is that which is memorized without the exercise of reason" (p. 293).

Insert Table 4 about here

Is there some way to generalize across these studies to conclude something about the content or structure of the knowledge base and to determine how to predict what will be retained and what will be lost? These studies suggest that certain types of knowledge in a domain are more likely to decay than other types. General principles or higher-level information is retained better than lower-level information. However, none of these studies of course material has tried to describe the structure of knowledge in any great detail and relate the structure to the retention function.

Other studies, however, have dealt with short-term retention of course-like material acquired in more controlled settings and have tried to clarify the link between structure and retention (e.g., Eylon & Reif, 1984; and Hannigan, cited in Bransford, 1979).

A study by Chi (1983) has gone even further in trying to characterize the structure of knowledge in a domain. Chi chose to study a child's knowledge of 40 dinosaurs. This is a limited knowledge base, but the knowledge was not acquired in a structured setting, and retention was measured over an interval of a year. Using verbal protocols, Chi was able to construct a semantic network of the child's knowledge of dinosaurs. She compared the child's knowledge for a lesser-known group of dinosaurs with his knowledge of a better-known group. The two knowledge sets differed in terms of the number of direct and indirect links among concepts about dinosaurs, the strength of linkages, and the particular pattern of intra- and interlinkages that delineated the cohesion of the groupings. Furthermore, retention performance after a year strongly favored the dinosaurs in the group that exhibited better structure.

Chi's results suggest promising uses for verbal protocols and other methodologies designed to describe the structure of the knowledge base in detail. Perhaps these methodologies can be used to determine how to predict the course of long-term retention functions from the characteristics of a given knowledge structure.

Long term retention of acquired skill

In this third and final section retention of skills will be discussed. Our use of the term "skill" is quite general and refers to acquired performance for some well-defined task domain. A critical element in most skills is the efficient organization and speed of the corresponding behavior. In evaluating someone's ability to perform mental calculations with 2- and 3-

digit numbers, the time to produce the correct answer is essential. For many perceptual motor skills, like accurately hitting a fast ball in baseball or juggling, the speed of the behavior and mediating information processes can be inferred from the successful performance itself.

The diversity in task structure and in the types of measures available for evaluating the level of skill for a task creates a number of problems for a review of retention for such skilled performance. Before turning to the actual review of studies of retention we must establish a framework for discussing studies of retention for different skills.

A framework for measuring level of skill

We will start by searching for a theory-free way of measuring the retention of skill and later discuss such a method in light of the current knowledge about the organization of skill.

The only method of measuring memory retention that is applicable to skill is the relearning method. With one constraint, that the learning and the acquisition of the skill can be segmented into equivalent trials, any measure of the skilled performance can be used to relate performance and amount of practice. Assuming a stable and monotonic relation between performance and practice during skill acquisition, the relearning method can be used to monitor the required amount of practice, that is, number of trials necessary to regain the original level of performance after some retention interval, as illustrated in Figure 27.

Insert Figure 27 about here

The standard measure of retention is $(T_o-T_r)/T_o)$, where T_r is the number of trials necessary to regain performance and T_o is the number of trials necessary to obtain the performance during original acquisition. In situations where complete relearning is not obtained, a modified measure of retention can be used. Under the assumption that performance in relearning will mimic performance in original learning, one can find the number of trials (T_a) in the original learning necessary to obtain a performance matching the initial performance of the delayed retention test. The corresponding modified measure of retention is then $[T_o-(T_o-T_a)]/T_a$, or T_a/T_o. The modified measure will, of course, normally yield much lower estimates of retention than the traditional measure.

Recent reviews of studies of skill acquisition show that improvement in performance can be described by monotonic functions of amount of practice (Newell & Rosenbloom, 1981). In fact, Newell and Rosenbloom show that a power function provides a remarkably good fit for a wide range of different skills. They propose a theoretical account in terms of the learning of chunks for the consistent relation between performance and practice across these dramatically different tasks and skills. Their account, which uses a simple uniform mechanism, is very attractive; but it appears inconsistent with other general analyses of skill acquisition that emphasize different stages.

In systematizing a large body of data on the process of acquiring skills, Fitts (1964) proposed three different stages. During the "cognitive stage," the subject attempts to understand the task and its demands and to gather knowledge about what information to attend to. The "associative stage" involves making the cognitive processes efficient to allow rapid retrieval and perception of required information. The "autonomous stage" allows performance to be generated essentially automatically with a minimum of conscious cognitive processes. More recently Anderson (1982) has provided a theoretical model with three different learning mechanisms corresponding to each stage. From his analysis Anderson was able to derive the power law for relating performance to amount of practice.

If the description of skill acquisition in terms of three different stages is correct, it should have a major impact on any analysis of retention of skill. Large differences in retention should be expected as a function of the stage attained for the skill in question.

It is by no means easy to assess the stages of skill acquisition attained by subjects in reported studies. The number of hours of practice leads to very different progression in the skill acquisition process depending on the complexity of the task and subject's familiarity with it. Fitts and Posner (1967) mention that the cognitive phase requires roughly 10 hours for learning how to fly a propeller airplane. In comparison, the cognitive phase would be quite short for college students learning to perform mental multiplication because they have extensive experience of the paper-and-pencil version of multiplication.

Our review is organized around different types of skills, and as much as possible we bring together studies of the same or related skills. For example, typing and tracking are two categories of skill for which there are a reasonable number of independent studies of retention. We have organized our discussion in terms of the amount of cognitive learning that skills require. We will begin by discussing skills that are relatively simple and that subjects can understand and perform without serious errors after a couple of minutes. We will then discuss skills of increasing complexity.

One of the theoretical issues that pervades the following review is whether skills, especially motor skills, have a fundamentally different retention function from the exponential decay function found for the memory performance reviewed previously. There are certainly some fundamental differences between the experimental situations used to measure retention of knowledge and those used to measure retention of skill. Most skills require that some equipment or even a specific apparatus be measured. The subject needs to be seated in the experimental situation, which should provide many additional cues for retrieval of relevant information acquired during the original acquisition phase. In addition, many experimenters allow their subjects to warm up and reacquaint themselves with the equipment. Most situations provide the subjects with feedback about the accuracy of their explorative behavior, which should be therefore considered part of the assessed amount of necessary relearning. The clearest contrast to these cue-rich situations is having subjects freely recall some material. The free recall situation need not share any cues with the original situation involved in acquisition.

Even when relearning involves some equipment, this equipment may have been used for a number of different activities or for memorization of several different sequences. Some of the research discussed in the following sections reports on retention of the production of fixed sequences. These studies are particularly interesting because they study retention of the speed of producing a sequence as opposed to ability to reproduce the individual steps. In the following sections we will comment on the similarities and differences between skilled performance and the memory performance previously reviewed. This consideration will enable us to discuss in detail whether the general statements made earlier about retention can be extended to skills.

Simple skills

The most complete study of the retention of a very simple task is reported by Ammons et al. (1958). Their task consisted of manipulating 15 different control units in an order given by a sequence of schematic drawings of each unit. The control units were mounted on a panel and were quite different from one another. They included a knife switch, a toggle switch, an automobile turn indicator, an automobile door handle, a rotary switch, a doorbell buzzer, and a sliding door latch. A red light remained on as long as the subject manipulated each control unit correctly in the right order. Manipulating the control incorrectly or in the wrong order caused the red light to go out, and before proceeding, the subject had to correct the error to get the red light to go on again. The time to complete this task was the only performance measure analyzed. Six retention intervals were used (1 minute, 1 day, 1 month, 6 months, 1 year and 2 years) for two different training conditions. The first condition consisted of five training trials corresponding to a total of about 10 minutes of practice, including the 1-minute rest intervals between trials. The other training condition consisted of 30 training trials corresponding to about an hour. After the six different retention intervals, the subjects' performance on 10 retest trials was recorded. Unfortunately, 10 trials were not sufficient for subjects in the 30-trial condition to regain their performance after retention intervals of a month and more. In Figure 28 the savings score for the five-trial condition is shown.

Insert Figure 28 about here

The surprisingly low savings score for the 1 minute retention interval is an artifact of including the trial on which the final training performance was obtained or surpassed in the number of retraining trials. Although no formal statistical analysis was conducted comparing performance at the tenth and last relearning trial, the subjects in the 30-trial condition appeared to show better performance than those in the 5-trial condition across most retention intervals.

The most consistent differences between retention conditions were observed for the time to complete the <u>first</u> relearning trial. Ammons et al. (1958) measured the improvement in completion time during practice as the difference in completion time between first and last practice trials. They then computed how much of this improvement was apparent on the first relearning trial, as shown in Figure 29.

Insert Figure 29 about here

The systematic pattern of retention as a function of the length of retention interval and amount of initial training is interesting and striking. However, for retention intervals longer than a day the time of the first relearning trial was always longer than that of the second practice trial, with one exception. In the 30-trial practice condition the retention after a month was comparable to that in the third practice trial. From a pragmatic point of view the savings on the first relearning trial were hence rather modest.

Ammons et al. (1958) collected verbal reports in a pilot study to gain insight into what was learned. They found three stages.

> Reports were in agreement that the first few trials were used to learn to identify all the items and their correct sequence, as well as how to manipulate them correctly. The next step was to speed up the run still using the schematic chart as a guide. Finally, S learned to go quickly through the sequence from memory without looking at the chart. (p. 319)

Unfortunately, to substantiate the claim about stages Ammons et al. (1958) report only that errors were eliminated with further practice. However, they note that manipulation errors were nonexistant and that sequential errors were low at the first relearning trial. This finding is an interesting hypothetical source of the faster completion times for the first relearning trial as compared to those for the first practice trial.

The psychological literature is full of studies showing dramatic improvements in performance on "simple" tasks (Salthouse & Somberg, 1982). However, very few studies have evaluated the retention of such acquired improvements. One exception is a study by Salthouse and Somberg (1982), in which improvements on four tasks were monitored over 50 sessions for 8 old and

8 young subjects. Four weeks after the end of practice the subjects were retested on the four tasks.

One task involved learning and detecting a configuration of dots among alternative arrangements of dots, and another was a memory scanning task. A visual discrimination task involved detecting specified target elements in two displays presented either in sequence or simultaneously. Considerable and reliable improvement was observed with practice. After the 4-week delay reliable decrements in performance on the retest compared to the last two practice sessions were seen for only one task--the task involving detection of the configuration of dots.

It would not be correct to accept Salthouse and Somberg's (1982) results as evidence for perfect retention. In Table 4 some estimates of the lower bound on retention are given. These estimates are based on assessing at what trial number during original learning the performance at the 4-week delay was first observed.

Insert Table 5 about here

It is of course clear that the actual retention lies somewhere between the estimated, modified savings, ranging from 6% to 36% for the young and 34% to 58% for the old, and the 100% suggested by the lack of reliable differences.

Tracking skills

A large number of studies have measured the retention of acquired tracking skill for a wide range of different devices. We will first survey the pattern of retention across the different studies and then use different pieces of information from the studies to assemble a more complete picture of skill acquisition as well as the structure of resulting skill.

The conditions for comparing retention as a function of interval of disuse across studies are far from optimal because each study has problems and peculiarities. Table 6 summarizes the results on retention from the different studies. The savings score derived from most studies require some comment.

Insert Table 6 about here

None of the studies summarized in Table 6 reports retention in savings, and hence we have tried to calculate the savings from other information presented. Some of the studies (Battig, Nagel, Voss, & Brogden 1957; Fleishman & Parker, 1962, instruction group) report no reliable loss. In light of the huge improvement during acquisition for both of these studies, the savings score must be close to 100%. Unlike the other studies, Fleishman and Parker's study (1962) discarded the first three 1-minute trials as warm-up. This omission may lead to an overestimate of retention in this study, but the effects on the estimates of savings will be small. Reynolds and Bilodeau (1952) compared two training groups with massed and distributed practice. Because retention testing was done under a massed condition, the performance of the massed training group was included in Table 6. During the massed retention test subjects in the distributed practice condition showed considerably slower relearning than the original learning. This result strongly suggests that longer intertrial intervals benefit performance factors, such as concentration and motivation, independent of skill. We estimated savings for the Reynolds and Bilodeau (1952) study from the published average performance on acquisition and relearning trials.

From a large study of over 400 subjects in conditions with different intertrial intervals ranging from 1 minute to 30 hours, Bell (1950) recruited the 47 subjects in his 1-year retention study. Bell's published results refer to averages across all types of subjects. Hence, an unknown number of subjects in his study acquired their skill within a single session. The estimates of savings in Table 6 for Bell's (1950) study are based on the average performance during acquisition and retention trials.

Ammons et al. (1958) encountered some interesting problems in their large-scale retention study. In particular the group with less initial practice had much better retention performance during the first 15-minute block on relearning than during the last block of practice. Ammons et al. cite several motivational factors to account for this counterintuitive result. Fortunately their study included an immediate retention test, and they published the number of relearning blocks required for subjects with longer retention intervals to match the performance of subjects in the immediate retention condition.

The retention results in Table 6 are remarkably consistent across studies and can be readily summarized. Only studies with limited amounts of practice show any clear decrement in performance for retention intervals up to a year. With practice, retention intervals of two years show a relatively small loss in performance on the first relearning trials, and this loss can be eliminated within a single session of 20 minutes to an hour.

In an attempt to evaluate the effects of overlearning Ammons et al. (1958) also studied the retention of six subjects with 40 hours of practice after 1 year. Among the subjects in the 1- and 8-hour practice and 1-year retention interval condition, 6 subjects were selected with matching performance for the first hour of practice. The performance on the first 15-minute block of the retention test was no different for subjects in the 8- and 40-hour practice conditions and both groups were reliably better than that in the 1-hour practice condition. Hence, at least for the tracking task used by Ammons et al. (1958), no benefit from 400% additional practice could be observed. It should be noted that Ammons et al. did not present any evidence that performance on this task could be improved beyond the level attainable after 8 hours of practice.

The studies of tracking discussed in this section contain some information about the skill acquisition process and the structure of the resulting skill. In a pilot study Ammons et al. obtained verbal reports from their subjects during the learning phase. These reports indicated that subjects went through three stages of learning. First they learned the relations of the model plane and single control movements; then these control movements were coordinated into integrated patterns, which in turn were automatized into movement patterns.

Fleishman and Parker (1962) performed a number of analyses that support the idea of differences in the factors that determine performance at different points in the skill acquisition process. Performance of individual subjects during most of the first hour of practice (up to the 8th 6-minute block) is unrelated to their final performance (50th 6-minute block) with nonsignificant correlations of 0.13 or less. On the other hand, practice sessions towards the end of training correlated 0.70 with final performance. Furthermore, correlations between subjects' final performance (based on the last three practice sessions) and their performance on the retest session were between 0.80 and 0.98 for the different retest conditions. The combined correlation for final learning and retention performance for the 109 subjects with retention intervals between 9 and 25 months is 0.80. In comparison the correlation between retention performance and retention interval is only 0.30.

Fleishman and Parker (1962) also attempted to predict retention performance from test scores on another tracking test as well as from a paper-and-pencil test measuring spatial orientation. The correlations were positive but did not reach significance. Battig et al. (1957) explored the generalizability of the tracking skill by changing the movement pattern to be tracked. Although performance with the changed pattern declined, the performance was clearly superior to that in the first part of practice. Battig et al. (1957) also demonstrated performance clearly worse than that during initial practice by reversing the directions of the control stick so that the original direction of the control leading to an upward movement was changed to yield a downward movement, and vice versa.

Typing

Typing is the only domain in which several investigators have studied long-term retention for the same skill. A couple of studies have even examined very long retention intervals from 4 and 25 years (Hill, 1934, 1957; Hill, Rejall, & Thorndyke, 1913). These studies provide some of the best available evidence on the retention of skills, and it is worthwhile to attempt to systematize that evidence despite some major differences between studies. There are five studies, which cover acquisition of typewriting with a subsequent retention interval. Because all five originally reported a different measure for typing performance, these measures have been converted in Table 7 to typing speed in words per minute.

Insert Table 7 about here

In Table 7 the typing performances for the first training session and mean performance for the first five sessions are given for the five individual subjects. Although the estimates are averages based on sessions of 10 minutes (Book, 1925), half an hour (Swift & Schuyler, 1907), a full hour (Swift, 1906), or typing a 100- or 300-word paragraph (Hill, 1934; Hill, Rejall, & Thorndike, 1913), these estimates are remarkably comparable across subjects. The number of sessions, total time of typing practice, and the final typing performance at the end of training differ across studies. The length of the retention interval ranges from almost 3 months to 16 years. In Hill's (1934) study the interval between training and retest is 25 years; but during the first 4 years of that period the subject used the typewriter occasionally for typing letters, and on a single occasion 9 years after training he copied a 75-page manuscript.

In Table 7 the typing performance at the first retest session and the number of retest sessions necessary for attaining the previous performance level are given. From the observations two measures of retention were calculated: the standard measure consisting of savings in relearning and the percent of improvement in typing speed maintained at the first retest occasion. Except for Book's 1908 study (Book, 1925), all savings in relearning range between 80% and 90%. The maintained improvement shown in the first retest session ranges from 39% to 53%, and the highest retention is observed for the longest retention interval and vice versa.

On the basis of the 4 subjects, retention of typing skill appears stable over vastly different retention intervals. As for Book's (1925) study of his subject X, unlike the other four studies, the retest sessions were different and consisted of only 10 minutes of timed typing, whereas the training sessions consisted of 10 minutes of timed typing followed by 50 minutes of unmonitored typing. Hence, in 40 minutes (four retest sessions) Subject X attained the typing speed achieved at the end of training. When retested after an additional year without practice subject X attained a typing speed clearly higher than the speed attained during the original training. From Book's (1925) report on introspections during relearning it appears likely that Subject X had about half an hour of warm-up before the first real typing test.

Also noteworthy in Table 7 are the different typing and forgetting rates in Rajall and Hill's studies for the repeated typing of the same 100-word paragraph compared to the rates for the always different 300-word paragraph. The higher typing rate for same 100-word paragraph shows a higher loss at the first retest session, and the subject Hill required about twice the number of relearning sessions to reacquire that rate (Hill in Hill, 1934). At the age of 80 Subject Hill attempted to reacquire his typing skill after an additional 25 years without practice. The task of typing an unfamiliar 300-word paragraph essentially reproduced the old acquisition curve, and no clear evidence of retention is apparent. Retention is clearly evidenced for the fixed 100-word paragraph, and a 65% savings in relearning was observed.

Interpretation of these results is difficult because of the subject's explicitly mentioned problems with eyesight.

Retention of typing skill has also been studied using groups of subjects who have completed courses in typing. In all three studies to be discussed here subjects were interviewed at retention tests, and all subjects who had practiced typing during the retention interval were discarded from further analysis.

Schroeder (1934) retested 20 of his students in a two-semester typing course at the beginning of the second semester. He compared their typing performance at the end of the spring semester (average typing speed 27 words per minute) with their first typing test performance in the fall after 3 months of summer vacation. Schroeder (1934) found no difference in frequency of errors and found that more students had improved their typing speed than decreased it. These results should be interpreted with some caution because the test was conducted on the 4th day of the typing class. Although no organized typing practice had been arranged, students were allowed to try out the typewriters during this time.

Hagman (1979) retested 38 military staff members after they had completed a typing course with an average typing speed of 28.3 words per minute. The retention interval ranged from 14 to 38 days with an average of 25 days. The retention of typing skill was tested twice during the session. The first test occurred directly after a 5-minute warm-up. After 25 minutes of practice and a 5-minute break a second test was administered. Subjects' retention-test performance in net words per minute was 27% worse. Most of the decrement in performance was due to an increase (86%) in typing errors; the gross typing rate decreased only 12%. The 25 minutes of typing practice led to a reliable improvement in gross and net words per minute, and Hagman (1979) estimated that final typing performance would be attained with 2.5 hours of further practice. Length of retention interval (14-38 days) was not reliably related to decrement in typing skill. As Hagman (1979) does not report any acquisition curves, it is not possible to derive any estimates of savings and modified savings.

The most complete account of retention of typing skill is provided by Baddeley and Longman's (1978) study, which was designed to evaluate the effects of massed vs. distributed learning of typing. Four groups with different training schedules were compared. Training sessions varied between 1 and 2 hours in length, and groups had either one or two training sessions per day. The most effective typing acquisition was attained by the group that had a single 1-hour session per day. The groups that had two 1-hour sessions or one 2-hour session were no different. The least effective training was observed for the group with two 2-hour sessions per day. Each group consisted of 18 subjects, 16 to 17 of whom were available for a retention test. A third of each group was retested at either 1, 3 or 9 months' delay. The retention test consisted of a few minutes of warm-up, an initial 15-minute test session, some practice, and a second 15-minute test. The performance on the second test for the four training conditions at the three retention intervals is given in Figure 30.

Insert Figure 30 about here

The decrease in typing performance shown in Figure 30 is relatively slow and especially slow between 3 and 9 months of disuse. The last part of the acquisition curves provided by Baddeley and Longman (1978) permits estimation of modified savings only for the 1-month and 3-month retention intervals, which show modified savings of about 75% and 60% respectively.

The results of the group studies suggest more retention of typing skill at the first test occasion after long retention intervals than are shown in the results of the case studies reported earlier. However, all the group studies allowed for warm-up and prior practice before the initial test; thus the discrepancy between group studies and case studies can be resolved.

The studies just discussed used traditional performance measures and hence provide little or no information on changes in learning as a function of further practice. In his classic analysis of the acquisition of typewriting skill, Book (1925) proposed that acquisition consists of three major stages. The first stage is called the letter association stage, in which typing is performed letter by letter. During the initial part of this stage the subject concentrates on a letter, locates the letter mentally on the keyboard, moves a finger to the key, and finally strikes the key. Later in this stage, these individual steps are automatized. The second stage is called the syllable or word association stage, in which typing is done by syllables or words. Most of Book's evidence comes from the retrospective reports written by subjects after each test session as does the following example:

> In the very beginning of this stage (April 25th) Y reported as follows: "I get my copy almost entirely by taking one letter at a time as they are being written. The exceptions to this are the easy words and combinations which the fingers take care of as a group. In my actual writing my copy getting is a mixture of these two procedures--taking it letter by letter and easy words

as a whole, or (when distracted by something) looking ahead a little to get and hold it as in the visual method of writing." Word associations had already begun to form. Later (June 17), when many word associations had been developed he wrote in his special observations notes: "In trying to get more detailed introspections on just how I proceed with the copy and writing under these general conditions its seems that I am, in a very incipient manner, repeating the copy, perhaps the letters, as I write them. In easy short words I do not think that I incipiently pronounce letters but only words. The tendency to pronounce increases with the difficulty of the writing. The grouping that occurs seems to be determined all along by the nature of the copy. An easy word is taken as a whole; a few short easy ones are taken in a connected series. A long difficult word may be broken up into groups of easy, familiar combinations, or in case of extreme difficulty it may be incipiently spelled through letter by letter, and written by attending to each individual movement." (Book, 1925, p. 45)

The final stage is called the expert stage, in which the type writing is absolutely continuous. The units in attention are now clauses and phrases, as illustrated by the following verbal report. This report was collected from one of Book's subjects typing the same sentence over and over, but it is similar to those obtained for expert typists.

"When I write several words in succession in this way they run more or less together. The movements no longer are separated into groups according to the words. I am no longer conscious of

> the words or groups of movements which they represent but have my attention on getting through with the sentence as a whole and on a general control of the successions of movements required to write it." (Book, 1925, p. 54)

Unfortunately, Book (1925) did not provide much converging evidence based on performance to substantiate the reality of the three stages of learning. More recent work has provided some evidence. For example, Grudin (1981) has demonstrated that advanced typists cannot recall the spatial location of letters on the keyboard, whereas beginners are both rapid and accurate in their recall. An expert typist can, of course, reconstruct the location of the letters by typing out the alphabet, but that kind of recall is very different from spatially organized recall. Furthermore, Grudin (1983) has demonstrated that the types of errors made vary as a function of the level of typing skill.

For obvious reasons only a small number of studies have looked at retention of beginning typing skill. Efficient typing requires a rather extensive amount of practice, and although stopping short of efficient performance would be of great theoretical interest, it would be of little practical value to the participating subjects. Two studies by Freeman and Abernethy (1930, 1932) are rare exceptions. Their studies were motivated by the earlier findings of superior retention of motor skills compared to memory for lists. They wanted to compare typing to a task with the same structure but without the motor component. The typing task involved typing a short paragraph with the keys covered and a diagram of the keyboard visible to the subjects. The structurally similar task involved translating the same paragraph into digits by means of a translation key that listed all letters and the associated digits to be substituted. Each group of subjects acquired these two tasks to a criterion of two correct successive translations of the paragraph without access to the diagram or translation key. Subjects relearned the tasks after two weeks (Freeman & Abernethy, 1930) and after ten weeks (Freeman & Abernethy, 1932). The savings at relearning for the two tasks and the two delay intervals are given in Figure 31.

Insert Figure 31 about here

Figure 31 shows an interesting interaction in the retention of the two tasks, with the typing task being better retained. It is equally interesting to compare the retention of "typing task" to retention of typing skill discussed earlier. Retention is much better for the fully acquired typing skills even when measured as a savings in relearning. This result is at least consistent with the view that retention of typing skills at different levels of acquisition is different.

A brief discussion of some of the introspective reports obtained from Book's (1925) Subject X will conclude this consideration of retention of typing. Subject X relearned typing after a year's retention interval. The keyboard was shielded with a cover so that Subject X could not see the keys. The locations of only a small number of keys could be visually determined from the keyboard.

> The _t_, _h_, _g_ and _y_ were mentally located by the thought or sight of these letters in the copy calling up at once the muscular image of the movements required for striking their corresponding keys. As soon as these letters were thought of an

actual feeling of strain was noticed in the fingers which seemed to guide attention to the proper position on the keyboard. (Book, 1925, p. 102)

Some keys could not be located even by the associated motor movements to individual letters but were recovered from the memory of typing familiar words.

> "Before beginning to write," he says, "I tried hard to locate the 'd' key but without success. Just now came to the word 'day' and was startled to find the right finger on the 'd' key before I was conscious of what I had done. I had actually written 'da' before my past trouble in trying to locate the 'd' took possession of consciousness, when I stopped to write this note." (Book, 1925, p. 103)

However, the retrieval of the acquired motor patterns was quite rapid.

> "By the time the first thirty words of page 145 of Munsterberg's "Psychology and Life" had been copied, the position of all the letters had been so well relearned that any key could be struck directly as needed. All letter associations and most word associations had been fully revived." (Book, 1925, p. 104)

These introspective reports on the reacquisition of a skill are available from only a small number of subjects and lack supporting and converging evidence based on performance. Future research will be necessary to evaluate their intriguing suggestions.

Other acquired skills

In the large body of studies of skill acquisition only a relatively small number contain information about retention after delay. Several of the studies with results on retention involve only a single subject. It is our judgment, however, that these studies are important to any attempt to generalize across skills in providing information on widely different types of skills.

In a series of studies Swift (1903, 1905, 1910) studied the retention of a motor skill for durations up to 4 years without any intervening practice. The motor skill consisted of juggling two balls with one hand. The performance was measured by the number of times either of the balls was caught in a single trial of juggling. Seven subjects participated in training sessions consisting of ten juggling trials each. Two of these subjects participated in the long-term retention study, and data from their performance will be described in more detail. The primary subject, using his right hand, was given 48 training sessions, during which his performance increased from an a couple of catches to over one hundred catches per juggling trial. At this point he was trained to use his left hand. In the eight training sessions that this subject was given for using his left hand, his performance equaled that obtained after 38 sessions with his right hand, a savings of 78%. All subjects except one showed clear savings for learning with the left hand, indicating a considerable lack of response specificity. The primary subject also verbally described his method for juggling as "keep the balls in two parallel columns in front of him" (Swift, 1903, p. 204). After this subject was trained with his left hand, he was retested each month for 5 months. Then he had a period of disuse for about 1-1/2 years. The retest at 2 two years since the end of practice (Swift, 1905) showed no reliable difference from the last session of practice even for the first juggling trial. Four years later (Swift, 1910) the subject was retested again, and his initial retest performance was no better than his performance on the first training session 6 years earlier. However, relearning was rapid. In 11 training sessions the original level of performance was obtained, which corresponds to a 77% savings. A second subject was retested at 2 years after practice, and his retention performance was, if anything, better than the last session of practice.

In another case study Braden (1924) studied the acquisition of ball-throwing skill. The subject had to hit a box with a 5-inch circular hole at a distance of 12 feet. Each session consisted of eight trials of throwing 25 balls or a total of 200 throws. The experiment consisted of 100 sessions. The accuracy of the subject's performance for the first five sessions was 20.2% and for the last 10 sessions, 42.5%. After almost 2 years of no practice the subject's initial performance during the first five retest sessions (26.6% accuracy) was comparable to that of the first five training sessions. However, after only 9 relearning sessions the retest performance was consistently higher than the average of the last 10 training sessions. This result indicates a savings of 90% after almost 2 years of disuse.

Johnson (1927) reports that subjects who had acquired the ability to walk on a 30-feet long tight wire were able to do so on their first trial even after 1 or 2 years. Subjects were given three trials per session. For each trial the distance traversed by subjects before they lost balance was measured. The criterion of walking the entire distance on three consecutive trials was attained by all 76 subjects within 17 sessions, and most subjects reached the criterion much earlier. For the subset of 20 subjects willing to take an IQ-test, a sizable negative correlation (-0.57) between IQ and number of required trials was obtained.

Kolers (1976) studied the retention of an acquired skill to read inverted text after more than a year of disuse. During the original acquisition students read 220 pages of text spaced out during several sessions and showed continuous improvement in their speed for reading inverted text. At the first retest session over a year later the subjects' reading speed was much faster than at the beginning of acquisition and was comparable to that at about the 168th page of original reading. This performance corresponds to a modified savings score of 76.4%. To acquire the reading speed they had achieved at the end of acquisition the subjects needed to read about 182 pages, a savings score of 17.3%. In contrast to the other retention studies the modified savings score in Kolers' (1976) study is much higher than the actual savings score. This means that learning from the first retest session on is slower at retest than during original acquisition. The most likely account of this finding is that the subjects used a different strategy at retest than at original learning.

Brooks (1924) gave 24 subjects 15 minutes of practice on each of three tasks for 14 sessions. The three tasks were mental multiplication, cancellation of specific letters in text, and inverted writing. After a 10-week delay the subjects' performance was retested. Brooks (1924) reports that the losses in percentage of gain in performance during practice were 43.12%, 6.37%, and 23.38% for mental multiplication, letter cancellation, and inverted writing respectively. These losses can be converted from the acquisition curves into modified savings of 50%, 78.6% and 64.3% respectively. The modified savings imply a much more similar retention across tasks. Brooks (1924) also calculated correlations between initial and final performance during acquisition and found correlations of 0.74 for mental multiplication and around 0.50 for both letter cancellation and inverted writing. He also correlated initial performance with amount of improvement during practice and found no correlation for inverted writing, a negative correlation (-0.47) for letter cancellation, and a positive correlation (0.52) for mental multiplication.

Hill (1914) reports on a case study of mirror drawing with the right arm during 48 sessions involving a single trial each. After 3 years the subject was retested and showed a modified savings of 64.6% on the first trial. The subject regained final practice performance with 91.7% savings. Hill's (1914) study provides some interesting results on transfer. Before and after training the subject was tested on the trained pattern and on a different pattern with both the right and left arm. As illustrated in the results in Figure 32 the transfer across the right to the left, unpracticed arm was almost perfect and transfer to the unfamiliar patterns for both arms was substantial and also equal.

Insert Figure 32 about here

These results are based on a single test trial and should be viewed with some caution.

Retention of basic skills during summer vacation.

The studies of retention of basic skills during summer vacation are relatively old and have been previously reviewed by Stroud (1940). Many of the studies described by Stroud (1940) do not report statistics and simply provide mean differences. Several studies used the same test before and after summer vacation and did not use control groups that would have allowed evaluation of retest effects. In this section we will briefly review only those studies relevant to basic skills like arithmetic and reading. Table 7 summarizes all studies of retention of arithmetic-related skills during summer vacation.

Insert Table 8 about here

The studies in Table 8 show a consistent and reliable decrease in calculational skills with the exception of Morrison's (1924) study, where no significant results were found at all. The loss in arithmetical reasoning was not found to be significant in any of the studies. Some of the investigators attempted to estimate how much time in school is required to regain loss in calculational skills. Morgan (1929) estimated 3-1/2 months and Bruene (1928) half a year. Nelson (1928) retested her students in the 5th and 7th grade with parallel forms of the original test and found that it took them 2 to 6 months to regain the level their final performance had attained in the spring.

Insert Table 9 about here

Table 9 contains results from studies of retention of reading-related skills, which were evaluated with parallel forms of tests. Retention tests based on retesting with the same tests are not acceptable because they have been reliably shown to increase the function they were designed to measure. However, even with parallel forms a number of studies show reliable increases. In no case is a reliable decrement observed. Elder (1927) argued that improvement in reading increases at about the same rate during summer vacation as during the regular school year. It is quite possible that students' reading activities during the summer mediate the observed improvements.

Summary on retention of skills

The number of studies providing information on retention of acquired skills after various intervals of delay is quite small. The studies on typing and tracking reviewed in this paper are consistent with the hypothesis of exponential decay even for acquired skills. In some of the cases, especially those involving extensive practice, decay appears to occur at slower rates than for many of the studies of memory performance discussed earlier. Our review also showed that a couple of factors can account for results in some studies that reported no reliable decay. In some cases the acquired performance was limited by ceiling effects (or floor effects for reaction time), and the lack of reliable decrement in performance after delay is often associated with a performance that corresponds to a modest amount of savings. Other studies discount performance early in the test session as warm-up and hence over estimate retention performance. Considering the cue-rich environment provided by equipment or test apparatus and the fact that skill is mostly measured by the speed of producing a response, the similarities between skill and memory performance are remarkable.

The conceptualization of skill acquisition in terms of stages appear to be quite fruitful in accounting for not only learning performance but also retention after delay. The rapid improvement during the first blocks, or even first trials, of original learning as well as relearning suggest theoretically distinct mechanisms at work. Further work using more detailed methodology to assess the structure of acquired skill and the cognitive processes involved in its reacquisition appear especially promising. A particular finding, which appears especially interesting to explore more systematically is Hill's (1934) result that the increased performance in typing a specific 100-word paragraph showed less retention than the typing skill for unfamiliar texts.

In some of the previous sections we have discussed the effects of overlearning and interference. The concept of overlearning is rarely used in the context of skill acquisition. Most of the measures of performance used in studies of skill acquisition show a steady improvement even with extensive amounts of practice. If skill acquisition is viewed as consisting of three stages characterized by different types of learning, a similar distinction can be made between learning and overlearning of skill. The closest correspondence with the concept of overlearning is the autonomous final stage, in which the acquired behavior patterns are no longer changing and the "learning" proceeds as automation of the behavior pattern.

Studies of motor skills provide remarkably consistent evidence for the systematic changes in the types of learning occurring throughout skill acquisition. Extensive reviews of the relation between individual differences in initial performance and individual differences in final performance consistently show no relationship (Kleinman, 1977; Marteniuk, 1974). The same reviews also note the lack of any general factors of motor ability and conclude that final performance can rarely be predicted from performance on psychometric tests of motor ability. Kleinman (1977) proposed the possibility that different parts of the neural organization of the motor system are involved in the different stages of skill acquisition.

These result showing a lack of relation between motor skills the performance of skills can be expressed as a lack of transfer and interference. Books on motor learning (e.g., Kerr, 1982) note the consistent lack of transfer and interference of most motor skills. One of the few clear demonstrations of interference between motor skills comes from a study of tracking skill (Lewis, McAllister & Adams, 1951), where interference was observed when the controls actions were reversed . Positive transfer appears to be observed mostly when the same general cognitive principles can be applied or when the same subtasks are shared (cf. the common elements theory).

The body of relevant research on transfer is too vast for a reasonable treatment here. There are, however, a couple of findings that strongly support the framework of stages, and they will be briefly discussed. McGeoch and Irion (1952) pointed to several studies showing that interference from practice on a similar task varies as a function of the amount of previous practice on that task. The interference found by Siipola and Israel (1933) on a subsequent, similar task is shown in Figure 33 as a function of four different amounts of practice.

Insert Figure 33 about here

With intermediate practice negative transfer is observed; but as amount of practice is increased, large amounts of positive transfer are found. Transfer between tasks yielding initial positive transfer appear to increase monotonically as a function of increased practice (McGeoch & Irion, 1952).

Another very interesting finding described by McGeoch and Irion (1952) comes from studies of retention of transfer. Bunch's (1936) study is one of the few using long retention intervals. The retention of a specific problem and transfer to a related problem is plotted as a function of delay interval in Figure 34.

Insert Figure 34 about here

It is remarkable that the transfer retention shows essentially no decline with the length of the retention interval. McGeogh and Irion (1952) propose that the transfer performance corresponds to the retention of general factors, such as mode of attack. In addition to providing some interesting results on different rates of decay for various types of information retained, the transfer paradigm suggests a method for decomposing performance, and especially retention, into different types of mediating components and representations.

General Summary

Several different research traditions have been reviewed to extract relevant information on the long-term retention of acquired information, knowledge, and skill. These research traditions have focused on different types of behavior and used quite different methodologies. However, the results on retention are remarkably similar. When obviously flawed studies are discarded, such as those that retest the same subjects multiple times, retention appears to decay as a logarithmic function of time, at least as a reasonable first approximation in all but a very small number of studies. The rate of decay has been found to be remarkably similar across different types of materials. This finding is particularly salient for the studies discussed above which compare the retention of different types of information, where the constant differences in retention levels across delay intervals imply the same slope of forgetting. The differences in retention between different materials correspond to differences in the intercepts, or equivalently in the initial level of retention or "strength" of association. Similar results have been obtained for comparisons across different methods of measuring retention.

It appears that rates of decay are different for different levels of skill acquired with very different amounts of practice. It is also only in the domain of skill acquisition that consistent and striking effects of "overlearning" occur for retention. Finally, as a function of amount of practice on a task, interference does not simply decrease subsequent learning but changes into facilitation.

The prospects of infering general characteristics or even "laws" describing or predicting the retention of performance as a function of acquired performance and length of retention interval, appear to be much better than was expected. We are currently involved in developing a general theoretical model of retention as well as a corresponding parametric fit to the empirically observed retention performance as a function of the length of the retention interval for the studies reviewed above. The theoretical implications of the existence of such a model would be great. The practical value of successfully predicting long-term retention of skills acquired in applied settings might be even greater. In addition, a general conceptual framework is needed for integrating the results from these different research traditions in a single unified framework.

Toward a general conceptual framework for studying long-term retention

The single greatest difference between the research traditions is in the types of observations made to measure retention. The selection of recorded observations is determined by the respective theories, which often consider only some aspect of behavior. Recently the human information processing theory of cognition has evolved as the major theory. Its advantage over the earlier theories is that a broad range of different types of observations are used to provide a convergent description of the cognitive processes under study.

Methodology for Describing Cognitive Processes

The methodology for obtaining detailed descriptions of cognitive processes is quite different from standard experimental methodology. To describe cognitive processes successfully, it is important to collect more observations than the traditional performance data of reaction time and accuracy of response. Figure 35 outlines a number of different types of observations that can be collected on any cognitive process.

Insert Figure 35 about here

At the top of Figure 34 cognition is schematically represented as a series of internal processing steps consistent with the human information processing theory of cognition. These internal processing steps cannot, of course, be directly observed through any of the sources of data shown in Figure 1; but the framework of information processing theory allows specification of hypotheses about the relation between the internal processing steps and observable behavior. For example, when a subject fixes his or her gaze on a specific item in a visually presented table of information, the corresponding internal step is hypothesized to involve processing of that information. Ericsson and Oliver (in press) have attempted to systematically survey these assumptions for different types of observations.

Traditionally, scientists using information processing models have been satisfied to identify a theoretical model that is sufficiently powerful to compute or produce the correct answer from the information given to human subjects performing a task. However, analysis of the task often identifies many possible models and associated sequences of processing steps that can account for the observed performance. If the processing model is required to regenerate not simply the answer or response but also other observations like eye-fixations, the number of adequate models is dramatically reduced.

Traditionally experimental psychologists have almost exclusively limited their observations to the accuracy of the produced answer or response. With one possible exception, all the work on memory for lists, texts, procedures, and knowledge that has been surveyed in this review analyzes only the ability of subjects to produce correct responses. The possible exception is relearning studies that use fixed presentation rates and for which the total relearning time is given or can be calculated. In the work on skill acquisition information on reaction time or the speed of responses is available. Studies of reaction time and attempts to account for differential latencies on different tasks in terms of the duration of component processes have been and remain a major direction in cognitive psychology. Only a relatively small number of investigators have compared reaction times and latencies at study and after delay on traditional memory tasks (Michotte & Portych in Woodworth, 1938; Mueller & Pilzecker, 1900). These investigators have all used memory tasks involving paired associates. The reported latencies for correct responses are given in Figure 36.

Insert Figure 36 about here

The retrieval speed is markedly increased by retention delay. It is important to note that the reaction time estimates are confounded by the fact that at delay the number of correct responses is decreased. At immediate test, that is, after a 5-minute retention interval, the proportion of correct responses is 57% and 79% for 5 and 14 repetitions respectively (Mueller & Pilzecker, 1900). At the delayed test 24 hours later the percentage of correct responses is reduced to 33% and 47% respectively. Hence the observed increases in reaction time are probably underestimated because, as no correct answer is generated, some of the weakest traces are not even included. One of the reasons that retrieval times have not been used much in memory research is that recall latencies are very difficult or even impossible to assess by serial and free recall methods. In addition to recording the recall latency, Michotte and Portych (Woodworth, 1938) also collected retrospective reports on the retrieval processes.

Although all the types of observations are equally valid, some of them contain more information related to the cognitive processes for certain tasks. For example, subjects' verbal reports of their thought processes are essential for studying their use of preexisting knowledge during memorization and recall.

Ericsson and Simon (1980, 1984) started out by examining the case in which the relation between thoughts and requested verbalizations is predicted to be optimal. The sequence of successive states shown in Figure 1 correspond to thoughts in attention; if subjects verbalize their new thoughts as they enter attention, their thoughts will become observable as verbalized segments.

The general relation between heeded thoughts and observable verbalizations is much easier to understand in the context of specific examples. Table 10 displays the think-aloud protocol of a subject mentally multiplying 36 times 24.

Insert Table 10 about here

Most of the verbalized information consists of generated intermediate steps, like "4," "carry the 2," and "144." There is no difference in principle between these intermediate steps and the final result, "864."

From this model of concurrent verbalization it is clear that the subject has to have time to complete the verbalization of the heeded information before new thoughts enter attention. For tasks in which subjects have extensive experience, the sequence of thoughts is so closely connected that a concurrent, sequential verbalization of these spontaneously occurring thoughts is not possible. In such situations, subjects can report their thoughts in retrospect by recalling the sequence of thoughts just after they complete the task. When the time required to complete the task is relatively short (a couple of seconds or less), retrospective report of all heeded thoughts will be relatively complete. For tasks with longer duration, concurrent reports, or think-alouds, will be more detailed than the corresponding retrospective reports.

In our review we found a small number of studies describing results of subjects' verbal reports. Some of the studies used quite different methods for eliciting verbal reports than those currently recommended (Ericsson & Simon, 1984). Also, the verbally reported information was rarely validated by convergence with other performance data or by specially designed experiments. Nevertheless, verbal reports may be critical to understanding the subjects' internal encoding of the presented information as well as the cognitive processes used in retrieval. Of particular importance, the verbal reports will be critical in determining what existing knowledge is used in encoding and how newly acquired information is stored.

In our discussion of nonsense syllables we noted the uncontested importance of preexisting knowledge for rapid acquisition of new information. If new knowledge is encoded in terms of old knowledge, what is the implication of forgetting old knowledge that has repeatedly been used to store new knowledge? Although the relevant empirical evidence has not been collected, it appears from the skill acquisition studies discussed earlier that such encodings using well entrenched, preexisting knowledge would improve rather than hinder retention of that knowledge. From detailed analyses of verbal protocols we hope that an understanding of the structure of acquired new knowledge will emerge and help us further understand retention in general.

References

Ammons, R. B., Farr, R. G., Block, E., Neumann, E., Dey, M., Marion, R., & Ammons, C. H. (1958). Long-term retention of perceptual-motor skills. Journal of Experimental Psychology, 55, 318-328.

Anderson, J. P., & Jordan, A. M. (1928). Learning and retention of Latin words and phrases. Journal of Educational Psychology, 19, 485-496.

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-406.

Ausubel, D. P., Robbins, L. C., & Blake, E., Jr. (1957). Retroactive inhibition and facilitation in the learning of school materials. Journal of Educational Psychology, 48, 334-343.

Ausubel, D. P., Stager, M., & Gaite, A.J.H. (1968). Retroactive facilitation in meaningful verbal learning. Journal of Educational Psychology, 59, 250-255.

Baddeley, A. D., & Longman, D.J.A. (1978). The influence of length and frequency of training session on the rate of learning to type. Ergonomics, 21, 627-635.

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. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. Journal of Experimental Psychology: General, 108, 296-308.

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. (1983). The cognitive map of a city: fifty years of learning and memory. The Psychology of Learning and Motivation, 17, 125-163.

Bahrick, H. P., & Bahrick, P. (1964). A re-examination of the interrelations among measures of retention. Quarterly Journal of Experimental Psychology, 16, 318-324.

Baltes, P. B., & Kliegl, R. (1985). On the dynamics between growth and decline in the aging of intelligence and memory. In K. Poeck, H. -J. Freund, & H. Gaenshirt (Eds.), Neurology. Heidelberg: Springer-Verlag.

Bassett, S. J. (1928). Retention of history in the sixth, seventh, and eighth grades with special reference to the factors that influence retention. (No. 12). Baltimore, MD: Johns Hopkins University Studies in Education.

Battig, W. F., Nagel, E. H., Voss, J. F., & Brogden, W. J. (1957). Transfer and retention of bidimensional compensatory tracking after extended practice. American Journal of Psychology, 70, 75-80.

Bell, H. M. (1950). Retention of pursuit rotor skill after one year. Journal of Experimental Psychology, 40, 648-649.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives. New York: Longmans, Green.

Book, W. F. (1925). The psychology of skill with special reference to its acquisition in typewriting. New York: Gregg.

Bower, G. H., & Hilgard, E. R. (1981). Theories of learning. Englewood Cliffs, NJ: Prentice-Hall.

Braden, S. R. (1924). An extensive experiment in motor learning and re-learning. Journal of Educational Psychology, 15, 313-315.

Bransford, J. D. (1979). Human cognition: Learning, understanding, and remembering. Belmont, CA: Wadsworth.

Briggs, L. J., & Reed, H. B. (1943). The curve of retention for substance material. Journal of Experimental Psychology, 32, 513-517.

Brooks, F. D. (1924). Learning in the case of three dissimilar mental functions. Journal of Experimental Psychology, 7, 462-468.

Bruene, E. (1928). Effect of the summer vacation on the achievement of pupils in the fourth, fifth and sixth grades. Journal of Educational Research, 18, 309-314.

Bunch, M. E. (1936). The amount of transfer in rational learning as a function of time. Journal of Comparative Psychology, 22, 325-337.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. The Psychology of Learning and Motivation, 16, 1-58.

Chi, M.T.H., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. Developmental Psychology, 19, 29-39.

Christiaansen, R. E. (1980). Prose memory: Forgetting rates for memory codes. Journal of Experimental Psychology: Human Learning and Memory, 6, 611-619.

Craig, C. S., Sternthal, B., & Olshan, K. (1972). The effect of overlearning on retention. The Journal of General Psychology, 87, 85-94.

Davis, R. A. (1935). Psychology of learning: A textbook in educational psychology. New York: McGraw-Hill.

Dietz, A. G., & Jones, G. E. (1931). Factual memory of school pupils for a short article which they read a single time. Journal of Educational Psychology, 22, 586-598, 667-676.

Dooling, D. J., & Christiaansen, R. E. (1977). Levels of encoding and retention of prose. The Psychology of Learning and Motivation, 11, 1-39.

Dooling, D. J., & Lachman, R. (1971). Effects of comprehension on retention of prose. Journal of Experimental Psychology, 88, 216-222.

Ebbinghaus, H. (1964). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Eikenberry, D. H. (1923). Permanence of high school learning. Journal of Educational Psychology, 14, 463-482.

Elder, H. E. (1927). The effect of the summer vacation on silent-reading ability in the intermediate grades. The Elementary School Journal, 27, 541-546.

Ericsson, K. A. (1985). Memory skill. Canadian Journal of Psychology, 39(2), 188-231.

Ericsson, K. A., & Oliver, W. (in press). Methodology for laboratory research on thinking: Task selection, collection of observations and data analysis. In R. J. Sternberg and E. E. Smith (Eds.), The psychology of human thought. New York: Cambridge University Press.

Ericsson, K. A., & Polson, P. G. (in press). Memory for restaurant orders. In M. Chi, R. Glaser, & M. Farr (Eds.), The nature of expertise. Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87, 215-251.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis. Cambridge, MA: MIT Press/Bradford.

Eylon, B., & Reif, F. (1984). Effects of knowledge organization on task performance. Cognition and Instruction, 1, 5-44.

Finkenbinder, E. O. (1913). The curve of forgetting. American Journal of Psychology, 24, 8-32.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), Categories of human learning. New York: Academic Press.

Fitts, P. M., & Posner, M. I. (1967). Human performance. Monterey, CA: Brooks/Cole.

Fleishman, E. A., & Parker, J. F., Jr. (1962). Factors in the retention and relearning of perceptual-motor skill. Journal of Experimental Psychology, 64, 215-226.

Freeman, F. N., & Abernethy, E. M. (1930). Comparative retention of typewriting and of substitution with analogous material. Journal of Educational Psychology, 21, 639-647.

Freeman, F. N., & Abernethy, E. M. (1932). New evidence of the superior retention of typewriting to that of substitution. Journal of Educational Psychology, 23, 331-334.

Freud, H., & Cheronis, N. D. (1940). Retention in the physical science survey course. Journal of Chemical Education, 17, 289-293.

Frutchey, F. P. (1937). Retention in high-school chemistry. Journal of Higher Education, 8, 217-218.

Gilbert, T. F. (1957). Overlearning and the retention of meaningful prose. The Journal of General Psychology, 56, 281-289.

Glasnapp, D. R., Poggio, J. P., & Ory, J. C. (1978). End-of-course and long-term retention outcomes for mastery and nonmastery learning paradigms. Psychology in the Schools, 15, 595-603.

Goldberg, S. L., Drillings, M., & Dressel, J. D. (1981). Mastery training: Effect on skill retention (Technical Report No. 513). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Graesser, A. C. (1978). How to catch a fish: The memory and representation of common procedures. Discourse Processes, 1, 72-89.

Graesser, A. C., Robertson, S. P., Lovelace, E. R., & Swinehart, D. M. (1980). Answers to why-questions expose the organization of story plot and predict recall of actions. Journal of Verbal Learning and Verbal Behavior, 19, 110-119.

Greene, E. B. (1931). The retention of information learned in college courses. Journal of Educational Research, 24, 262-273.

Grudin, J. T. (1981). The organization of serial order in typing. Unpublished doctoral dissertation, Department of Psychology, University of California, San Diego.

Grudin, J. T. (1983). Error patterns in novice and skilled transcription typing. In W. E. Cooper (Ed.), Cognitive aspects of skilled typewriting. New York: Springer-Verlag.

Gillette, A. L. (1936). Learning and retention: A comparison of three experimental procedures. Archives of Psychology, 28(Serial No. 198).

Hagman, J. D. (1980). Effects of presentation- and test-trial training on motor acquisition and retention (Technical Report No. 431). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hagman, J. D., & Rose, A. M. (1983). Retention of military tasks: A review. Human Factors, 25, 199-213.

Hagman, J. (1979). Typewriting: Retention and relearning (Research Report No. 1211). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ERIC Document Reproduction Service No. ED 173 659)

Hall, J. F. (1955). Retroactive inhibition in meaningful material. Journal of Educational Psychology, 46, 47-52.

Hill, D. S. (1914). Minor studies in learning and relearning. Journal of Educational Psychology, 5, 375-386.

Hill, L. B. (1934). A quarter century of delayed recall. Journal of Genetic Psychology, 44, 231-238.

Hill, L. B. (1957). A second quarter century of delayed recall or relearning at 80. Journal of Educational Psychology, 48, 65-68.

Hill, L. B., Rejall, A. E., & Thorndike, E. L. (1913). Practice in the case of typewriting. Pedagogical Seminary, 20, 516-529.

Hovland, C. I. (1951). Human learning and retention. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York: Wiley.

Hurlock, R. E., & Montague, W. E. (1982). Skill retention and its implications for navy tasks: An analytical review (NPRDC Special Report No. 82-21). San Diego, CA: Navy Personnel Research and Development Center.

Johnson, G. B. (1927). A study in learning to walk the tight wire. Pedagogical Seminary, 34, 118-128.

Johnson, P. O. (1930). The permanence of learning in elementary botany. Journal of Educational Psychology, 21, 37-47.

Kennedy, L. R. (1932). The retention of certain Latin syntactical principles by first and second year Latin students after various time intervals. Journal of Educational Psychology, 23, 132-146.

Kerr, R. (1982). Psychomotor learning. Philadelphia, PA: Saunders College.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Kintsch, W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of the propositions in the base structure of sentences. Cognitive Psychology, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. Journal of Verbal Learning and Verbal Behavior, 14, 196-214.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kleinman, M. (1977). Ability factors in motor learning. Perceptual and Motor Skills, 44, 827-836.

Kolers, P. A. (1976). Reading a year later. Journal of Experimental Psychology: Human Learning and Memory, 2, 554-565.

Krueger, W.C.F. (1929). The effect of overlearning on retention. Journal of Experimental Psychology, 12, 71-78.

Landauer, T. K., & Ainslie, K. I. (1975). Exams and use as preservatives of course-acquired knowledge. Journal of Educational Research, 69, 99-104.

Layton, E. T. (1932). The persistence of learning in elementary algebra. Journal of Educational Psychology, 23, 46-55.

Lewis, D., McAllister, D. E., & Adams, J. A. (1951). Facilitation and interference in performance on the modified Mashburn apparatus: I. The effects of varying the amount of original learning. Journal of Experimental Psychology, 41, 247-260.

Luh, C. W. (1922). The conditions of retention. Psychological Monographs, 31(Whole No. 142).

Marteniuk, R. G. (1974). Individual differences in motor performance and learning. Exercise and Sport Sciences Reviews (Vol. 2). New York: Academic Press.

McGeoch, J. A. (1942). The psychology of human learning: An introduction. New York: Longmans, Green.

McGeoch, J. A., & Irion, A. L. (1952). The psychology of human learning. New York: Longmans, Green.

McGeoch, J. A., & McKinney, F. (1934a). Retroactive inhibition in the learning of poetry. American Journal of Psychology, 46, 19-33.

McGeoch, J. A., & McKinney, F. (1934b). The susceptibility of prose to retroactive inhibition. American Journal of Psychology, 46, 429-436.

McGeoch, J. A., & Melton, A. W. (1929). The comparative retention value of maze habits and of nonsense syllables. Journal of Experimental Psychology, 12, 392-414.

Morgan, L. D. (1929). How effective is special training in preventing loss due to the summer vacation? Journal of Educational Psychology, 20, 466-471.

Morrison, J. C. (1924). What effect has the summer vacation on children's learning and ability to learn? Educational Research Bulletin, 3, 245-249.

Mueller, G. E., & Pilzecker, A. (1900). Experimentelle Beitraege zur Lehre vom Gedaechtniss. Zeitschrift fuer Psychologie und Physiologie der Sinnesorgane: Ergaenzungsband 1.

Mueller, G. E., & Schumann, F. (1894). Experimentelle Beitraege zur Untersuchung des Gedaechtnisses. Zeitschrift fuer Psychologie, 6, 81-339.

Murdock, B. B., Jr. (1974). Human memory: Theory and data. Potomac, MD: Erlbaum.

Myers, G. C. (1917). Delayed recall in American history. Journal of Educational Psychology, 8, 275-283.

Myrow, D. L., & Anderson, R. C. (1972). Retroactive inhibition of prose as a function of the type of test. Journal of Educational Psychology, 63, 303-308.

Naylor, J. C., & Briggs, G. E. (1961). Long-term retention of learned skills: A review of the literature. ASD Technical Report No. 61-390.

Nelson, M. J. (1928). How much time is required in the fall for pupils of the elementary school to reach again the spring level of achievement? Journal of Educational Research, 18, 305-308.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Erlbaum.

Osborn, W. C., Campbell, C. H., & Harris, J. H. (1979). The retention of tank crewman skills (Research Report No. 1234). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Postman, L. (1971). Transfer, interference and forgetting. In J. W. Kling & L. A. Riggs (Eds.), Woodworth and Schlosbergs's experimental psychology. New York: Holt, Rinehart, and Winston.

Postman, L. (1976). Methodology of human learning. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 3). Hillsdale, NJ: Erlbaum.

Radosavljevich, P. R. (1907). Das Behalten und Vergessen bei Kindern und Erwachsenen nach experimentellen Untersuchungen. Paedagogische Monographien, No. 1.

Reynolds, B., & Bilodeau, I. M. (1952). Acquisition and retention of three psychomotor tests as a function of distribution of practice during acquisition. Journal of Experimental Psychology, 44, 19-26.

Rubin-Rabson, G. (1941). Studies in the psychology of memorizing piano music. VII. A comparison of three degrees of overlearning. Journal of Educational Psychology, 32, 688-696.

Salthouse, T. A., & Somberg, B. L. (1982). Skilled performance: Effects of adult age and experience on elementary processes. Journal of Experimental Psychology: General, 111, 176-207.

Schaie, K. W. (Ed.). (1983). Longitudinal studies of adult psychological development. New York: Guilford.

Schendel, J. D., & Hagman, J. D. (1980). On sustaining procedural skills over prolonged retention intervals (Research Report No. 1298). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Schendel, J. D., Shields, J. L., & Katz, M. S. (1978). Retention of motor skills: Review (Technical Paper No. 313). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Schroeder, L. J. (1934). The effect of summer vacation on ability in typewriting. Journal of Applied Psychology, 18, 282-287.

Shields, J. L., Goldberg, S. L., & Dressel, J. D. (1979). Retention of basic soldiering skills (Research Report No. 1225). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Siipola, E. M., & Israel, H. E. (1933). Habit-interference as dependent upon stage of training. American Journal of Psychology, 45, 205-227.

Singer, M. (1982). Comparing memory for natural and laboratory reading. Journal of Experimental Psychology, 111, 331-347.

Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of interference. Cognitive Psychology, 10, 438-464.

Stevens, S. S. (Ed.). (1951). Handbook of experimental psychology. New York: Wiley.

Stroud, J. B. (1940). Experiments on learning in school situations. Psychological Bulletin, 37, 777-807.

Swift, E. J. (1903). Studies in the psychology and physiology of learning. American Journal of Psychology, 14, 201-251.

Swift, E. J. (1905). Memory of a complex skillful act. American Journal of Psychology, 16, 131-133.

Swift, E. J. (1906). Memory of skillful movements. Psychological Bulletin, 3, 185-187.

Swift, E. J. (1910). Relearning a skillful act: An experimental study in neuro-muscular memory. Psychological Bulletin, 7, 17-19.

Swift, E. J., & Schuler, W. (1907). The learning process. Psychological Bulletin, 4, 307-310.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology, 9, 77-110.

Thorndyke, P. W., & Hayes-Roth, B. (1979). The use of schemata in the acquisition and transfer of knowledge. Cognitive Psychology, 11, 82-106.

Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. Journal of Memory and Language, 24, 595-611.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. Journal of Memory and Language, 24, 612-630.

Tyler, R. W. (1933). Permanence of learning. Journal of Higher Education, 4, 203-204.

Underwood, B. J. (1949). Experimental psychology. New York: Appleton-Century-Crofts.

Underwood, B. J. (1957). Interference and forgetting. Psychological Review, 64, 49-60.

van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Watson, R. I. (1938). An experimental study of the permanence of course material in introductory psychology. In R. S. Woodworth (Ed.), Archives of psychology. New York: Science Press.

Weaver, F. J., Ramirez, A. G., Dorfman, S. B., Raizner, A. E. (1979). Trainee's retention of cardiopulmonary resuscitation. Journal of the American Medical Association, 241, 901-903.

Welborn, E. L., & English, H. (1937). Logical learning and retention: A general review of experiments with meaningful verbal materials. Psychological Bulletin, 34, 1-20.

Wert, J. E. (1937). Twin examination assumptions. Journal of Higher Education, 8, 136-140.

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory. Cognitive Science, 5, 87-119.

Woodworth, R. S., & Schlosberg, H. (1954). Experimental psychology. New York: Holt, Rinehart, & Winston.

Woodworth, R. S., & Schlosberg, H. (1965). Experimental psychology. New York: Holt, Rinehart, & Winston.

Worcester, D. A. (1928). The permanence of learning in high school subjects--algebra. Journal of Educational Psychology, 19, 343-345.

Footnote

[1]The preparation of this review has been supported by a grant (No. F$1689-84-0002 Task Authority 744-022) to the first author from the Airforce Human Resources Laboratory, Brooks Air Force Base, Texas. We gratefully acknowledge the comments and suggestions of Janet Grassia on an earlier version of this report.

Table 1

<u>Tasks and task steps tested. Basic skills retention</u>.

<u>First Aid</u> - CPR - Cardiopulmonary Resuscitation

PERFORMANCE STEPS		PERCENT GO*
1.	Positions victim on back	100
2.	Tilts head back, one hand on forehead one under neck	93
3.	Positions close to victim's side	93
4.	Places heel of hand on lower half breastbone	90
5.	Spreads and raises fingers of hand	86
6.	Places other hand on top of first	86
7.	Brings shoulders over victim's breastbone keeping arms straight	86
8.	Presses downward 1-1/2 to 2 inches	86
9.	Releases pressure immediately	86
10.	Does not remove hands	73
11.	<u>AFTER 15 COMPRESSIONS PLACE HAND BEHIND NECK</u>	45
12.	<u>PINCHES NOSTRILS TOGETHER WITH OTHER HAND</u>	45
13.	<u>TILTS VICTIM'S HEAD BACK, BLOWS 2 BREATHS</u>	45
14.	<u>RATIO OF HEART PUMPS TO BREATHE</u>	50

* No coaching, no practice since training soldiers (from Shields, Goldberg, & Dressel, 1979).

Table 2

Retention of course growth expressed in percentages.

Objectives	End of Years One	Two	Three
(1)	(2)	(3)	(4)
Naming animal structures	23	18	10
Stating function of structures	59	47	41
Terminology	57	54	51
Other facts	70	68	49
Interpreting new experiments	111	114	119
Applying principles to new situations	127	154	158

Table 3

Percent retained of material learned in Zoology (from Tyler, 1933).

Type of Examination Exercise	% Retained
1. Naming animal structures pictured in diagrams	50
2. Identifying technical terms	81
3. Recalling information	
a. Structures performing function in type forms	87
b. Other facts	86
4. Applying principles to new situations	100
5. Interpreting new experiments	112
Average for all exercises in the exam	85

Table 4

Percent retained of gain made in high school chemistry (from Frutchey, 1937)*.

Type of Test Question		% Retained
1.	Selection of facts	84
2.	Application of principles	92
3.	Terminology	66
4.	Symbols, formulas, and valence	70
5.	Balancing equations	72

* Note: Gain is equal to final test minus pretest and amount retained is equal to retention test minus pretest.

Table 5

Retention of four acquired laboratory skills (Salthouse & Somberg, 1982).

Task	Type of subjects	Statistical test loss after delay	First trial at which performance at delay was observed	Estimated modified savings (%)
Detection of configuration of dots	young	$p < 0.05$	12	12/24
	old	$p < 0.05$	29	29/58
Memory scanning	young	ns	6	6/12
	old	ns	17	17/34
Detection of target element				
successive	young	ns	3	3/6
	old	ns	23	23/46
simultaneous	young	ns	18	36
	old	ns	21	42

Table 6

Summary of long-term retention of acquired tracking skill.

Task	Study	No. of subjects	Length of sessions	No. of sessions	Retention interval	Retention
bi-dimensional tracking	Battig et al. (1957)	3	10x1 minute trials	110	8 months	no reliable loss
rudder control test	Reynolds & Bilodeau (1952) (massed practice group)	84	12x30 sec	1	10 weeks	70% initial performance showed modified savings of 50%
pursuit rotor	Bell (1950)	47	1 minute	20	1 year	60% initial performance showed modified savings of 40%
complex tracking task	Fleishman & Parker (1962) no formal instruction condition	9	21x1 minute	17	9 months	99%
		7	21x1 minute	17	14 months	99%
		8	21x1 minute	17	24 months	94%
	Fleishman & Parker (1962) instruction	10	21x1 minute	17	1 month	no reliable loss
		10	21x1 minute	17	5 months	no reliable loss
		9	21x1 minute	17	9 months	no reliable loss
		9	21x1 minute	17	14 months	no reliable loss
	Ammons et al. (1958) low practice group group	740	15x1 minute	4	1 month	74%
		740	15x1 minute	4	6 months	67%
		740	15x1 minute	4	1 year	58%
		740	15x1 minute	4	2 years	57%
	high practice	740	15x1 minute	32	1 month	96%

Table 7.

Summary chart of long-term retention of typing performance as observed on individual subjects.

Investigator	Initial performance 1st	Initial performance Ave first 5	Final performance Ave test 10	No of hours practice	Duration	Delay	1st	Regain final ??????	Relearning savings	Initial session % of gain
Swift (1906)	360	420	1040	50	50 days	2 yrs 35 days	700	10 hrs	90%	45%
(words/nouns)	6	7	17.33				11.66	5 hrs		
							to regain			
Swift & Schyler	1250	1260	3140	32.5	75 days	84 days	2000	8 days	87.3%	39%
(1907)	8.3	8.4	20.93				13.33			
Strokes/30 min										
Book X	745	1475	87	1 yr	6 months	1365	4x10 min	95.3%	90%	
5 strokes=1 wrd	9.5	29.3	50	1 month	1 addit yr	27.3	9x10 min			
			<u>60</u>			1390				
			187			27.8				
Rejall										
100 wrds fixed	6.23	10.33	41.91	7.45 (fixed)	5 months	4-1/2 yrs	20.5	15 sessions	86.7%	32%
		(s=2.615)	(s=1.748)	31.75 (vary)				5 hours		
300 wrds vary	6.84	9.19	25.54	113 sessions			15.92			42%
		(s=1.88)	(s=1.42)							
Hill										
100 wrds fixed	10	10.62	34.51	127 sessions		25 yrs	21.05	25 sessions	80.3%	44%
		(s=0.78)	(2=1.36)							
300 wrds vary	7.14	8.92	23.81			16.84			535	
		(s=1.23)	(s=0.88)							

Note: Book (1908) uses a complex measure to calculate strokes with a half stroke for hitting the space bar and two strokes for carriage return. Hill (1934) aruged that 15 re-test sessions were required for relearning but the performance on the two tasks is differentially reacquired. Hill (1934) uses an adjusted score compensating for number of errors. Notes with transofrmations to typing speed. Notes giving the calculational formula for savings measures.

Table 8

Retention over summer vacation by different investigators for arithmetic calculation and reasoning.

	N	test	change	p-value two-tailed	grade
Schrepel & Laslett (1936)					
arithmetic fundamental	51	same	decrease	$p<0.001$	8
arithmetic reasoning	74	same	none	ns	8
arithmetic reasoning	72	parallel forms	none	ns	8
arithmetic computation	72	parallel forms	decrease	$p<0.05$	8
arithmetic reasoning	49	parallel forms	none	ns	9
arithmetic computation	49	parallel forms	decrease	$p<0.01$	9
Bruene (1928)					
arithmetic fundamentals	15	parallel forms	decrease	**	4
arithmetic reasoning	15	parallel forms		ns	4
arithmetic fundamentals	26		decrease	***	5
arithmetic reasoning	26			ns	5
arithmetic fundamentals	28		decrease	***	6
arithmetic reasoning	28			ns	6
Morgan (1929)					
arithmetic fundamentals (control group)	38	same test	decrease	***	6
Morrison (1924)					
	32	same test		NS	2
	27	same test		ns	3
	28	same test		ns	4
	24	same test		ns	5
	30	same test		ns	6
	48	same test		ns	8

Table 9

Retention over summer vacation by different investigators for reading related tests.

	N	test	change	p-value two-tailed	grade
Schrepel & Laslett (1936)					
paragraph meaning	72	parallel tests	decrease	ns	8
word meaning	72	parallel tests	increase	p<0.001	8
dictation	72	parallel tests	increase	ns	8
language usage	72	parallel tests	increase	ns	8
paragraph meaning	49	parallel tests	increase	p<0.01	9
word meaning	49	parallel tests	increase	p<0.001	9
dictation	49	parallel tests	increase	ns	9
language usage	49	parallel tests	increase	ns	9
Elder (1927)					
silent reading test	42	parallel tests	23-23 (12)	ns	3
silent reading test	46	parallel tests	34-12(11)	p<0.001	4
silent reading test	55	parallel tests	32-23(18)	p<0.005	5
silent reading test	60	parallel tests	31-29(11)	0<0.01	6
Bruene (1928)					
reading	15	parallel tests	6-9(6)	ns	4
reading	26	parallel tests	10-16(13)	ns	5
language usage	26	parallel tests	13-13(12)	ns	5
reading	28	parallel tests	17-11(9)	ns	6
language usage	28	parallel tests	10-18(16)	ns	6

Table 10

A transcript of a thinking-aloud protocol from a subject mentally multiplying 36 times 24. On the right side, the same multiplication is performed using the traditional paper and pencil method.

OK	36
36 times 24	24
um	144
4 times 6 is 24	720
4	864
carry the 2	
4 times 3 is 12	
14	
144	
0	
2 times 67 is 12	
2	
carry the 1	
2 times 3 is 6	
7	
720	
720	
144 plus 72	
so it would be 4	
6	
864	

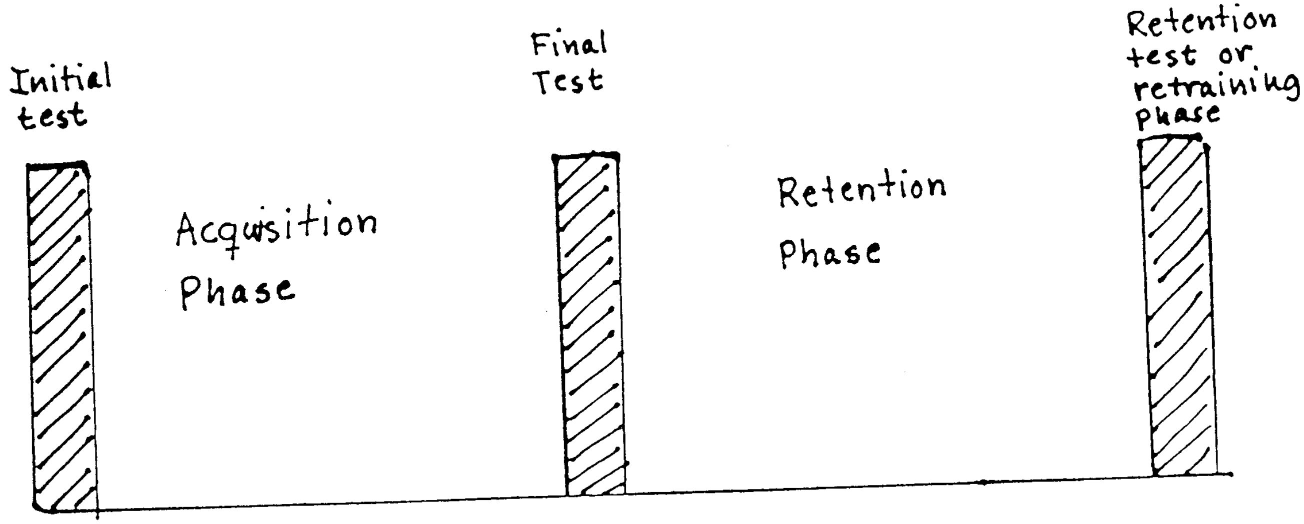

Fig. 1 A schematic outline of the time course of a study of retention.

Fig 2

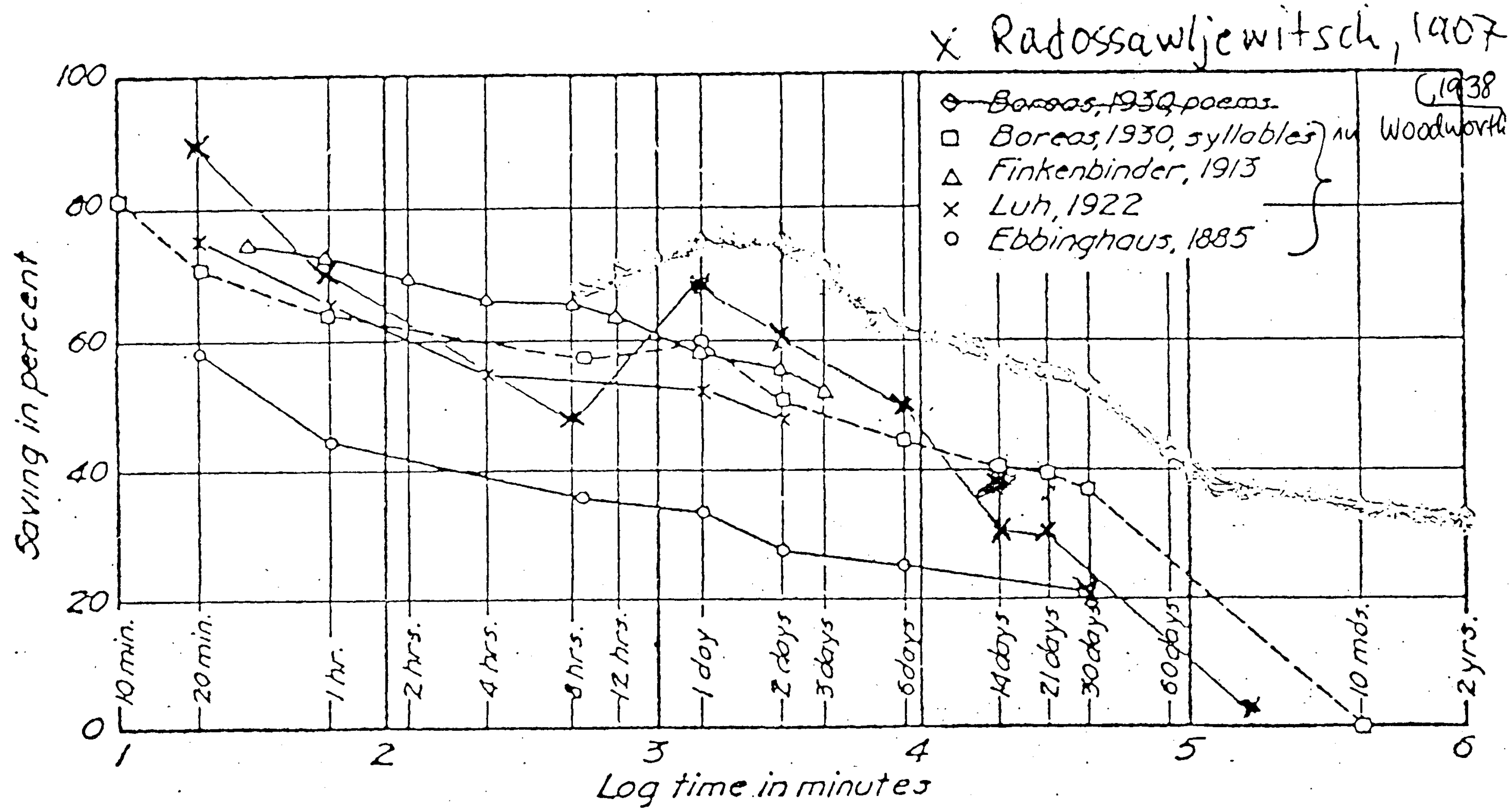

Fig. 2 Several retention curves plotted on a logarithmic abscissa.

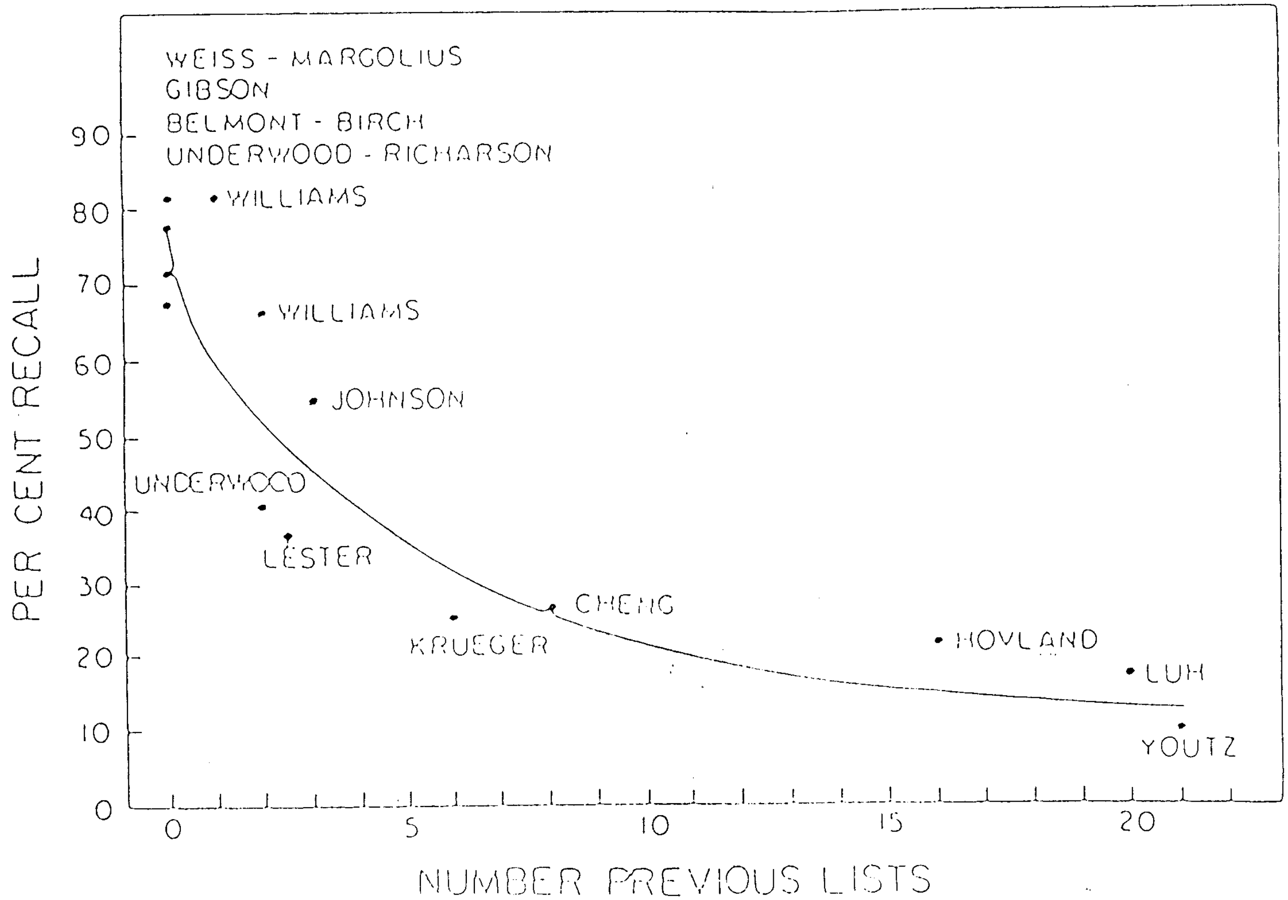

FIG. 3 Recall of a single list after a one-day retention interval as a function of the number of lists previously learned and recalled. (from Underwood, 1957.)