THE PLACE OF COGNITIVE ARCHITECTURES IN A RATIONAL ANALYSIS

Technical Report AIP - 48

John R. Anderson

Departments of Computer Science
and Psychology
Carnegie Mellon University
Pittsburgh, Pa., 15213

14 July 1988

This research was supported by the Computer Sciences Division, Office of Naval Research and DARPA under Contract Number N00014-86-K-0678. Reproduction in whole or in part is permitted for purposes of the United States Government. Approved for public release; distribution unlimited.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION AVAILABILITY OF REPORT Approved for public release; Distribution unlimited
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AIP - 48	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Carnegie-Mellon University 6b OFFICE SYMBOL (*If applicable*)	7a NAME OF MONITORING ORGANIZATION Computer Sciences Division Office of Naval Research
6c ADDRESS (*City, State, and ZIP Code*) Department of Psychology Pittsburgh, Pennsylvania 15213	7b ADDRESS (*City, State, and ZIP Code*) 800 N. Quincy Street Arlington, Virginia 22217-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Same as Monitoring Organization 8b OFFICE SYMBOL (*If applicable*)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0678
8c. ADDRESS (*City, State, and ZIP Code*)	10 SOURCE OF FUNDING NUMBERS p4000ub201/7-4-86 PROGRAM ELEMENT NO: N/A; PROJECT NO: N/A; TASK NO: N/A; WORK UNIT ACCESSION NO: N/A

11. TITLE (*Include Security Classification*)

The Place of Cognitive Architectures in a Rational Analysis

12 PERSONAL AUTHOR(S)
Anderson, John Robert

13a. TYPE OF REPORT	13b. TIME COVERED	14 DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Technical	FROM 86Sept15 TO 91Sept14	1988 July 14	38

16 SUPPLEMENTARY NOTATION
To appear in K. VanLehn (Ed.) Architectures for Intelligence, Hillsdale, NJ: Erlbaum

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
Memory, categorization, rational analysis, cognitive architecture

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

It is argued that human cognition can be predicted from the assumption that it is optimized to the information-processing demands that are placed on it. Results that are taken in support of particular architectures (PDP, ACT*, SOAR) are shown to be consequences of this rationality principle of human cognition. Implications of this rationality principle for cognitive architectures are discussed.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION
22a NAME OF RESPONSIBLE INDIVIDUAL Dr. Alan L. Meyrowitz	22b TELEPHONE (*Include Area Code*) (202) 696-4302 22c. OFFICE SYMBOL N00014

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

The basic goal of a theorist in specifying a cognitve architecture is to specify the mind's principles of operation and organization much like you would specify those of a computer. Any cognitive phenomena should be derivative from these principles As this conference gives witness. there are many cognitive architectures. This paper will try to make some claims about the role of architectures generally in psychological theory. but it will do this by taking as examples three of the architectures which figure prominently at Carnegie Mellon University. There is the Soar architecture of Laird. Newell. and Rosenbloom (in press) my own ACT* architecture (Anderson. 1983), and the PDP architecture of McClelland and Rumelhart (Rumelhart & McClelland. 1986; McClelland & Rumelhart. 1986).

Now that there are numerous candidates for cognitive architectures. one is naturally led to ask which might be the correct one or the most correct one. This is a particularly difficult question to answer because these architectures are often quite removed from the empirical phenomena which they are supposed to account for In actual practice one sees proponents of a particular architecture arguing for that architecture by reference to what I call signature phenomena. These are empirical phenomena which are particularly clear manifestations of the purported underlying mechanisms. The claim is made that the architecture provides particularly natural accounts for these phenomena and that these phenomena are hard to account for in other architectures.

In this paper I will argue that the purported signature phenomena tell us very little about what is inside the human head. Rather they tell us a lot about the world in which the human lives. The majority of this paper will be devoted to making this point with respect to examples from the SOAR, ACT*. and PDP architectures. At the end of the paper I will turn to the issue of the consequences of this point for the role of cognitive architectures.

As a theorist who has been associated with the development of cognitive architectures for 15 years I should say a little about how I came to be advocating this position. I have been strongly influenced by David Marr's (1982) metatheoretical arguments in his book on vision which are nicely summarized in the following quote:

> An algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examing the mechanism (and the hardware) in which it is solved.

Marr made this point with respect to phenomena such as stereopsis where he argued that one will come to an understanding of the phenomena by focusing on the problem of how two two-dimensional views of the world contained enough information to enable one to extract a three-dimensional interpretation of the world and not by focusing on the mechanisms of stereopsis. He thought his viewpoint was appropriate to higher-level cognition although he did not develop it for that application. As recent as a few years ago I could not see how his viewpoint applied to higher-level cognition (Anderson, 1987). However, in the last couple of years I have come to see how it would apply and have realized its advantages. Before specifying its application let us briefly note three advantages of focusing on the information-processing problem and not the information-processing mechanisms:

(1) As Marr emphasized, the understanding the nature of the problem offers strong guidance in the proposal of possible mechanisms. If anything this is more important in the case of higher-level cognition where we face a bewildering array of potential mechanisms and an astronomical space of their possible combinations which we must search in trying to identify the correct architecture.

(2) Again as Marr emphasized, this allows us a deeper level of understanding of these mechanisms. We can understand why they compute in the way they do rather than regarding them as random configurations of computational pieces.

(3) Cognitive psychology faces fundamental indeterminacies such as that between parallel and serial processing, the status of a separate short term memory, or the format of internal representation. Focusing on the information-processing problem allows us a level of abstraction that is above the level where we need to resolve these indeterminacies.

A Rational Analysis

The basic point of Marr's was that if there is an optimal way to use the information at hand the system will use it. I have stated this as the following principle:

Principle of Rationality. The cognitive system optimizes the adaptation of the behavior of the organism.

One can regard this principle as being handed to us from outside of psychology--as a consequence of basic evolutionary principles. However, I do not want to endorse the principle on such an evolutionary basis because there are many cases where evolution does not optimize. Of course, there are many cases where it does (for a recent discussion see Dupre, 1987). On one hand there are the moths of Manchester and on the other hand, as Simon notes in his companion article, there are the fauna of Australia. It is an interesting question just where and how we would expect evolution to produce optimization, but this is an issue that I neither have space nor competence to get into. Rather, I view the principle above as an empirical hypothesis to be judged by how well theories that embody the principle of rationality do in predicting cognitive phenomena.

On the empirical front it might seem that the principle of rationality is headed for sure disaster in accounting for human cognition. It is the current wisdom in psychology that man is anything but rational. However, I think many of the purported irrationalities of man disappear when we take a broader view of the human situation. Among the relevent considerations are the following three:

1. We have to bear in mind the cost of computing the behavior. Many think that the problem with the traditional rational man model of economics is that it ignores the cost of computation. Thus we need something like Simon's (1972) bounded rationality where one includes computation cost in the function to be optimized. For instance, in principle a rational person should be able to play a perfect game of chess told the rules of chess but this ignores the prohibitive cost of a complete search of the game tree.

2. The adaptation of the behavior may be defined with respect to an environment different than the one we are functioning in. For instance, one might wonder why human learning mechanisms do so poorly at picking up knowledge in a school environment. I think the answer is that it is not a school environment that they are adapted to.

3. One must recognize that traditional tests of human rationality typically involve normative models that make no reference to the adaptiveness of the behavior. For instance, normative models typically advocating maximizing wealth while the evidence is that there is a negative correlation between wealth and number of surviving offspring (Vining, 1986). The implication is that one must look critically at the functions which we are trying to optimize in a rational analysis.

With these caveats it is my claim that one can use a rational approach as a framework for deriving behavioral predictions. Developing a theory in a rational framework involves the following 6 steps:

1. Precisely specify what the goals of the cognitive system are.

2. Develop a formal model of the environment that the system is adapted to (almost certainly less structured than the experimental situation).

3. Make the minimal assumptions about computational costs.

4. Derive the optimal behavioral function given (1)-(3).

5. Examine the empirical literatures to see if the predictions of the behavioral function are confirmed.

6. If predictions are off, iterate.

The theory in a rational approach resides in the assumptions in (1) - (3) from which the predictions flow. I refer to these assumptions as the framing of the information processing problem. Note this is a mechanism-free casting of a psychological theory. It can be largely cast in terms of what is outside of the human head rather than inside. As such it enjoys another advantage which is that its assumptions are potentially capable of independent verification.

What I would like to do in the majority of the paper is to apply this rational analysis to one signature phenomenon for each of the three architectures mentioned in the introduction--SOAR, ACT*, and PDP.

SOAR--Power Law Learning

The signature phenomenon I would like to consider for the SOAR theory is power-law learning which is referenced in many of the SOAR publications. Figure 1 illustrates data from the Siebel (1963) task which Rosenbloom, Laird, & Newell (in press) have simulated within SOAR. In this task subjects were presented with a panel of ten lights, some of which were lit. They had to press the corresponding fingers on their hands. Subjects saw all configurations of lights except the one in which no lights were lit. Figure 1 plots their performance time against the amount of practice which they had. Both scales are logarithmic. As can be seen the relationship is linear implying that the performance measure is a power function of practice. As Newell and Rosenbloom (1981) discuss, such

power functions are ubiquitous. The effects can be quite extensive. The data plotted by Seibel covers 40.000 trials.

Insert Figure 1 about here

In the Soar model the power law falls out of the chunking learning mechanism plus some critical auxiliary assumptions. Chunking refers to the collapsing of multiple production firings into a single production firing that does the work of the set. In the Seibel task subjects might chunk productions that will press subsets of lights simultaneously rather than separately. It is assumed that each chunk produces a performance enhancement proportional to the number of productions eliminated. Chunks are formed at a constant rate--either on every opportunity or with equal probability on every opportunity. The final critical assumption is that as chunks span larger and larger units the number of potential chunks grows exponentially. This is fairly transparent in the Seibel task where there are 2^n productions needed to encode all chunks of n lights. As a consequence of the last assumption, learning will progress even more slowly because it takes more experience to encounter all of the larger chunks.

I have always had a number of haunting doubts about the SOAR explanation of the Seibel task. Some of these were expressed in Anderson (1982 and 1983). One is that the exponential growth in chunks does not seem true of simple memory experiments (such as paired-associate learning) which produce beautiful power law learning functions. The second is that the analysis has no place for forgetting effects which must be taking place. So we know by the time of the 40.000 trial of the Seibel task the benefit of the first trial should be fading. Third, the model has no provision for massing effects. We know that as many trials are massed together they loose their effectiveness. Note that the massing effect and forgetting effects are at odds with each other. One is optimized by massing the trials

together and the other by spacing them apart.

I will offer a rational analysis of power law learning which will also explain the forgetting and massing functions. This will be part of a larger rational analysis of human memory which is the topic of the next section

A Rational Analysis of Human Memory

The claim that human memory is rationally designed might strike one at least as implausible as the general claim for the rationality of human cognition. Human memory is always disparaged in comparison to computer memory--it is thought of as slow both in storage and retrieval and terribly unreliable. However, such analyses of human memory fail both to understand the task faced by human memory and the goals of memory. I think human memory should be compared with information-retrieval systems such as the ones that exist in computer science. According to Salton and McGill (1983) a generic information retrieval system consists of four things:

(1) There is a data base of files such as book entries in a library system. In the human case these files are the various memories of things past.

(2) The files are indexed by terms. In a library system the indexing terms might be keywords in the book's abstract. In the human case the terms are presumably the concepts and elements united in the memory. Thus, if the memory is seeing Willie Stargell hit a home run the indexing terms might be Willie Stargell, home run, Three Rivers Stadium, etc.

(3) An information retrieval system is posed queries consisting of terms. In a library system these are suggested keywords by the user. In the case of the human situation it is whatever cues are presented by the environment such as when someone says to me "Think of a home run at Three Rivers Stadium".

(4) Finally there are a set of target files desired by which we can judge the success of the information retrieval.

. One thing that is very clear in the literature on information retrieval systems is that they cannot know the right files to retrieve given a query. This is because the information in a query does not completely determine what file is wanted. The best information retrieval systems can do is assign probabilities to various files given the query. Let us denote the probability that a particular file is a target by P[A].

In deciding what to do informational retrieval systems have to balance two costs. One is what Salton an McGill call the precision cost and which I will denote C_P. This is the cost associated with retrieving a file which is not a target. There must be a corresponding cost in the human system. This is the one place where we will see a computational cost appearing in our rational analysis of memory.

The other cost Salton and McGill call the recall cost and we will denote it C_R. It is the cost associated with failing to retrieve a target. Presumably in most cases it is much larger than the precision cost for a single file or memory.

Given this framing of the information processing problem we can now proceed to specify the optimal information processing behavior. This is to consider memories (or files) in order of descending P[A] and stop when the expected cost associated with failing to consider the next item is greater than the cost associated with considering it or when

$$P[A]\,C_R < (1\text{-}P[A])\,C_P \qquad (1)$$

We now have a complete theory of human memory except for one major issue--how should the system go about estimating P[A]. I propose that the system should use the item's past history of usage and the elements in the current context to come up with a

Bayesian estimate of that probability. A particularly transparent way of stating this is with the Bayesian odds ration formula which we can state

$$\frac{P(A|H_A\&Q)}{P(\bar{A}|H_A\&Q)} = \frac{P(A|H_A)}{P(\bar{A}|H_A)} \cdot \prod_{i \in Q} \frac{P(i|A)}{P(i|\bar{A})} \qquad (2)$$

where $P(A|H_A\&Q)$ is the posterior probability that the memory is needed given its past history and the cues in the current context. $P(\bar{A}|H_A\&Q)$ is 1-$P(A|H_A\&Q)$. $P(A|H_A)$ is the posterior probability given just the history. $P(\bar{A}|H_A) = 1\text{-}P(A|H_A)$. $P(i|A)$ is the conditional probability that i would be in the current context if A is needed, and $P(i|\bar{A})$ is the conditional probability if A is not needed.

This way of formulating the relationship nicely breaks up the need probability into the product of a history factor $P(A|H_A)/P(\bar{A}|H_A)$ plus a context factor the product involving the $P(i|A)/P(i|\bar{A})$. Note that in this context factor we are assuming the individual cues are independent of one another in order to obtain a product. I neither want to argue that this is really true nor that the human system actually acts as if it is. I am only using this as an approximation to get an indication of what the rational predictions are.

It should be pointed out that P[A] is the probability that A is needed, not the probability that A will be recalled if needed which is presumably much higher. The basic assumption in the discussion that follows is that the need probability will be monotonically related to observed dependent variables such as probability of recall and latency of recall. Elsewhere (Anderson, in press) I have developed detailed, and I think plausible, proposals about how need probability is related to these dependent variables but the points I will make here do not really depend on this level of detail.

The History Factor

In investigating the implications of this rational analysis for the power-law learning function we need to focus on the history factor in the above equation. In particular we need to specify $P(A\ H_A)$ To determine this we need to know how the past history of usage of a memory trace predicts whether it will be currently used. To determine this in a truly valid objective way we would have to follow people around, determine when they use particular facts, and induce what the empirical relationship is. It is nearly impossible to imagine collecting such objective statistics in the human case but such statististics are available for other information retrieval systems. For instance, there is data about how past borrowings from a library predict future borrowings (Burrell, 1980; Burrell & Cane, 1982). There is data about how past accesses to a file predict future accesses (Stritter, 1977). The data for these different domains is quite similar in terms of the nature of the functional relationship between past use and current use. I propose that these relationships are true of all information retrieval systems including the human one.

The basic point of my argument might be lost in the mathematics that follows so let me state it up front: I will show that an information retrieval system optimized in the sense defined earlier and faced with the statistics of library borrowings or file usage would produce the practice functions, retention functions, and spacing functions associated with human memory. Thus, if we accept the premise that human memory faces the same statistics as these objectively observable information-retrieval systems, we can predict its behavior with no further assumptions. The power of this analysis is that the statistics of information presentations are objectively observable and do not have to be postulated. This is in sharp contrast to a mechanistic theory where the critical structures are unobservable.

Burrell (1985) developed a model for library borrowings which provides a good analytical starting point. There are three basic assumptions in Burrell's model. The first is

that the items in a system vary in their desirability. Burrell assumes that the distribution of desirability is a gamma distribution with parameter b and index v. He is able to basically show such a distribution of borrowings in the case of a library system. The second assumption that Burrell makes is that there is an aging process such that items will decay in their borrowing rate with the passage of time. Again he can empirically validate that such an aging process does occur. This means that if we take an item from the gamma distribution with initial desirability λ its desirability after time t will be $\lambda r(t)$ where r(t) describes the rate of decay. Burrell uses a simple exponential decay in rate of the form. The third assumption of Burrell is that borrowings are a Poisson process and that times until next borrowing are exponentially distributed with rate $\lambda r(t)$.

With these assumptions we can derive what I call the recency-frequency function RF(n,t) which is the probability that an item introduced t time units ago and used n times over that period will be needed in the current time unit. It has the form:

$$RF(n,t) = \frac{v+n}{M(t)+b} r(t) \qquad (3)$$

where n is the number of borrowings in the past and M(t) is defined

$$M(t) = \int_0^t r(s)ds \qquad (4)$$

This gives us a linear relationship between number of uses, n, and need probability. This is a special case of a power function where the exponent is 1. Newell & Rosenbloom (1981) note that such hyperbolic functions give reasonable fits to human practice functions. Under the transformation from need probability to latency proposed in Anderson (in press) the power function relationship remains although there are a family of functions with different exponents.

To account for the spacing effect I have found it necessary to augment Burrell's model with two further assumptions both of which can be verified in the case of library systems but which were unimportant for Burrell's concerns. One is that there is variation in rate of decay. In the library system this is the distinction between the classics and the flash-in-the-pans. The second assumption is that items undergo periodic revivals in which they return to their original rate of usage. At Carnegie-Mellon, for instance, this happens when a course is offered which the book is relevent to. In my modelling I have simply assumed that there was an exponential distribution in decay rates and that revivals were also a Poisson process. Unlike Burrell's original assumptions I have no evidence that these forms are accurate for library systems or any other information retrieval system. Therefore, these additional assumptions must be viewed as approximate.

These additional assumptions eliminate the simple closed form solutions of Equation 3 but do not upset the prediction of power-function practice. Figures 2 illustrate the results derived from the more complex model. In addition to the practice function, it is also the case that the theory predicts typical retention functions. Despite the fact the decay process, r(t), is exponential, the effect of the revival component is to slow down the forgetting function to approximate the power-function relationship that is typically obtained between delay and retention. These figures illustrate the predictions for the dependent variables of probability (Figure 2) and (Figure 3) but similar functions are obtained if we look at the underlying need probability.

Insert Figures 2 & 3 about here

With these assumptions in hand I tried to model the classic data of Glenberg (1976) on the spacing effect. He varied the interval between two presentations of an item and looked at the effect on the recall of the item. He showed that the effect of the spacing

interval interacted with the time between the second study and test His data and the predictions of the theory are shown in Figure 4. In both cases at short test lags there is a negative relationship between spacing and recall while at long test lags there is the more common positive relationship. (Glenberg's data is a little strangely behaved at 0 and 1 study lags apparently because of inattentiveness).

Insert Figure 4 about here

Thus, we have shown that power law learning, forgetting, and the spacing (or massing) effect can all be predicted from a single rational perspective which sees human memory as adapting to the statistics of information use. Thus, it is what is outside the human head not what is inside that is controlling the memory performance. I should emphasize that this does not deny that chunking may be one of the mechanisms the mind uses to achieve this adaption. However, the argument is that the real explanation is in the outside world and not in the internal mechanisms.

ACT*--The Fan Effect

Now I would like to turn to the second architecture, ACT*, and consider a signature phenomenon which has played a key role in its development. This is the fan effect (Anderson, 1983). The fan effect has been most typically studied in a sentence recognition experiment where the subject is asked to study a set of sentences such as the following:

1. The doctor is in the bank (1-1)
2. The fireman is in the park (1-2)
3. The lawyer is in the church (2-1)
4. The lawyer is in the park (2-2)

In these materials we are manipulating the number of facts studied about the person and the location. Each sentence above is followed by two numbers giving its classification according to number of facts associated with subject and location.

Figure 5 shows the network representation that we assume in the ACT* theory that the subject sets up to encode this material. There are proposition nodes which are connected by labelled associations to each of the concepts. Note that as we increase the number of facts associated with a concept we increase the number or fan of arrows leading from the concept.

Insert Figure 5 about here

A typical experiment is focused on the subject's ability to recognize these sentences after they have been learned. A subject might have to recognize these sentence when mixed in with distractors like "The doctor is in the church". According to ACT*, upon being presented with a sentence such as "The lawyer is in the park" the subject activates the concepts in the sentence such as lawyer, in, and park. Activation spreads from these concepts along various network paths. The time to recognize a sentence is a function of the amount of activation reaching the proposition node. The critical additional assumption in the ACT* theory is that the amount of activation that can spread out of a node is fixed and that the more paths emanating out of a concept the less activation can go to any one proposition and so the slower recognition will be. Table 1 shows some data confirming this prediction. There we have data classified according to the fan associated with subject and with location.

Insert Table 1 about here

A Rational Analysis of the Fan Effect

We can extend our rational analysis of memory to accomodate the fan effect. Here we will be interested in analyzing the context factor rather than the history factor since we are manipulating properties of the memory cues that we presented to subjects. That is we

want to focus on the quantities $P(i|A)/P(i|\bar{A})$. We can rewrite these as

$$\frac{P(i|A)}{P(i|\bar{A})} = \frac{P(A|i)P(i)/P(A)}{P(\bar{A}|i)P(i)/P(\bar{A})}$$

The P(i) drop out. Since $P(\bar{A})$ must be near one (there are millions of traces and no one can be very probable) it can also be ignored. To an approximation we can also ignore $P(\bar{A}|i)$. This is a good approximation to the extent that the probability of needing a trace remains low even in the presence of a predictive cue. If we allow this approximation we get the following which is very easy to analyze:

$$\frac{P(i|A)}{P(i|\bar{A})} \simeq \frac{P(A|i)}{P(A)} \tag{5}$$

Our claims do not depend on making this approximation. It is just that they are a lot easier to see with the approximation.

In our experiments P(A) is basically constant for all items and so the critical factor turns out to be the probability that the trace is relevent given a particular cue. This is precisely what is manipulated by fan in a typical experiment. The more facts associated with a particular concept the less likely any one is given the concept. Basically if the fan is n the probability is 1/n. Anderson (1976) did an experiment that decorrelated fan and probability by manipulating the probability of testing various facts associated with a particular concept. That experiment showed conclusively that the critical factor is probability and not fan.

Thus, the fan effect is a consequence of memory using the correlation between cues and a memory's relevence to predict when the memory is needed. It may be that spreading activation is one of the mechanisms that the mind uses to compute the correlation. However, for current purposes the critical fact is that once again the

explanation of the phenomena lies in what is outside of the human head and not what is inside.

PDP -- Categorization

PDP models involve representing knowledge in a distributed form where specific experiences do not have specific encodings. On the other hand PDP models do learning locally such that changes in strengths of connection between specific elements must underlie these distributed encodings. This leads PDP models to naturally produce generalization phenomena such that they extract central tendencies out of the experience of specific instances. In introducing PDP models, McClelland, Rumelhart, & Hinton (1986) give a lot of play to categorization phenomena which is the identification of common categories in a set of tendencies. It receives more page space in their article than any other phenomena. There is a substantial literature in cognitive psychology on categorization behavior. McClelland et al. do not actually simulate any specific experiment in this literature but rather offer a simulation of the extraction of the characteristics of the members of two gangs (the jets and the sharks) as a prototype of the experiments in the literature.

To develop a rational analysis of categorization behavior the first thing we need to ask is what are the goals of the cognitive system in forming categories. In much of the experimental literature on categorization one gets the feeling that the driving force behind categorization is some sort of social conformity--that we need to learn to use the same labels to describe objects as do other people. However, this clearly cannot be all of the picture, particularly because people can learn to identify categories in the absence of any labels. I think the real function of categorization is to maximize the system's ability to predict properties of objects including their labels. Clearly, a system that can make accurate predictions will be in a position to maximize its goals.

The reason people form categories to maximize prediction is because of the nature of objects in the external world. Formally, the following is the characterization that I will assume in my rational derivations. I will assume that the world seen so far has consisted of n objects which are partitioned into s disjoint sets or categories. Each object can be classified according to some r dimensions (for simplicity I will only consider cardinal dimensions) where each dimension i has some m_i values. The members of a category belong in that category by virtue of possessing theoretical probabilities p_{ij} that they will display value j on dimension i. These probabilities provide the intensional definition of a category in contrast to its extensional definition which can be gotten simply by listing the category members.

These assumptions are intended as descriptions of the external world not just of the perception of the world in the human head. One can ask why the objects in the world should partition themselves in disjoint partitions defined by conjunctions of features. I cannot say I know the total answer but there are some obvious things to point at. For instance there is the genetic phenomenon of species which enforces a disjoint (no crossbreeding) partitioning of conjunctively defined categories (the common genetic code within a species). Other types of objects like physical elements and tools tend to produce similar disjoint partitionings of conjunctively defined categories. One can also question the probabilistic definition of category membership since this is in contradiction to the tradition in the artificial intelligence work on categories. However, I think it is indisputable that category members do display their features with only certain probabilities. Most labradors are black and have four legs but neither feature is displayed universally.

An Ideal Algorithm for Categorization

Given the formalization above we can go to characterizing what the ideal algorithm would be for categorization ignoring computational costs. Our basic situation is that the

system has observed n objects and their features and is presented with a n+1 st object with at least one feature missing and must predict the probability that it will display value j on dimension i. The following equation is the obvious one for that prediction

$$Pred_{ij} = \sum_{X} P(X|F)\, P(ij|X) \qquad (6)$$

where the summation is over all partitions X of the n+1 objects into categories. P(X|F) is the probability of that partition given the observed features of the n+1 objects, and P(ij|X) is the probability that the n+1st object will display value j on dimension i if X is the partitioning.

The problem with this ideal solution is the number of partitions grows exponentially with n. I have not been able to find the closed form expression but the number of partitions is approximately $(n+2)!/(3*2^n)$. Thus, for instance, there are the following 15 partitions of the 4 objects abcd: (abcd) (a,bcd) (b,acd) (c,abd) (d,abc) (ab,cd) (ac,bd) (ad,bc) (a,b,cd) (ab,c,d) (a,c,bd) (ac,b,d) (a,d,bc) (ad,b,c) (a,b,c,d).

It is entirely unreasonable to suppose that the human system could correspond with the prescriptions of this algorithm if that meant computing the value exactly. The human system may have some way of approximating the ideal algorithm. I have no proof that computing the quantity in Equation 6 is np-complete. For all I know there is an equivalent calculation which is computationally tractible.

An Iterative Algorithm for Categorization

Despite the lingering possibility that the ideal algorithm may have attractable form, research in machine learning has failed to find such an algorithm and the new trend is for iterative algorithms (e.g., Lebowitz, 1987; Fisher, 1987). I have worked with the following iterative algorithm:

1. Initialize the partitioning to be the empty set.

2. Given a partitioning of the first m objects calculate for each category K the probability P_k that the m + 1st object comes from category K. (Let P_0 be the probability that the object comes from a new category.)

3. Create a partitioning of the first m + 1st objects with the object assigned to the category with the maximum probability.

4. To predict value j on dimension i for the n + 1st object

$$Pred_{ij} = \sum_K P_K P(ij|K) \tag{7}$$

To apply this algorithm we need to derive rational formulas for P_K and P(ij|K). The latter is involved in the former so I will simply present a rational analysis of P_K. Again we can derive a Bayesian analysis of this quantity:

$$P_K = P(K|F_{m+1}) = \frac{P(K)P(F_{m+1}|K)}{\sum_I P(I)P(F_{m+1}|I)} \tag{8}$$

where $P(K|F_{m+1})$ is the probability that the m + 1st object belongs to category K given that it has feature structure F_{m+1}.

where P(K) is the prior probability that the object comes from category K

$P(F_{m+1}|K)$ is the probability of feature structure F_{m+1} given the object comes from category K

In deriving P(K) we are interested in the prior probability that two objects will be in the same category in advance of information about their features. A reasonable constraint to place on any formula for P(K) is that the probability that two objects find themselves in

the same category be independent of the the total number of objects to be categorized. Let this be the coupling probability which we will call c. It can be shown that there is only one formula satisfying this constraint and this is

$$P(K) = \frac{cn_K}{(1-c)+cm} \tag{9}$$

where c is a coupling probability
n_K is the number of objects in category K
m is the total number of objects

In addition, we need the following formula for P(0) the probability that the m + 1st object comes from an entirely new category

$$P(0) = \frac{1-c}{(1-c)+cm} \tag{10}$$

The remaining quantity to specify is the conditional probability $P(F_{m+1}|K)$ that the m + 1st object will display its feature structure given that it comes from category K. In developing an analysis of this quantity we will assume as an approximation that the probability of displaying a value on one dimension is independent of the probability on another dimension. If so we can have the following mathematical development:

$$P(F_{m+1}|K) = \prod_I P(ij|K) \tag{11}$$

where P(ij|K) is the probability of displaying value j on dimension i. This turns on our assumptions about the joint density function $f_i(x_1, x_2, \ldots x_m)$ which is the probability density that $p_{i1} = x_1$, $p_{i2} = x_2, \ldots p_m = x_m$.

Recall that p_{ij} is the theoretical probability that an item in a category will display value j on dimension i. If we assume a uniform density.

$$P(ij|K) = \frac{c_{ij} + 1}{n_K + m_i} \qquad (12)$$

where n_K is the number of objects in category K

c_{ij} is the number of objects in category K with the same value as the object to be classified

m_i are the number of dimensions on dimension i

It turns out the iterative algorithm so defined does a very credible job of categorization. It does a good job of classifying the 630 soybeans diseases of Michalski and Chilausky (1980) which have been a standard test case in artificial intelligence. Table 2 illustrates a simpler example structure which it has been applied to. Here we have 20 animals classified according to 10 binary dimensions. The values were made up by me and in retrospect they have some errors. Nonetheless, depending on the values set for c it merges all 20 into one category, breaks the 20 into two sets of the 10 mammals and 10 birds, further subdivides whales, humans, and seals as a subcategory of animals (by accident and mistake I gave these three mammals the same binary feature description), or divides the objects into 20 separate categories.

Insert Table 2 about here

Psychological Accuracy

Of more interest than how this does as an artificial intelligence algorithm is the question of how well it does as a model of human categorization behavior. I have applied it to the now classic data of Medin and Schaffer (1978) where it did better than their original model using only a single parameter, c, rather than their many. I have also applied it to the long series of experiments involving the Posner and Keele (1968) stimuli using an encoding of these materials developed by Hintzman. It accounts for all the phenomena that

Hintzman lists for these materials. I have also successfully predicted the results of a complicated experiment of Elio and Anderson (1981) which no model before Hintzman's was able to account for.

Rather than discussing the specific experiments in detail it is worthwhile listing some of the major phenomena that are known about human categorization and explaining how the model accounts for each:

1. Clearly the research indicates that, to a degree, people extract the central tendency of a set of instances in that their behavior is a function of the distance from that central tendency. This simply reflects a sensitivity to the statistical correlation between features and category identity which amounts to using conditional probabilities in a Bayesian analysis.

2. In addition to distance from a central tendency the literature has found an effect of distance from specific examples (e.g., Medin & Schaffer, 1978). This is produced by the tendency of the model to break diverse categories into subcategories where the features cluster together. The reason for this is that predictive power is gained by such decomposition.

3. It has shown that when a category has multiple central tendencies subjects can pick this up (Neumann, 1977). As with point (2) this is produced by the tendency to break a large diverse category into smaller categories that increase predictability.

4. Research such as that of Medin & Schaffer has shown that categorization is a non-linear function of similarity--that the increase in performance as we go from two to three matching features is greater than the increase in going from one to two. This can be traced back to equation (11) where probabilities (measuring similarity) multiply rather than

add.

5. There is an effect of category size as was discussed with respect to the Posner & Keele task. This is simply a sensitivity to base rates.

6. Rosch, Mervis, Gray, Johnson, & Boyes-Braem (1976) have documented the many circumstances in which there appear to basic level categories. The existance of such categories in our framework is simply a consequence of the fact that these categories maximize the predictability of the world--which is basically Rosch's original point.

7. It is not necessary for feedback on category membership to be given in order for categories to emerge (Fried & Holyoak, 1984). Categories will emerge any time they increase the predictability of the universe. However, by applying category labels we increase the amount of structure that can be predicted and so enhance the value of category membership. So, labels should enhance categorization but are not essential.

8. The more things that can be predicted from category membership the more likely a category is to be formed even though this means one has to learn more about a category (Billman, 1983).

Thus it seems that categorization phenomena can be again explained from a rational perspective assuming that the controlling factor is the structure of the world and not the structure in the human head. Note again this analysis does not deny that PDP mechanisms may be the way that the mind implements this rational analysis. However, it denies that PDP models provide an adequate explanation of the phenomena.

Conclusions

In summary we have looked at three cognitive architectures. For each we have have taken a signature phenomenon and developed a reasonable model of the world in which that phenomenon occurs and the goals of humans operating in that world. We have made a few assumptions about computational costs which are not at all mechanism specific. We have derived the signature phenomena as solutions to the optimization problems we defined. In each case this rational analysis led to an account that was as accurate or more accurate than the original mechanistic account.

Now we come to the hard question of what the implications are of these demonstrations. I am not really sure what the implications are but I will hazard two guesses. However before I do I want to forestall misunderstanding by disavowing two possible interpretations of the point of this paper.

One possible reaction to the relative good showing of the rational analysis might be renewed effort to develop a better cognitive architecture. There is a tendency to view this rational analysis as a first-order approximation which any self-respecting architecture ought to do better than. These results might be thus be taken as damning the three architectures we considered rather than praising rational analysis. However, I think simply looking for a better architecture really misses the point. First it does not deal with the fundamental identifiability problems that haunt our search for such mental mechanisms. More important it looses the essential insight that it is no accident that architectures which correspond to human behavior compute in the way they do. They do so because this is in fact what is optimal given the world in which they reside.

A second reaction might be to take this as an indictment of mechanistic accounts of mind and a call for a retreat to behaviorism. After all, the argument might go, we have

shown that human behavior can be predicted by reference to the environment without reference to what is in the head. However, to retreat back to behaviorism would be to leave us with the same computationally inadequate models of mind that we abandoned 30 years ago. The simple fact is that the optimal behavior is often going to be computationally complex and mechanistic accounts give us a way of expressing that complexity and simple stimulus-response associations do not.

While I am confident that the above two are the wrong reactions I am less certain about the positive proposals I have to make, but here they are: My first guess is that cognitive architectures should be viewed as notations for expressing the behavioral functions that emerge as the solutions to the optimization problems in a rational analysis. The real theory lies in the assumptions made in the statement of the optimization problem--i.e., the assumptions about the goals, the world, and the computational limitations. These assumptions do not have the same identifiability problems that the mechanistic models do and lead to a much deeper explanation of the phenomena at hand. However, something computationally powerful like a Turing-equivalent architecture is necessary if we are going to be able to express the solution to these optimization problems.

Thus the theory is in the framing of the information processing problem and the architectures provide notation for expressing the solutions to the optimization problems. I see a one-to-many mapping between framings and architectures. That is, one can take a single framing and for every architecture find some configuration of its mechanisms that enable the optimal behavior to be computed. Choice among architectures is then not to be determined by veracity of empirical predictions. Rather it is to be determined by how easy it is to work out the optimal behavior in that architecture. Ease of use is the classic criterion for selecting among notations. Empirical veracity is reserved for theories.

My second guess (which is a variation on the first guess) is that architectures in some form may play a role in actually framing the optimization problem. Recall earlier that step (3) in developing a rational analysis was to make some assumptions about computational cost. In the case of memory, the assumption was that there was a retrieval cost. In the case of categorization the assumption was that a certain function was not computable and another was. These were relatively bland assumptions but they do reflect the architecture. It is possible in other applications of a rational analysis, the computational assumptions might be richer. On this view, much of the detail we associate with an architecture might be just theoretical notation, but there may be some core, contentful assumptions. This view would encourage us to sift the notation from the content in our architectures, using relevence to rational analysis as a basis for making that discrimination. Basically, architecture would define the bounds on optimization in a rational analysis or, in Simon's hillclimbing metaphor, define the contours of the surface on which the local optimization takes place.

Table 1

Person-Location Experiment (A hippie is in the park)—Mean Verification Times and Error Rates for Trues and Falses

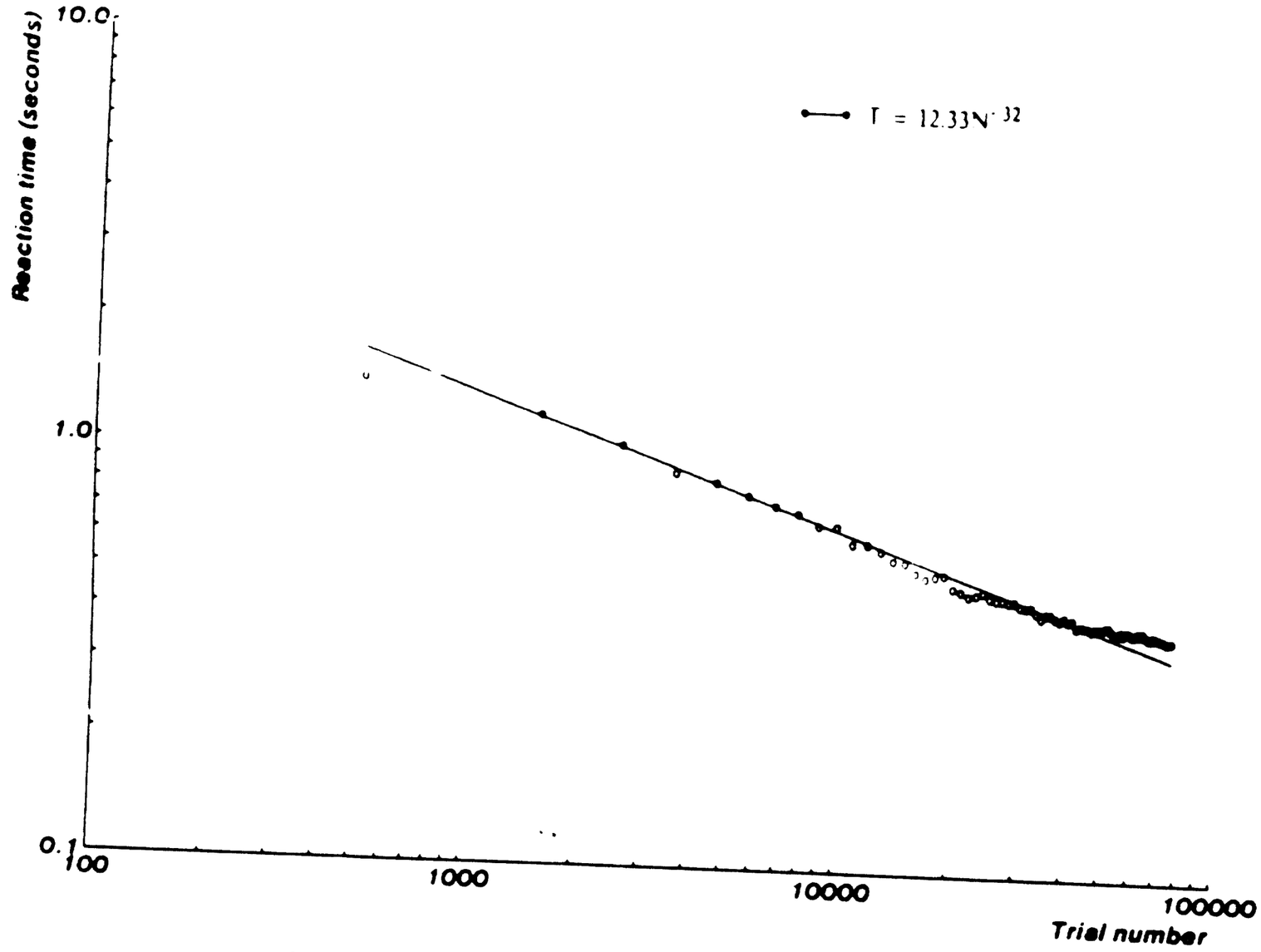
		Trues: Number of propositions per person				Falses: Number of propositions per person			
		1	2	3	Mean	1	2	3	Mean
Number of propositions per location	1	1.111 (.051)	1.174 (.042)	1.222 (.046)	1.169 (.046)	1.197 (.019)	1.221 (.042)	1.264 (.030)	1.227 (.030)
	2	1.167 (.065)	1.198 (.056)	1.222 (.060)	1.196 (.060)	1.250 (.014)	1.356 (.037)	1.291 (.044)	1.299 (.032)
	3	1.153 (.063)	1.233 (.044)	1.357 (.054)	1.248 (.054)	1.262 (.042)	1.471 (.079)	1.465 (.051)	1.399 (.057)
	Mean	1.144 (.059)	1.202 (.048)	1.267 (.054)	1.204 (.053)	1.236 (.025)	1.349 (.053)	1.340 (.042)	1.308 (.040)

Table 2

Features I

Animals	Hair	Light	Color	Lay	Eggs	Mammals	Four	Legged	Fins	Big	Aggressive	Has A	Beak	Gives	Milk
WHALES	0	1	0	1	0	0	1	0	0	1					
SEALS	0	1	0	1	0	0	1	0	0	1					
DOGS	1	0	0	1	1	0	1	1	0	1					
CATS	1	0	0	1	1	0	0	1	0	1					
HORSES	1	0	0	1	1	0	1	0	0	1					
BEARS	1	0	0	1	1	0	1	1	0	1					
BATS	1	0	0	1	0	1	0	0	0	1					
HUMANS	0	1	0	1	0	0	1	0	0	1					
MICE	1	1	0	1	1	0	0	0	0	1					
PLATYPUS	1	0	1	1	1	0	0	0	1	1					
CHICKENS	0	1	1	0	0	0	0	0	1	0					
PENGUINS	0	0	1	0	0	0	1	0	1	0					
ROBINS	0	0	1	0	0	1	0	0	1	0					
OSTRICHES	0	1	1	0	0	0	1	0	1	0					
CROWS	0	0	1	0	0	1	0	0	1	0					
PARROTS	0	0	1	0	0	1	0	0	1	0					
SPARROWS	0	1	1	0	0	1	0	0	1	0					
EAGLES	0	1	1	0	0	1	1	1	1	0					
HAWKS	0	1	1	0	0	1	0	1	1	0					
SEAGULLS	0	1	1	0	0	1	0	0	1	0					

Figure 1



Learning in a ten finger, 1023 choice task (log-log coordinates).
Plotted from the original data for Subject JK (Seibel, 1963).

Figure 2

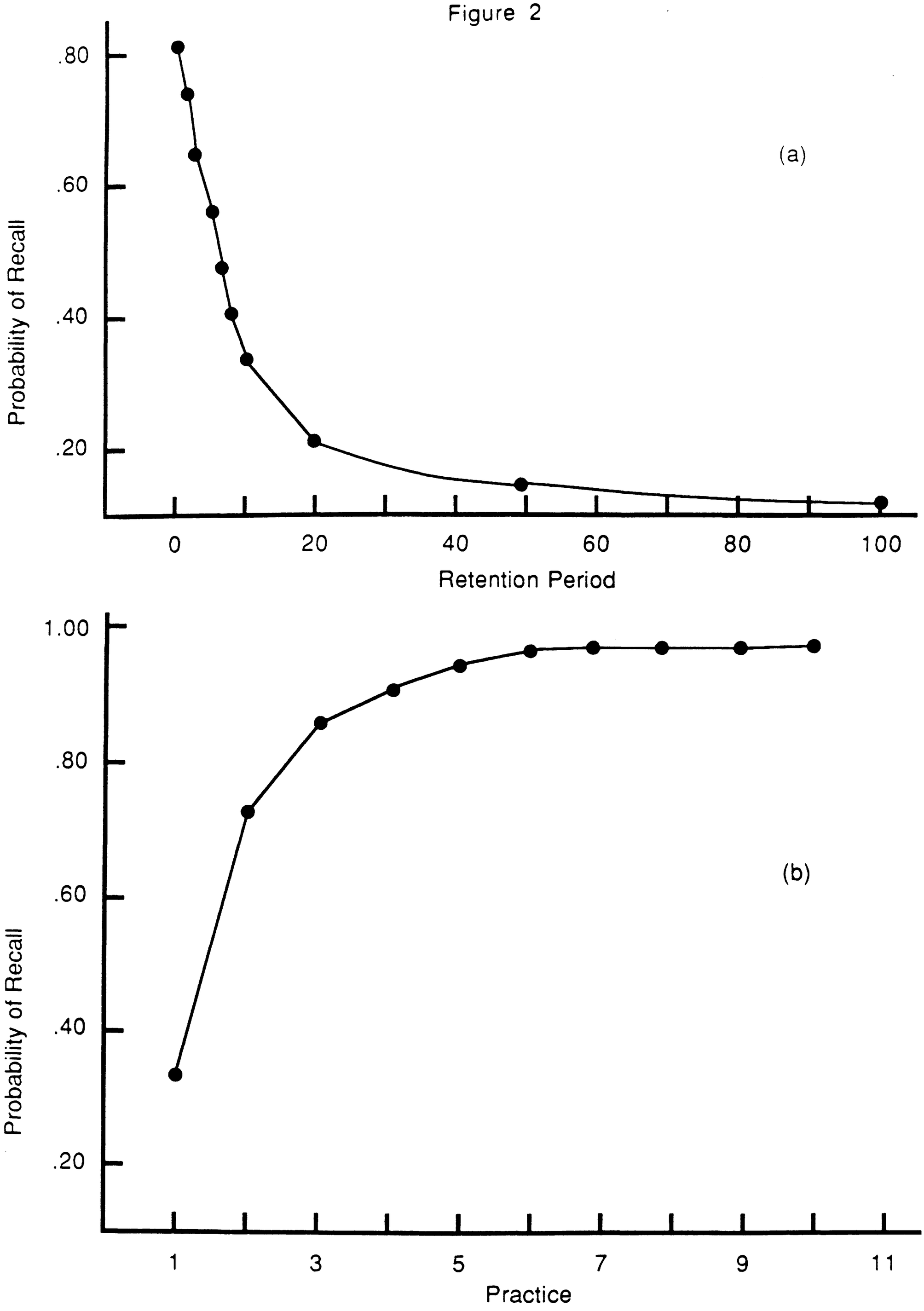

.80
.60
.40
.20
Probability of Recall
0
20
40
60
80
100
Retention Period
(a)
1.00
.80
.60
.40
.20
Probability of Recall
1
3
5
7
9
11
Practice
(b)

Figure 3

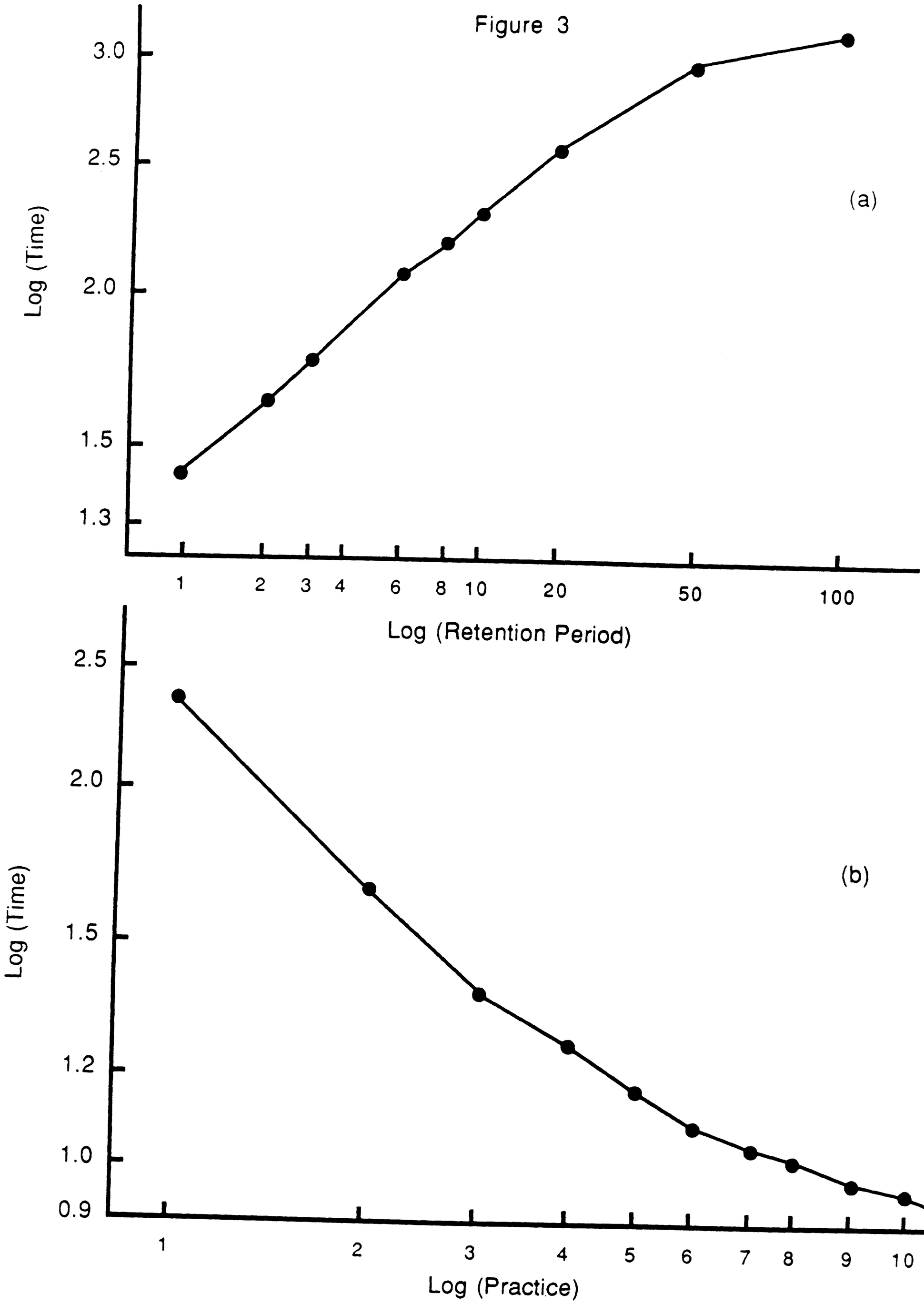
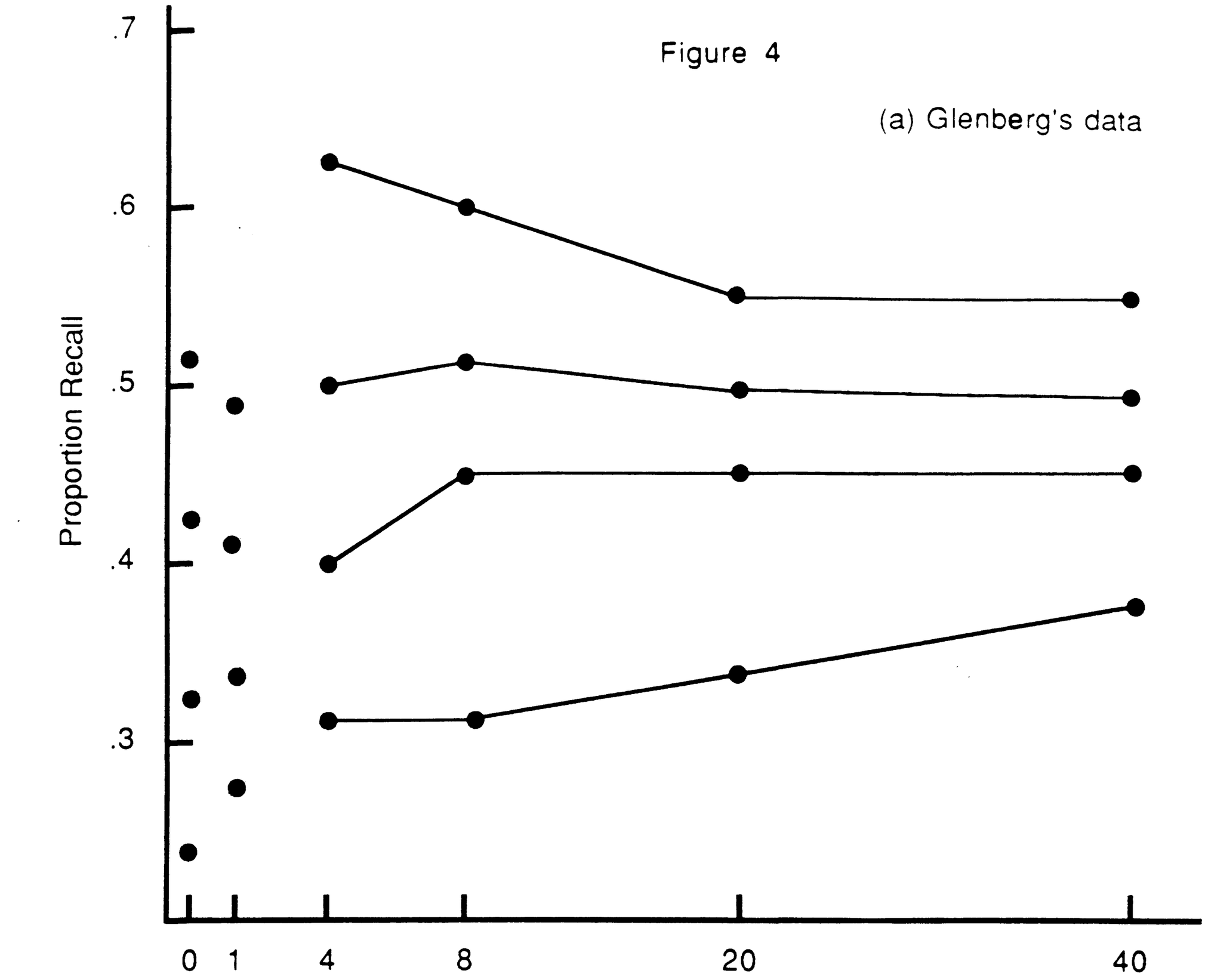

Log (Time)
3.0
2.5
2.0
1.5
1.3
1
2
3
4
6
8
10
20
50
100
Log (Retention Period)
(a)
Log (Time)
2.5
2.0
1.5
1.2
1.0
0.9
1
2
3
4
5
6
7
8
9
10
11
Log (Practice)
(b)

Figure 4

(a) Glenberg's data



(b) Theory

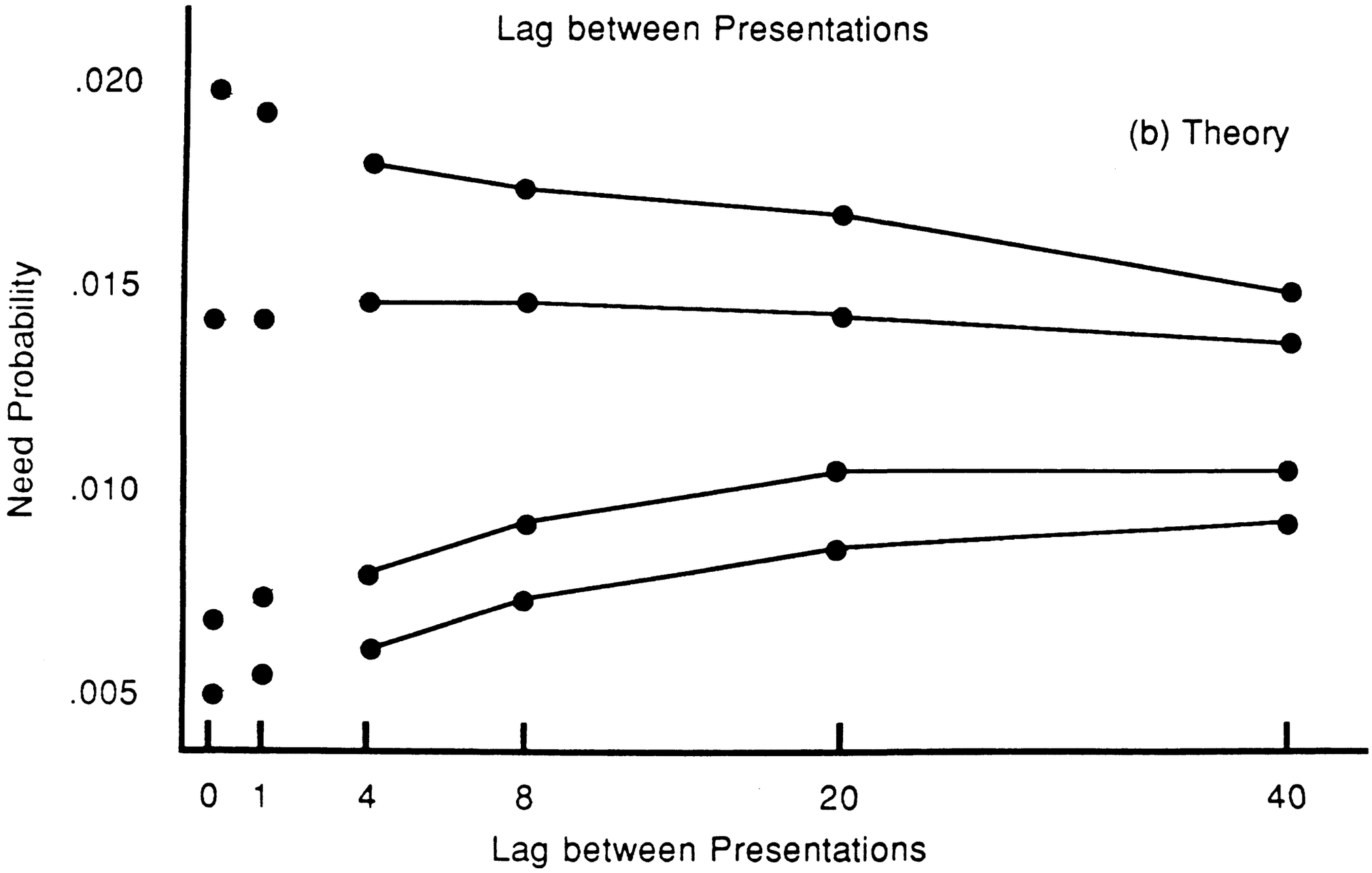
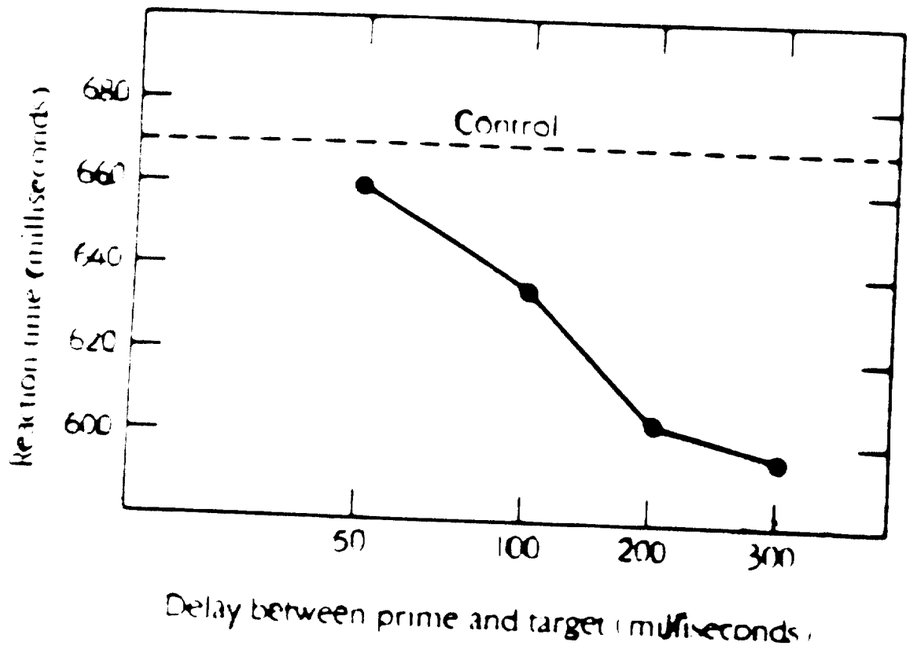


Figure 5



Difference between primed and control conditions as a function of the interval between priming word and target word

Figure Captions

Figure 1 Data from Siebel (1963) plotting time to respond against amount of practice

Figure 2 Relationship between probability of recall and retention interval (a) and practice (b)

Figure 3 Relationship between latency of recall and retention interval (a) and practice (b). These are log-log plots to show the characteristic power functions.

Figure 4 (a) Glenberg's data showing the interaction between retention interval and study log: (b) Predictions of the theory for Glenberg's experiment.

Figure 5 ACT* propositional network representation of the fan material.

References

Anderson, J. R. (1976). *Language memory and thought*. Hillsdale NJ: Erlbaum.

Anderson, J.R. (1982). Acquisition of Cognitive Skill. *Psychological Review*, 89, 369-406.

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J.R. (1983). Retrieval of information from long-term memory. *Science*, *220*, 25-30.

Anderson, J. R. (1987). Methodologies for studying human knowledge. *The Behavioral and Brain Sciences*, *10*, 467-505.

Anderson, J. R. (in press). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Billman, D. (1983). *Inductive learning of syntactic categories*. Doctoral dissertation, University of Michigan, Ph.D. dissertation.

Burrell, Q. L. (1980). A simple stochastic model for library loans. *Journal of Documentation*, *36*, 115-132.

Burrell, Q. L. (1985). A note on aging on a library circulation model. *Journal of Documentation*, *41*, 100-115.

Burrell, Q. L. & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society, Series A*(145), 439-471.

Elio, R. & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 397-417.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139-172.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for

classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234-257.

Glenberg, A. M. (February, 1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*(1), 1-16.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning, 2*, 103-138.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McClelland, J. L., Rumelhart, D. E., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: Bradford Books.

McClelland, J. L., Rumelhart D. E., & Hinton, G. E. (1986). *Parallel Distributed Processing*. Vol. 1: *The appeal of parallel distributed processing*. In D. E. Rumelhart & J. L. McClelland (Eds.).

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Michalski, R. S. & Chilausky, R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems, 4*, 125-161.

Neumann, P. G. (1977). Visual prototype information with discontinuous representation of dimensions of variability. *Memory & Cognition, 5*, 187-197.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition, pp. 1-55*. Hillsdale, NJ: Erlbaum.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of*

Experimental Psychology, 77, 353-363.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 7, 573-605.

Rosenbloom, P. S., Laird, J. E., & Newell, A. (In Press). *Working Models of Human Perception: The chunking of skill in knowledge*. London: Academic Press. In H. Buoma & B. A. G. Elsendoorn (Eds.).

Rumelhart, D. E., McClelland, J. L., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: Bradford Books.

Salton, G., & McGill, M. J. (1983). *Introduction to modern infromation retrieval*. New York: McGraw-Hill.

Siebel, R. (1963). Discrimination reaction time for 1023--alternative task. *Journal of Experimental Psychology*, 66, 215-226.

Simon, H. A. (1972). *Decision and Organization. Theories of bounded rationality*. Amsterdam: North-Holland. In C. B. Rander & R. Radner (Eds.).

Stritter, E. P. (1977). *File migration*. Doctoral dissertation, Stanford University. Stanford: Stanford Linear Accelerator Center.

Vining, D. R. (1986). Social versus reproductive success: The central theoretical problem of human sociobiology. *The Behavioral and Brain Sciences*, 9, 167-216.