IMPLICIT CONCEPT FORMATION

Zoltan Dienes

Wolfson College

BA (Hons) University of Cambridge

MA (Hons) Macquarie University

A thesis submitted for the degree of Doctor of Philosophy

University of Oxford

For Jules

Acknowledgements

I am indebted to my supervisors, Donald Broadbent and Dianne Berry, for their conscientious guidance and constant support in shaping this thesis. Both have been valuable mentors in the ways of being a psychologist.

My thanks to the people who assisted at the various stages of conducting this research. I am grateful to Don Dulany for making his raw data available to me. I am grateful to Alexandros Trevis and Vernon Dobson for reading Chapter Five and providing valuable feedback. I am grateful also for the informative discussions with Alan Allport, Cheng Chan, Jon Driver, Jeanette Garwood, Miles Glen, Joncks, Jeff Miller, and Peter McLeod.

Implicit concept formation

Zoltan Dienes

Wolfson College, Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term, 1990

Abstract

This thesis provides a conceptual and empirical analysis of implicit concept formation. A review of concept formation studies highlights the need for improving existing methodology in establishing the claim for implicit concept formation. Eight experiments are reported that address this aim. A review of theoretical issues highlights the need for computational modelling to elucidate the nature of implicit learning. Two chapters address the feasibility of different exemplar and Connectionist models in accounting for how subjects perform on tasks typically employed in the implicit learning literature.

The first five experiments use a concept formation task that involves classifying "computer people" as belonging to a particular town or income category. A number of manipulations are made of the underlying rule to be learned and of the cover task given subjects. In all cases, the knowledge underlying classification performance can be elicited both by free recall and by forced choice tasks.

The final three experiments employ Reber's (e.g., 1989) grammar learning paradigm. More rigorous methods for eliciting the knowledge underlying classification performance are employed than have been used previously by Reber. The knowledge underlying classification performance is not elicited by free recall, but is elicited by a forced-choice measure. The robustness of the learning in this paradigm is investigated by using a secondary task methodology. Concurrent random number generation interferes with all knowledge measures.

A number of parameter-free Connectionist and exemplar models of artificial grammar learning are tested against the experimental data. The importance of different assumptions regarding the coding of features and the learning rule used is investigated by determining the performance of the model with and without each assumption. Only one class of Connectionist model passes all the tests. Further, this class of model can simulate subject performance in a different task domain.

The relevance of these empirical and theoretical results for understanding implicit learning is discussed, and suggestions are made for future research.

TABLE OF CONTENTS

ABSTRACT

This thesis aims to provide a conceptual and empirical analysis of implicit concept formation. Chapter One reviews relevant empirical research and previous theoretical developments. The review proceeds by defining implicit knowledge according to the specificity of its transfer and its distinct properties and learning characteristics as compared to explicit knowledge. A review of concept formation studies highlights the need for improving existing methodology in establishing the claim for implicit concept formation. Accordingly, the first major aim of this thesis is to develop concept formation tasks to test the specificity of the knowledge underlying classification performance, and to determine its distinctive characteristics. Chapters Two, Three, and Four address this aim.

A review of theoretical issues highlights the need for computational modelling to elucidate the nature of implicit learning. Specifically, the debate over the role of exemplar storage and deployment in implicit learning could be advanced by determining the fit of simulations of exemplar models to implicit concept formation data. Further, although it has been claimed that the Connectionist algorithms result in implicit knowledge, there has not been a sustained attempt to simulate with Connectionist networks tasks that are often used in the implicit learning literature. Two tasks for which a wealth of performance data already exist are artificial grammar learning (e.g., Reber, 1989) and the control tasks (e.g., Berry & Broadbent, 1984), and so these tasks would be suitable for modelling. Thus, the second major aim of the thesis is to determine the feasibility of different exemplar and Connectionist

models in accounting for how subjects perform the artificial grammar learning and control tasks. Chapters Five and Six address this aim.

Chapter Two presents the results of two experiments that extend existing methodology to provide more complete transfer tests of the knowledge underlying classification. Only one previous study, by Hayes (1987). demonstrated a lack of transfer of classification performance to another structured knowledge measure; unfortunately, even this measure failed to tap a highly plausible source of knowledge that could be used for classification. Thus, Experiment One employs the Residents Task of Hayes. Hayes asked subjects to classify "computer people" into one of three towns. Each computer person was described by four phrases, including one unique and three shared phrases. Subjects classified to a criterion of two thirds of the computer people correct two thirds of the time. At this point a transfer test was given in which subjects were asked to recognize the unique phrases and to indicate the town associated with each. Hayes found that the number of correct answers to this measure was significantly less than classification performance. However, a problem with the Hayes study is that in addition to the unique phrase, the combination of shared phrases was also unique to each person. Accordingly, Experiment One uses the same procedure as Hayes, with an additional transfer test in which subjects are asked to recognize unique combinations of shared phrases and to indicate the town associated with each. The main result is that knowledge of unique phrases is sufficient to account for classification performance. Although this result seems inconsistent with that of Hayes, a computer model of classification on the Residents Task reveals that in some situations chance factors can increase

classification performance. In fact, this model shows that the results of both Experiment One and Hayes are consistent with the classification and transfer tests tapping the same level of underlying knowledge.

A problem with the Residents Task, even with the extended transfer task, is that there are different probabilities for performing well on the classification and the transfer tests by chance alone. Experiment Two is a first attempt to test for implicit concept formation by appropriately adjusting these probabilities. Additionally, subjects are exposed to the information incidentally to discourage an explicit mode of learning. The stimuli consist of twelve computer people, each described by four phrases and living in one of three towns. One phrase is unique to a town (rather than to a person) and the other phrases are counterbalanced across the towns. Subjects are exposed to one person at a time and are asked to count the number of different letters in the description of the person. They are then given two blocks of classification performance, followed by the transfer test. In the transfer test, subjects underline the important sentence(s) for each town and indicate the rule connecting the sentence(s) to the town. The results show that subjects perform on the classification and transfer tests at a chance level.

Chapter Three reports three experiments that continue to probe for the possibility of implicit learning in simple concept formation tasks by addressing possible problems with Experiment Two. The task used in Experiment Two may not induce implicit learning because of the way the stimuli are displayed, because of the underlying rule to be learned, or because the cover task biases

subjects to process the stimuli at an inappropriate level (letters rather than words). Thus, the tasks used in Chapter Three address the above three points. First, the categories are not discrete (like towns) but continuous: Computer people have one of twelve graded incomes. People might tend to deal with continuous rather than discrete variables in an implicit way. Second, in addition to the rule used in Experiment Two, a new underlying rule is introduced in which two phrases, rather than one, determine category membership. Multiple rather than single cues might be best dealt with implicitly rather than explicitly. And third, the cover task involves processing words rather than letters. Experiments Three, Four, and Five incorporate these changes; Experiment Three differs from Four and Five in the nature of the cover task.

In Experiment Three, subjects have a partial report task. Four phrases are displayed describing a computer person, followed by a blank screen, and then an income. The value of the income indicates which two phrases to report. Like Experiment Two, this procedure results in chance classification and transfer test performance. In Experiments Four and Five, subjects memorize which income went with which person. Experiment Four is exploratory and has unequal probabilities for performing well on the classification and transfer tasks by chance alone. When the underlying rule involves a single unique phrase per category, performance on the classification and transfer tasks is good and equivalent on both. When the rule involves two phrases determining category membership, performance is better on the classification than the transfer task. Experiment Five introduces additional classification blocks to equalize baseline probabilities on the classification and transfer

tasks; when this is done, performance on the classification and transfer tasks is equivalent. Thus, the apparant lack of transfer in Experiment Four is artifactual.

To summarize, Chapters Two and Three present the results from five experiments that indicate an equivalence between classification performance and a structured transfer test. These results may indicate the absence of an effect to be found or the failure to establish the necessary conditions to find implicit learning. Rather than continuing to probe this issue with new paradigms, Chapter Four returns to a paradigm in the concept formation literature where it is claimed that the necessary conditions for implicit learning have been established. Specifically, the artificial grammar learning paradigm of Reber (e.g., 1976) is employed for Experiments Six, Seven, and Eight reported in Chapter Four. Reber (e.g., 1989) claims that this task is learned implicitly, but it has yet to be demonstrated that classification does not transfer to a structured knowledge test. Thus, in Chapter Four a structured transfer test similar to those used in Chapters Two and Three is applied to the artificial grammar learning task.

In the artificial grammar learning task, subjects are first asked to memorize strings of letters generated by a finite state grammar, and then are asked to classify grammatical and nongrammatical strings. In Chapter Four, two transfer tasks are used: Free recall, a distinctively explicit knowledge test, in which the subject is asked to describe as completely as possible how she actually classified; and the test of Sequential Letter Dependencies (the SLD test), the structured transfer test in which

the subject is presented with allowable initial letter sequences, of length 0 letters upwards, and asked to indicate which letters can occur next. The SLD test is similar to the Hayes (1987) transfer test and to the transfer test used in Experiments Two to Five in that it assesses the subject's ability to apply her knowledge to elements of exemplars presented in isolation. In Experiment Six, classification knowledge transfers to the SLD test, but not to free recall. This is not suprising: Free recall is not expected to be a sensitive test. More importantly, free recall fails to correlate with either classification or SLD, which do correlate with each other. This pattern of results provides only tentative evidence for distinct knowledge bases underlying free recall on the one hand, and classification and SLD performance on the other.

Experiment Seven attempts to strengthen the case for separate knowledge bases by employing a dual task methodology used by Hayes (1987; and as reported in Broadbent, 1989). Hayes found that concurrent random number generation (RNG) did not interfere with learning the artificial grammar (as measured by classification) when incidental learning instructions were used; but concurrent RNG did interfere if subjects were asked to search for rules during learning. Thus, concurrent RNG may interfere with an explicit but not implicit mode of learning. Thus, one interesting possibility is that under incidental instructions, RNG may interfere with free recall, indexing explicit knowledge, but not classification and SLD, potentially indexing implicit knowledge. In fact, Experiment Seven shows that with concurrent RNG <u>all</u> knowledge tests suffer. This result raises the possibility of different task priorities (as a function of different task demands) accounting for the interference

found by Hayes for explicitly but not implicitly instructed subjects. But the results are also consistent with distinct implicit and explicit learning modes, depending on the exact nature of the Performance Operating Characteristics (POCs) for the different knowledge measures for implicitly and explicitly instructed subjects. In Experiment Eight, task priorities are systematically manipulated for both implicitly and explicitly instructed subjects; and classification, Free Recall, and SLD measures of artificial grammar learning are taken. The results show dual task conditions interfere with all knowledge measures for both implicitly and explicitly instructed subjects. The priority manipulation per se has no influence on classification performance. Further, the results indicate that Hayes may have failed to find an effect of dual task conditions on implicitly instructed subjects simply because his study lacked power. In summary, Chapter Four shows that classification knowledge in the artificial grammar learning task can transfer to recognition judgements of part exemplars, and that there is no evidence that it represents a different type of knowledge to that elicited by free recall. Nonetheless, in that classification knowledge fails to transfer to free recall, it can be regarded as "implicit" at an everyday level of explanation.

In order to investigate the sorts of mechanisms that could learn the tasks used in Experiments Six to Eight, and to elucidate the theoretical issues raised in Chapter One, the next two chapters turn to computational modelling. Chapter Five investigates the performance of different auto associators and exemplar models in learning the artificial grammar used in Experiments Six to Eight.

An auto associator is a mechanism that attempts to predict each part of a presented pattern according to the remainder of the pattern. An auto associator can classify grammatical and nongrammatical strings according to how well each part of a string is completed. A range of auto associators are considered differing along four dimensions: The learning rule used (Hebb or Delta); letter vs digram coding; sensitivity to cooccurrence or contingency; and simultaneous vs successive prediction. Two exemplar models are considered: The MINERVA 2 model of Hintzman (1986) and the array model of Estes (1986). Both exemplar models involve determining an overall similarity between each test string and the stored exemplars. Each model is used to produce a rank ordering of string difficulty, and this is compared to experimental rank orderings (derived from Experiments Six to Eight and from Dulany, Carlson, & Dewey, 1984). The results show that the simultaneous delta rule auto associator models fare best in accounting for the empirical data. These models pass a number of tests failed by the other models. It is shown that these models could be regarded as abstracting a set of representative but incomplete rules of the grammar. Further, the knowledge used by these models for classification performance does transfer to the SLD test.

In order to determine the generality of auto association in providing a possible mechanism for implicit learning, Chapter Six investigates the performance of different Connectionist models, including auto association, in learning the dynamic control tasks. The dynamic control tasks involve the subject attempting to control the level of one or more variables by deciding on which of a number of actions to take. The subject attempts to reach and maintain the

variables at prespecified target levels; the subject's ability to do this is the measure of performance. Subjects are also asked to indicate how these variables will be influenced by different possible subject actions; the subject's ability to do this is a measure of predictive knowledge. A key finding in the implicit learning literature is that performance and predictive knowledge need not be related (e.g., Berry & Broadbent, 1984).

Two Connectionist architectures are considered: Reinforcement learning (e.g., Barto, Sutton, & Brouwer, 1981). and auto association. In reinforcement learning, context units coding the current situation are connected to units coding the different possible actions. In a given situation, the action unit with the most activation is fired. The weight from the context units to the relevant action units is then altered according to whether the action moves the variables closer to target or not. It is shown that such an architecture can learn the dynamic control tasks. Such a system, like subjects, has poor predictive knowledge, but that is because it is hard-wired that way. A more interesting architecture is the auto associator, where all units are connected to all others. Thus, it is logically possible for the auto associator to learn not only what action to produce in a given context (performance), but also what context results from a given action (predictive knowledge). In fact, it is shown that when the dynamic control tasks are simulated by an auto associator learning by the delta rule, it can learn to perform with little predictive knowledge in some situations.

Chapter Seven draws the results together from the experimental and computational work. On concept formation tasks,

both subjects and models can respond appropriately to elements of exemplars presented in isolation (the SLD test) as well as in context (classification). Also, in both subjects and models, the knowledge need not be represented in propositional form. The implications of these results are indicated and directions for future research are suggested.

Chapter One

Introduction: Implicit learning and concept formation

Introduction

There appear to be many examples in everyday life of us learning to respond appropriately according to criteria we can readily state; for example, in learning the rules of chess. But this is not always so. There also appear to be cases of us learning to respond in some rule-like way, without us being able to say what the rules are that govern our behaviour. For example, we learn to recognize and produce grammatical utterances without being able to say what the rules of grammar are. Concept learning may provide another example. Polanyi (1969) has argued that people are generally unable to state the necessary perceptual attributes for identifying instances of a category. He argued that we may easily recognize a face, yet only inadequately describe by what particulars we recognized it; or we may diagnose a disease, and yet be unable to identify, let alone describe, the great number of details we do in fact notice in making the diagnosis. Thus, there may not be a common database for all output processes; some tasks may develop an associated database that can be elicited only by that task (Broadbent, Fitzgerald, & Broadbent, 1986). The claim that there exists such specific ("non-manipulable") knowledge, and that it has distinctive properties or learning characteristics, defines the claim for the existence of implicit knowledge, as investigated by this thesis.

This thesis provides both a conceptual and empirical analysis of implicit concept formation. The present chapter reviews relevant research. Initially, the work of the Broadbent group

indicating implicit and explicit modes of learning on dynamic control tasks is overviewed. Then, literature suggesting an equivalent implicit mode in specifically concept formation tasks is reviewed. This review highlights the need for an investigation to more firmly establish the nature, or even existence of, implicit concept formation. Next, theoretical perspectives on implicit learning are discussed. It is argued that an elucidation of the theoretical issues could usefully employ computational modelling. Finally, the logic and aims of the series of models and experiments conducted for this thesis are specified.

Implicit learning and control tasks

A substantial body of evidence for the existence of distinct implicit and explicit learning modes comes from a series of studies by the Broadbent group on "dynamic control tasks". The dynamic control tasks involve the subject attempting to control the level of one or more variables by deciding on the level of one or more others; the subject has continuous feedback of the variables to be controlled while she manipulates the controlling variables. Interest centers on the relationship between "performance", that is, the subject's ability to reach or maintain the variables to be controlled at prespecified target values, and "verbal knowledge", that is, some other measure of knowledge that could be used to account for performance. Verbal knowledge typically refers to the subject's ability to predict changes in the variables to be controlled given specified changes in the controlling variables. Such predictive knowledge is assessed by questionnaires and is regarded as tapping an explicit working model that could be used in controlling the task. Various control tasks with various different

relationships between variables have been used in this research. For example, in the person interaction task (Berry & Broadbent, 1984, 1988; Hayes & Broadbent, 1988), subjects had to manipulate a computer person's friendliness by varying their own friendliness towards the computer person on a 12 point scale. With the Clegg computer personality (Berry & Broadbent, 1984), the friendliness of Clegg is given by twice the value of the subject's friendliness minus Clegg's previous friendliness. Or in the transport task (Broadbent, 1977; Broadbent et al., 1986), subjects had to control bus load and spaces in a car park by manipulating bus schedule and parking fee, where the relationships are defined by simple linear equations.

Initial work on transport systems (Broadbent, 1977) and simulated economic systems (Broadbent & Aston, 1978) indicated that with practice performance improved but predictive knowledge did not. This result has recently been replicated by Sanderson (1989; Experiment 1, Group Int/4 trials), using a more complete questionnaire to assess predictive knowledge. The simplest explanation for these results might be that predictive knowledge is a less sensitive measure than performance of a common data base; however, Berry and Broadbent (1984) and Broadbent et al. (1986) obtained double dissociations between performance and predictive knowledge. Specifically, using nonsalient relationships between controlling and to be controlled variables, practice led to an improvement in performance but not predictive knowledge (Berry & Broadbent, 1984; Broadbent et al. 1986), as found earlier; but detailed verbal instructions on how to perform (Berry & Broadbent, 1984) and practice with a dramatic and runaway relationship (Broadbent et al. 1986) lead to an improvement in predictive

knowledge but not performance. Additionally, zero or negative correlations have been found between performance and predictive knowledge (e.g., Berry & Broadbent, 1984).

Thus, there is evidence for distinct knowledge bases on the grounds of (a) double dissociations; and (b) stochastic independence, or negative dependence. Recently, Dunn and Kirsner (1988) have argued that a "reversed association" more compellingly demonstrates separate underlying processes. A reversed association obtains when two variables are positively related in one condition, but negatively related in another. This result can be derived from Broadbent et al. (1986) and Berry and Broadbent (1988); it can also be derived from Sanderson (1989). In Broadbent et al. (1986), comparing the Single (S) with the Together (T) group of the transport studies, an increase in performance (from S to T) is accompanied by a decrease in predictive knowledge (from S to T). In Berry and Broadbent (1988), comparing the Nonsalient Control (NSC) with the Salient Experimental group (SE), an increase in performance (from NSC to SE) is accompanied by an increase in predictive knowledge (from NSC to SE). Similarly, Sanderson (1989), using the transport task of Broadbent et al. (1986), found a negative correlation between performance and predictive knowledge when all variables were displayed graphically (Experiment 2), but a positive correlation when there was extended practice with the task without graphical display (Experiment 1). These reversed associations could be interpreted in terms of a general knowledge base that can be used for both performance and verbal knowledge, but also a specific knowledge base that can be used for performance but not verbal knowledge.

The specificity of the performance knowledge base was

further highlighted by Berry and Broadbent (1988) and Berry (in preparation). Berry and Broadbent (1988) found positive transfer of performance from one task to another of the same underlying equation, if the tasks were perceptually similar (both transport tasks or both person interaction tasks) but not if they were perceptually dissimilar. The failure of transfer did not arise simply because subjects failed to realize they could apply their knowledge; informing subjects that the tasks were based on the same equation actually deteriorated performance. Berry (in preparation), using the Clegg and sugar production tasks, found positive transfer of performance to nominally the same task only if the specific way the subject responded was kept constant; that is, if the subjects typed their responses on both the first and second block of trials. There was no transfer of performance knowledge if on the first block of trials the subject spoke the responses and on the second block typed them. In sum, these results suggest that performance relies on a type of knowledge that can be applied only in certain perceptually defined situations.

Berry and Broadbent (1988) used a person interaction task to investigate different possible learning modes that could be associated with implicit and explicit knowledge. A salient computer person responded according to the last input behaviour entered by the subject; and a nonsalient computer person responded according to the input behaviour entered by the subject on the previous trial. The greater amount of implicit knowledge normally used by subjects for the nonsalient rather than salient person was shown by the low verbal knowledge for subjects who interacted with the nonsalient rather than the salient person. After initial experience with the task, instructions to search for the rule connecting the computer

person's responses to the subjects' improved performance with the salient person and deteriorated performance with the nonsalient person. Thus, different modes of learning can be applied that are more or less suitable for different types of task; Berry and Broadbent (1988) argued that for nonsalient relationships subjects learn most readily in some implicit way. Hayes and Broadbent (1988) further investigated the nature of implicit and explicit modes of learning by looking at their interaction with a secondary task. Concurrent random number generation interfered with learning an altered relationship with a salient person, but facilitated learning an altered relationship with a nonsalient person. The results imply that acquiring explicit but not implicit knowledge is retarded by a concurrent memory demanding task.

Research on the control tasks has relied heavily (though not exclusively) on predictive knowledge as the test of verbal knowledge. To the extent that performance does not require, at any level, being able to predict the consequences of one's actions (except in so far as they bring one closer to target), the lack of relationship between performance and predictive knowledge is uninformative with respect to the specificity of the knowledge underlying performance. If performance knowledge transfered to different targets, there would be evidence that the subject had implicitly acquired predictive knowledge. However, the evidence for transfer to different targets is contradictory (this evidence is mentioned on page 53), and so there is a need for using different measures of verbal knowledge with the control tasks. As a start to addressing this issue, Berry and Broadbent (1984,1987) reported that retrospectively asking subjects to write down how they went about attempting to reach and maintain target elicited uninformative

answers.

A recent study by Stanley, Mathews, Buss, and Koter-Cope (1989) used a more complete methodology for measuring verbal knowledge. Stanley et al., using the sugar production and Clegg tasks of Berry and Broadbent (1984), asked subjects after every 10 trial blocks to give complete instructions on how to perform the task. The informativeness of these instructions was assessed by the performance of yoked subjects requested to follow the transcribed instructions. Stanley et al. demonstrated that sudden improvements in performance by the original learners were not associated with simultaneous increases in the informativeness of the instructions. In fact, instructions helped the performance of yoked subjects only if the instructions were taken at least four blocks after the improvement in performance. These promising results need confirmation with different knowledge elicitation techniques. A problem with using yoked subjects is that they may not apply the provided instructions systematically. In fact, Stanley et al. found that when subjects were given a rule that if consistently applied would lead to 83% performance on average, they only performed at 60%. A more reliable measure of the validity of instructions may be provided by a computer simulated performance based on the instructions. This technique was used by McGeorge and Burton (1989) with the sugar production task. After 90 trials, 11 out of 28 subjects achieved simulated performances equal to or greater than actual performances. Unfortunately, they did not determine the time course of simulated performance knowledge; the point at which it develops relative to actual performance is important in assessing the claim for implicit knowledge.

In summary, work with the control tasks provides good

evidence for distinct knowledge bases and learning modes. With nonsalient relationships, subjects learn most readily in a way that allows them to control the task but not to form a context-free model of how the system works; an implicit mode of learning leads to implicit knowledge. With salient relationships, subjects learn to control the task at least partly by forming a model that enables them to predict future states of the system. In this case, an explicit mode of learning leads to explicit knowledge. However, the control tasks are relatively complex compared to most tasks in psychology; this may provide them with both the advantage of investigating how people learn in complex environments (Broadbent, 1977) and the disadvantage of working out exactly how subjects do it (for example, tests of predictive knowledge are only informative as to the specificity of performance knowledge if predictive knowledge is actually used "implicitly" during performance). Thus, the question arises whether the phenomena discussed above, indicating different implicit and explicit modes of learning, could be found with simpler tasks; such converging evidence would help rule out the possibility of task specific artifacts. Also, in the control tasks, what set of events defines the stimulus to which the subject responds on each trial is not clear (see Hayes, 1987, p. 88); a simpler task might allow greater precision in specifying the stimulus and therefore in defining what is learnt. A simpler task may allow greater control in investigating the properties of implicit learning. Accordingly, we now turn to consider concept formation tasks, where each trial involves a discrete stimulus and a discrete response.

Concept formation tasks

A concept formation task involves the subject attempting to classify exemplars. By analogy with implicit learning on the control tasks, implicit concept formation firstly implies that the subject can classify exemplars (performance) but the knowledge used for classification cannot be fully elicited by other measures or in different situations. This has typically been investigated with reference to the subject's ability to justify categorical choices (verbal knowledge). For example, after classification, the subject might not be able to say what attributes of an exemplar are relevant or in what way they are relevant. Secondly, for the knowledge underlying classification to be distinctively implicit, there should be evidence that such knowledge, as compared to more general and flexible knowledge (explicit knowledge), is of a fundamentally different type. Evidence for this second characteristic, unlike the first, requires demonstrating, or arguing, that the additional measure or measures tap explicit knowledge. The evidence for these two characteristics of implicit concept formation is discussed in turn.

Performance relies on a specific database.

In demonstrating the specificity of the performance database it is not enough simply to show that there is another measure that seems to be relevant to the task, but on which the subject does not perform well. This is due partly to the problem of simply not asking for the same knowledge that the subject used to classify exemplars (as in the case of correlated hypotheses; Dulany, 1961) and partly to the problem of differential test sensitivity (that is, a more thorough version of the same measure might elicit the same

knowledge as performance); see Ericsson and Simon (1984) for a thorough review of these issues with respect to verbal reports. Ideally, a battery of measures would be used to show the specificity of the implicit knowledge. This has rarely been done, but it is a start to show that a single plausible measure fails to tap the same degree of knowledge as performance.

Because of the wealth of studies reported on artificial grammar learning, this area is considered first in a section of its own. Then studies relevant to other concept formation paradigms are discussed.

(a) Artificial grammar learning.

Reber has long been exploring the learning of finite state grammars, which he claims occurs in some implicit way. In a typical study, subjects first memorize grammatical strings of letters, and then are informed of the existence of the complex set of rules that constrains letter order, and are asked to classify grammatical and nongrammatical strings. In an initial study, Reber (1967) found that the more strings subjects had attempted to memorize, the easier it was to memorize novel strings, indicating that they had learned to utilize the structure of the grammar. Subjects could not give a free account of the rules of the grammar (but when probed with specific letters, they could indicate which letters could start and which could finish a string). Reber has often asked for introspections regarding classification (e.g., Abrams & Reber, 1988; Allen & Reber, 1980; Reber & Allen, 1978; Reber & Lewis, 1977; Reber, Kassin, Lewis, & Cantor, 1980), but, unfortunately, he has rarely given a detailed analysis of the rules reported by subjects. Reber and Allen (1978) asked subjects to give justifications of their decisions during classification, but subjects did so on only

821 out of 2000 trials. With 50% guessing accuracy on 2000 - 821 = 1,179 trials, one might expect 821 + 589.5 = 1410.5 trials correct. In fact, there were 1,620 correct trials. Reber and Allen concluded that although subjects emerged with a small but solid body of articulated knowledge, they still could not tell all that they knew. Unfortunately, the appropriate analyses were not done to justify this conclusion. It appears that subjects knew quite a bit:

> Specific aspects of the letter strings were often cited as important in decision making . . . first and last letters, bigrams, the occasional trigrams and recursions were mentioned . . . Introspections after [observation learning] abound with references that have abstract rule-like qualities. Subjects refer to what can (and cannot) be . . . (p. 202).

There is no evidence that if these rules were applied on trials where no justification was reported, the predicted would not match the actual classification performance. The same applies to statements such as Abrams and Reber's (1988) that retrospective reports after classification " . . . yielded few data of value . . . few accurate characterizations of the rule system were given (p. 432)." Reber and Lewis (1977) also compared subjects rules and justifications with the subjects' classification performance. They cite several cases where a subject retrospectively claimed a letter in a certain position was acceptable, but the subject correctly rejected a nongrammatical item where the only violation was the letter in question. However, there is no evidence that the nongrammatical item did not violate some other of the subject's rules, and so there may be no inconsistency between stated rules and classification performance

A more systematic procedure to investigate the validity of the subjects' freely stated rules was used by Mathews, Buss,

Stanley, Blanchard-Fields, Cho, and Druhan (1989). Experimental subjects first briefly studied a set of exemplars, and then classified for 800 trials with feedback. After each 10-trial block, subjects were asked to give complete instructions on how to classify (the free recall measure of their knowledge). The validity of these instructions was assessed by the classification performance of yoked subjects requested to follow the transcribed instructions. The yoked subjects performed significantly (and substantially) above chance, and their performance gradually increased across blocks in the same manner as experimental subjects. This last result is consistent with the classification and free recall measures accessing the same (or a partly overlapping) knowledge base in experimental subjects. However, the yoked subjects always performed significantly worse than experimental subjects, suggesting that the knowledge base was not equally accessible by the classification and free recall measures. While these results are suggestive of a specific knowledge base underlying classification, there are some problems with using yoked subjects to assess the validity of the rules stated by experimental subjects. If the instructions contain exemplars, or parts of exemplars, implicit learning on the yoked subjects part may lead to an overestimation of the explicit rule content of the instructions. Conversely, application errors by the yoked subjects may lead to an underestimation of the validity of the instructions. As mentioned on page 17, Stanley et al. (1989) showed that giving subjects a rule for performing the control tasks perfectly lead to only 60% performance: Subjects could not have systematically applied the rule. If the instructions of experimental subjects contained many rules, the problem would have been accentuated: Yoked subjects may have assigned priorities in an

idiosyncratic and inconsistent way. A more systematic way of assessing the validity of rules elicited in free recall would be to directly use them to simulate classification performance.

The knowledge underlying classification of grammatical and nongrammatical strings has been shown to transfer to a number of tasks, including anagram problems. Reber and Lewis (1977) argued that the anagram task could rely on implicit knowledge, and that the subjects' anagram performance could not be accounted for by the subjects' statements of the grammatical rules. Unfortunately, Reber and Lewis did not assess whether there was complete transfer of classification knowledge to anagram performance. Reber and Lewis initially presented subjects with 15 grammatical strings for memorization. Next, subjects were presented with 28 sets of shuffled letters to be rearranged into grammatical strings; on this anagram task, subjects were given no feedback. On the second, third, and fourth days, subjects were initially re-exposed to the 15 grammatical strings presented earlier, and performed the anagram task. On the fourth day, subjects classified grammatical and nongrammatical exemplars, and were encouraged to give reasons for their decisions. Finally, subjects wrote down what they thought the rules of the grammar were and any strategies they used in going about the tasks.

Reber and Lewis (1978; p. 353) indicated various discrepancies between an average subject's reported rules and his anagram performance. Thus, the subject cited XXV as a possible ending, but of the 48 occasions where it was possible to construct a sequence ending in XXV, he did so on only five. However, anagram performance improved over the four days (from 31% to 68% correct on average), indicating that the relevant knowledge changed with time;

this certainly complicates the comparison of anagram performance with stated rules at the end of four days (see Ericsson & Simon, 1984, p. 153). Thus, some rules realized late in the experiment may not be reflected in many of the subject's anagram solutions, but this would not provide evidence for separate databases for stating rules and solving anagrams. Only one example was given by Reber and Lewis of the subject saying a letter sequence could occur when it did not in his solutions and opportunities were presented, and no examples of the subject saying a sequence could not occur when it did in his solutions.

Another task to which the knowledge underlying classification of grammatical and nongrammatical strings does transfer was introduced by Dulany, Carlson, and Dewey (1984). They asked subjects during classification to score that part of a string that "made it right" if it was classified as grammatical, or that part that violated the rules if it was classified as nongrammatical. For each feature thus scored by each subject, a validity was calculated: P(String is in the correct category / Feature i is in the string). The mean validity for each subject predicted proportion of correct classifications with a slope not significantly different from 1.00 and an intercept not significantly different from 0.00; this provides strong evidence that the scoring and classification tasks tapped the same database with about the same sensitivity. Another similar task to which the classification knowledge appears to transfer, at least partially, is the filling in of one or two blank spaces of otherwise grammatical strings (McAndrews & Moscovitch, 1985).

Perruchet and Pacteau (1990) investigated the hypothesis that classification knowledge involved knowledge only of (1) first

and last letters, and (2) permissible digrams, but not their positional dependency. Perruchet and Pacteau found that subjects exposed only to digrams could classify as well as subjects exposed to complete exemplars. Further, when subjects exposed to complete exemplars were asked to rate isolated digrams for their legitimacy, their ratings could predict classification performance without significant error. Unfortunately, Perruchet and Pacteau did not use a control group to assess performance on this transfer task in the absence of exposure to grammatical exemplars. This may be a problem because the legitimacy of different digrams was not counterbalanced across subjects.

There is a further problem in concluding that subjects only had knowledge of permissible digrams independent of position. The classification task employed exemplars that could be discriminated by the digrams used independent of their position. Indeed, Perruchet and Pacteau (1990) found that when subjects were exposed to nongrammatical exemplars that contained only permissible digrams, they could classify significantly above chance. The conclusion that the classification knowledge base includes constraints on the positions of features is further supported by the results of Dulany et al. (1984): Their scoring task predicted classification performance without significant error only when the feature scored was taken to include position. To summarize: The knowledge of permissible digrams underlying classification performance appears to transfer to rating isolated digrams; the knowledge of the positional dependency of digrams may or may not transfer to rating incomplete strings.

Reber (1976) found that the memory advantage of prior experience with the strings (Reber, 1967) was maintained if the

letters were changed but the same grammar was used. Similarly, Mathews et al. (1989) found substantial transfer to different letter sets when the classification task was used (when feedback was given), although transfer was significantly less with the different rather than the same letters. Thus, identical visual information during acquisition and testing is not necessary to access the classification knowledge base. When no feedback was given, transfer to different letter sets was poor, but above chance. Thus, unsurprisingly, a major part of the transfer to different letter sets appears to depend on the subject being able to determine the mapping between old and new letters. It is therefore possible that strings are at least partly processed in terms of an acoustic or articulatory code. If both visual and acoustic/articulatory codes could be rendered inapplicable, more specific transfer effects might be found.

To summarize work with artificial grammar learning, the knowledge underlying classification performance transfers at least partially to anagram solving (Reber & Lewis, 1977), to rating incomplete strings (Perruchet & Pacteau, 1990), to filling in missing letters (McAndrews & Moscovitch, 1985), and to new strings made with different letters (Mathews et al., 1989); and this knowledge transfers completely to scoring the relevant features of complete strings (Dulany et al. 1984). A "conceptual task" could be suggested that underlies performance, to varying degrees, on all the above tasks: Recognizing the well-formedness of strings or parts of strings. Specificity may emerge if the recognition of parts of strings is complete only when they are embedded within whole strings. The claim has also been made by Reber (e.g., Reber & Allen, 1978; Reber, Allen, & Regan, 1985) that the knowledge does

not transfer to free justifications of classification choices, where the justification is not itself constituted by a recognition judgement. Mathews et al. (1989) investigated this issue using a yoked subjects design, and provided some support for the claim. However, a problem with using yoked subjects is that there is no control over how they use the free recall information. Future research could assess the validity of rules elicited in free recall by directly using them to simulate classification performance. Such research would have to take into account the greater sensitivity of recognition to recall (e.g., Tulving, 1983) to provide a strong test of the hypothesis of separate databases for classification and justification.

(b) Other concept formation paradigms.

Many of the studies reported in this section measure the transfer of classification knowledge only with a free recall test. A problem is that both classification knowledge and the free recall test may access the same knowledge base with the same d', but the classification task may force the application of low confidence knowledge that is below the subject's criterion for free report. Thus, failure of classification knowledge to transfer to free recall provides only weak support for the existence of a distinctive classification knowledge base. Failure of classification knowledge to transfer to a recognition measure provides stronger support.

Smoke (1932) was one of the first to examine whether ability to define a concept matched classification performance. After exposure to instances of geometrically defined concepts, subjects gave a definition to the concept and then sorted 16 exemplars. The key result was that, considering only cases where the sorts were perfect, 30% of the definitions were defective. Smoke concluded

that subjects may be unable to give accurate verbal formulations of learned concepts. However, given any independent variance in the efficiency with which each measure taps a common database, selecting for perfect performance on one measure would give less than perfect performance on the other. The appropriate analysis of this data would involve deriving a predicted classification performance from each definition and comparing this to actual performance. Unfortunately, this analysis has not been conducted, and, in fact, has not been conducted for many of the studies reviewed here. Heidbreder (1934,1936) also reported inadequate verbal formulations of concepts subjects could sort, though the details of this work are not available.

Brunswick and Herma (1951) asked subjects to hold a series of weights in each hand. In one experiment, one hand held the heavier weight two thirds of the time. On trials where the weights were equal, subjects judged weights in the hand that normally received the lighter weights as heavier. When subjects were asked on the completion of the experiment which hand tended to hold the heavier weight, most subjects answered incorrectly. Brunswick (1956) argued that the results showed "learning against awareness", but for our purposes the important point is that subjects' incorrect responses accounted for their performance on test trials; thus, no evidence is provided for implicit concept formation.

In a study by Rommetveit (1960a), subjects ranked according to personal preference descriptions of stimulus persons constructed to vary systematically along 3 dimensions (intelligence, looks, and honesty). Many subjects ranked along a single dimension only; for example, the subject may have preferred all honest persons. Most such subjects did not report adopting any systematic strategy at

all, and they had particularly poor memory of the stimulus items upon which their conceptual sorting had apparantly been based. It remains possible, however, that subjects could report "unsystematic" strategies that could account for their performance. Rommetveit (1960b) asked boys to play a wheel of fortune game. Subjects put money in a slot which resulted in a figure appearing. Some figures were "good" and resulted in money being released. Other figures were "bad" and gave no prize. After 620 presentations, subjects classified 12 figures as good or bad, and then gave their free account of good and bad characteristics. Out of 14 subjects, 6 who were sorting significantly above chance gave "vague" and "misleading" verbal accounts. Rommetveit and Kvale (1965b), using the same incidental conditions, asked subjects to verbalize the rule after 81 presentations, and found that all subjects who could classify could also state the rule. Thus, for this task, all subjects may have been able to state the rule when they first learned it (see the discussion of automatization on p.43).

In a study by Phelan (1965), subjects sorted plates into four categories with feedback, until they could do so once without error. They were then tested on a transfer task that relied on the same sorting principle. Sixty-nine out of 110 subjects sorted without error on the transfer task. Of these only 23 stated a rule that would lead to perfect performance on both the original and transfer tasks.

Wilson (1975) investigated the incompleteness of reported hypotheses in a concept learning task by looking at subjects' ability to re-sort exemplars according to their reported hypotheses. Each experimental subject initially sorted exemplars with feedback. She was then asked to write down her hypothesis of the

experimenter's rule and to immediately re-sort the cards according to that hypothesis. The written hypothesis was shown to a control group of subjects, and to the same experimental subject a week later; all subjects were instructed to sort according to the written hypothesis. The correlation between experimental and control subject sorts was greater when the experimental sorts occurred a week after learning rather than immediately after. That is, experimental subjects probably initially sorted according to information that was not fully contained in their reported hypotheses, and that was forgotten after a weeks delay.

Kellogg has found in a series of studies (Kellogg, 1982; Kellogg, Robbins, & Bourne, 1978, 1983; see also Schroth, 1987) that subjects often fail to recognize hypotheses stated on the immediately preceding trial, even if the hypothesis was successful. This suggests that the short term store may be cleared after response or that responses are not controlled by hypotheses maintained in short term store (e.g., Restle, 1962); but as the subject's response was always consistent with the hypothesis just stated by the subject, it seems likely that both response and hypothesis were generated from the same database.

Lewicki (1986) has produced a body of experimental work relevant to implicit learning. Several of the studies conducted by Lewicki may be regarded as concept formation studies (Lewicki, 1986; Experiments 4.7, 4.8, and 7.4; also, Lewicki, Hill, & Sasaki, 1989) and so are relevant here. In Experiments 4.7 and 4.8 of Lewicki (1986), subjects heard recorded self-descriptions in which the pitch level of the speaker was consistently related to the speaker's warmth or capability. Subjects' subsequent warmth or capability ratings of speakers were influenced by pitch, but subjects indicated

that pitch did not affect their ratings. However, Lewicki (in Experiment 4.8) argued that other aspects of the stimuli probably covaried with pitch and thus subjects may have been acting on the basis of correlated hypotheses. Additionally, subjects may have been acting explicitly on the basis of analogy to stored exemplars. Both of these possible sources of knowledge could have gone undetected by Lewicki's procedure. In Experiment 7.4 of Lewicki (1986), subjects were exposed to descriptions of women along with their photographs. One woman differed from the others both in hairstyle and in being described as good at maths. Subjects who could not recall the hairstyle of the women nonetheless were influenced by hairstyle in subsequent ratings of maths ability. This experiment pits a recall measure against a possible recognition measure, and so does not provide a strong test for separate databases.

Lewicki, Hill, and Sasaki (1989) presented subjects with artificial "brain scans" that purportedly came from intelligent or non-intelligent people. The brain scans from intelligent rather than non-intelligent people had a slightly greater proportion of a certain graphics character in them. Subjects could later classify brain scans significantly above chance, even though in a post-experimental interview they did not mention the graphics character or its relevance. The same results were replicated in a second experiment in which subjects were presented with "words" that purportedly came from a Polynesian language, and belonged to one of a number of categories (e.g., food, tool). The words from one of the categories was displayed on a slightly different location on the screen. Subjects could later classify words significantly above chance, even though in a post-experimental interview they did not

mention word position or its relevance.

McGeorge and Burton (in press) also found a failure of classification performance to transfer to later free recall. In three experiments, subjects were initially exposed to 30 four-digit numbers while performing an incidental cover task (for example, mental arithmetic). Each set of digits contained at least one "3". Subjects later classified as "old" slightly more numbers that contained one "3" rather than no "3". In a post-experimental interview, no subject reported the common feature of the learning set. When told what it was, all subjects reported that they had not been aware of it.

An unpublished study by Hayes (1987; Experiment 5) did attempt to show that the knowledge underlying classification could not be elicited by a recognition measure. Subjects classified "computer people" into one of three towns (the "Residents task"). Each computer person was described by four phrases, including one unique phrase. Each non-unique phrase occurred once in each town. With an average classification performance of 60%, subjects were informed of the presence of unique phrases, and asked to judge the uniqueness of each phrase. For phrases identified as unique, subjects were further asked to name the associated town (Hayes calls the number of correct responses to this question the "VPA" measure). Subjects recognized as unique and produced the correct town to only 38% of the unique phrases, significantly underpredicting actual classification performance. Although the study is to be commended for attempting to show a disparity between two systematic measures (classification and VPA), it is debatable if the measures were really testing for identical knowledge. The computer people were constructed so that all pairs and triplets of phrases associated

with a person were unique. Thus, a computer person could be correctly classified one hundred percent of the time without using the unique phrase if the subject remembered any pair or triplet of nonunique phrases and the associated town. Future research could usefully employ Hayes' paradigm, but also apply the VPA measure to combinations of phrases to provide stronger evidence for a distinctive database underlying classification.

There are a number of studies not designed to test for implicit concept formation but in which the authors have incidentally noted the incompleteness of subjects' justifications for classification decisions. Hanfmann (1941) reported that some subjects solved the Vygotsky block test according to just "what looked right". Walk (1952) found that subjects gave incomplete justifications, and also that reversing the attributes of exemplars and nonexemplars slowed learning even though subjects did not report noticing the reversal. Posner and Keele (1968) found that after learning a concept formation task based on random dot patterns, some subjects did not state any classification rules when asked. Elio and Anderson (1981, p. 413) reported that subjects appeared to use information that they did not later articulate, and Medin and Edelson (1988, p. 73) mentioned subjects' reports that the answer just "popped into awareness".

Finally, there are also a number of studies designed to test for implicit concept formation that have found subjects' justifications adequate to account for their classification performance (Eriksen, 1962; Carlson & Dulany, 1985; Mathews, Buss, Chinn, & Stanley, 1988; Schwartz, 1966). Eriksen tachistoscopically flashed either nonsense syllables or nonsense syllables with a letter replaced by a dash to subjects. Subjects were to report

whether they thought the brief stimulus was a pleasant or unpleasant word. Half the subjects were told "correct" when they said "pleasant" to an altered nonsense syllable, and the other half were told "correct" when they said "pleasant" to a standard nonsense syllable. One third of the subjects responded significantly above chance in the predicted direction, but all of these subjects could correctly indicate, either by recall or forced-choice, whether the "pleasant" or "unpleasant" stimuli were associated with the dash. In a further experiment, subjects were to judge which of two lines was longer. The lines were printed in different colours and designed so that when one particular colour occurred it was always with the longer line. After several hundred judgements, test trials were interspersed where the lines were equal. Only subjects who later indicated that they used the color cue showed any "illusory effect" on the test trials.

Carlson and Dulany (1985) presented subjects with letter strings falling into one of two ill-defined categories, with individual letters occurring in each category with probabilities ranging from .1 to .9. After each classification decision, subjects underlined the letter or letters that they thought indicated category membership. The mean validity of the underlined features (i.e. P(String is in the correct category / Feature i is in the string)) for each subject predicted the proportion of correct classification decisions without significant error. The slope of the regression line was close to 1.00 and the intercept to 0.00.

Mathews et al. (1988) investigated the learning curves of both classification performance (on the "Bouthilet task") and verbal protocols obtained after each ten trial block. On each trial, subjects were shown a keyword followed by 5 choice words, where the

words were English words or Chinese characters for different subjects. The subject selected one of the choice words, and was then told which was the correct one. The correct choice was the one that contained only letters (elements) that were in the keyword. With 100% correct feedback, the trial block in which performance began to rise above chance was not significantly different from the block in which the verbal protocols first indicated partial knowledge; and the block in which performance reached asymptote was not significantly different from the first block for which the protocols contained a statement of the complete rule (Experiment 2). With partially incorrect feedback, the development of verbal knowledge preceded the onset of correct performance (Experiment 3).

Mathews et al. (1988) had a resolution of only 10 trials. Schwartz (1966) asked subjects to state a hypothesis on each trial of a card sorting task and found that the intraclass correlation between predicted and actual number of correct placements averaged .94 over several experimental conditions.

To summarize work relevant to showing a specific database underlying classification performance, there have been several suggestive findings (e.g., Hayes, 1987; McGeorge & Burton, in press; Reber & Allen, 1978), but they have provided only weak tests of this characteristic of implicit concept formation. Hayes is commendable in attempting to compare recognition tests of both justification and classification, but he may have failed to test for identical knowledge with his two measures. Future research could usefully follow up on Hayes' paradigm, but extend the "VPA" measure to cover more of the knowledge that could have been used for classification.

In terms of the studies failing to show implicit concept formation, several interpretations are available. First, they may

have provided a problem unsuitable for implicit learning. For example, Mathews et al. (1988) suggested that their "Bouthilet" problems may have been inappropriate because they did not involve a family resemblance structure. The associations used by Eriksen (1962) may have been too salient and thus too easily detected by other learning mechanisms (Berry & Broadbent, 1988). Second, they may have presented the problem in a way that did not elicit implicit learning. For example, asking for a hypothesis on every trial (Schwartz, 1966) may inhibit implicit learning (Berry & Broadbent, 1988). Or third, they may have tested for subjects' justifications in an appropriately sensitive way, demonstrating the existence of the common database that had been there all along. In contrast to studies providing evidence for implicit learning, these studies have often used forced-choice (Carlson & Dulany, 1985; Eriksen, 1962) or concurrent verbalization (Mathews et al. 1988; Schwartz, 1966). Because of these ambiguities, further work could probably most productively focus on paradigms that already provide some evidence for implicit concept formation and increase the sensitivity of the transfer tests; for example, in the manner suggested for the Hayes (1987) experiment.

Because of the problems of assessing whether the transfer test was really as sensitive as classification performance, or really tested for the same knowledge, in this area there is no definitive experiment but rather an ongoing research programme. The claim of implicit concept formation on a particular task is strengthened when plausible transfer tests elicit inferior knowledge, and weakened when such tests elicit equivalent knowledge. Also, the pattern of transfer helps to define just what is the conceptual task that alone elicits the associated knowledge;

defining the conceptual task sharpens the predictions of which experimental tasks should elicit the knowledge and which should not.

Performance relies on a different type of knowledge.

Showing that implicit knowledge is of a distinct type (or learned in a distinct way) involves comparing it to a presumed measure of explicit knowledge (or a presumed explicit learning strategy). It can be assumed that knowledge the subject can freely state in the form of rules is explicit knowledge; presumably the subject could apply it to many different situations if he was so instructed. The status of knowledge elicited in more structured ways is uncertain: Is the scoring task of Dulany et al. (1984), or the VPA measure of Hayes (1987), explicit or not? This is an empirical question, and ideally, for our purposes, relevant research should also demonstrate that the knowledge elicited by the structured measure can be regarded as explicit; that is, as tapping a database that can be applied to a range of different tasks.

Three ways of showing that the knowledge underlying performance can be of a different type to explicit knowledge are considered. The first is to show that performance and explicit knowledge are affected by different variables. The second way is to show that inducing an explicit learning mode changes or interferes with performance. And the third way is to show that performance on a task assumed to rely on implicit knowledge is affected by different variables than one assumed to rely on explicit knowledge.

1. Performance and explicit knowledge affected by different variables.

A study by Hull (1920) has been quoted with respect to implicit concept formation for almost seven decades now. Hull (1920) presented his subjects with a list of Chinese characters with

which they were to associate nonsense syllables. A given nonsense syllable was associated with all characters that had a common element, though subjects were not informed of this. Experiment E used a condition in which the common elements replaced the characters (concept given outright), as well as the condition in which the elements were embedded within characters. The subjects attempted to produce the appropriate nonsense syllable in response to the elements or characters until they responded perfectly for two successive passes through the stimuli. Five subjects were then asked to draw the common elements, and the drawings were rated by judges for their "efficiency". Giving concepts outright as compared to embedding them in characters improved the rated efficiency of drawings (38% compared to 71%) but not the efficiency of classification (47% compared to 53%). Hull noted this interaction but did not test its significance. Calculation of the F for this interaction yields $F(1,4)=3.06$, ns.

Verplanck and Oskamp (Verplanck,1962) produced a dramatic dissociation between hypotheses and classification performance. Subjects sorted illustrated cards to the left or right, and stated on each trial what their hypothesis was. Subjects were reinforced either for making a correct placement or for stating a correct hypothesis. When placements were reinforced, subjects placed the cards correctly on 71.8% of the trials and stated a correct or correlated hypothesis on only 48.4%. When hypotheses were reinforced, subjects stated a correct or correlated hypothesis on 94.2% of the trials, but placed the cards correctly on only 76.8%. Dulany and O'Connell (1963) pointed out two artifacts that jointly accounted for the results. When placements were reinforced, subjects produced many uncorrelated hypotheses and a correction for

guessing should have been applied. This in fact yielded a predicted performance of 67.8%, not significantly different from the 71.8% obtained. When hypotheses were reinforced, most hypotheses were correct or correlated and a correction for stimulus ambiguity should have been applied. Control subjects asked to sort according to the experimental hypothesis misclassified on 14.2% of the trials. The experimental subjects predicted performance should therefore be 80%, not significantly different from the 76.8% obtained.

Using the Residents task mentioned earlier (see p.34), Hayes (1987) found a number of manipulations that increased the VPA measure more than performance. These included using fewer irrelevant phrases, perceptually blocking the irrelevant phrases, informing subjects prior to classification of the existence of unique phrases, and giving complete rather than incomplete feedback. Unfortunately, the interpretation of these effects is clouded by the incompleteness of the VPA measure: The manipulations could plausibly have induced subjects to use unique phrases rather than pairs or triplets, so that the knowledge content (rather than type) assessed by performance and VPA simply overlapped more.

2. Inducing an explicit mode.

Phelan (1965) asked subjects to state the sorting principle either after successfully completing one sorting task and before starting a transfer task that relied on the same sorting principle, or else after finishing the transfer task. In the first case, 28 out of 55 subjects transferred perfectly, in the second case 41 out of 55. A chi-square on these numbers is significant (X^2=6.57, df=1, $p<.05$). That is, being asked to state the rule interfered with applying it. Rommetveit and Kvale (1965a) asked boys to play the "wheel of fortune game" described earlier (p.29), where some figures

indicated reward, and others no reward. Half the subjects were informed that they would later be asked to report the difference between "good" and "bad" figures. This instruction deteriorated subjects' ability to sort verbal descriptions of the basis of good and bad figures (but not the ability to sort actual good and bad figures).

Reber (1976) instructed one group of subjects to simply memorize a set of grammatical strings (implicit instructions), and another group to search for rules to assist their memorization (explicit instructions). The implicitly rather than explicitly instructed subjects subsequently classified more test strings correctly. However, data provided by Experiment One of Reber, Kassin, Lewis, and Cantor (1980) indicated no difference between implicit versus explicit instructions when the stimuli were presented in a random order, as in Reber (1976), $F(1,12)=1.90$, $p>.1$ (the F can be calculated from the information presented in the paper), and an advantage of explicit over implicit instructions when the stimuli were presented in an order that highlighted the grammatical rules (cf. Berry & Broadbent, 1988). Further, Dulany et al. (1984), Hayes (1987; and as reported in Broadbent, 1989), and Mathews et al. (1989; Experiments 1 and 2). all following a similar procedure to Reber (1976), failed to find an effect of implicit versus explicit instructions on randomly presented stimuli. Mathews et al. (1989; Experiment 3) used a much stronger implicit/explicit manipulation and still found no difference: For the implicit task (the match task), subjects held a string in memory for a few seconds and then had to recognize it among one of several choices. For the explicit task (the edit task), subjects were informed of the presence of rules and were asked to underline possible incorrect

letters in invalid strings. After each invalid string the correct string was presented. Classification performance was equally good after the match or edit task alone or in combination. Thus, a more fine-grained measure of performance than overall accuracy is needed to distinguish different possible implicit and explicit learning modes in the artificial grammar learning task.

Such a fine-grained analysis could be provided by model fitting. For example, Nosofsky, Clark, and Shin (1989) fitted an additive version of Medin and Schaffer's (1978) exemplar model and a rule based model to the performance of subjects classifying perceptual stimuli. The performance of subjects explicitly asked to search for rules was fitted better by a rule based rather than exemplar model; the performance of control subjects, not so instructed, was fitted better by the exemplar rather than rule based model.

3. Implicit and explicit tasks.

This section reviews the final sort of evidence for showing that the knowledge underlying performance can be of a different type to explicit knowledge: Comparing performance on tasks assumed to rely on implicit and explicit knowledge, respectively. The relevant studies here involved the artificial grammar learning paradigm. Hayes (1987; and as reported in Broadbent, 1989) instructed subjects either to simply memorize grammatical strings (implicit instructions), or to search for rules to assist their memorization (explicit instructions). Asking subjects to generate random numbers at the same time as they were exposed to the grammatical strings deteriorated subsequent classification performance for the explicitly but not implicitly instructed group (cf. Hayes & Broadbent, 1988). One problem with this study is that subjects were

not given priority instructions, and how implicitly and explicitly instructed subjects assigned priorities may have varied as a function of different task demands. Unfortunately, performance on the random number generation task was not measured to assess the extent of this possible artifact.

Abrams and Reber (1988) found that psychiatric patients classified grammatical and nongrammatical strings similarly to normals after exposure to grammatical strings, but were inferior to normals on a task that required determining a mapping between letters and numbers, and that was regarded as requiring explicit knowledge. Similarly, Mathews et al. (1989; Experiment 3) compared artificial grammar learning with learning a "biconditional" rule that determined the mapping between corresponding letters in the first and second halves of a string. As described above on page 40, with the artificial grammar, the match and edit tasks produced identical classification performance. With the biconditional rule, on the other hand, classification performance was better after the edit rather than the match task; performance on the match task was next to chance. Mathews et al. hypothesized that only the finite state grammar involved a family resemblance structure that could be learned implicitly.

The converging evidence presented in these three sections implies the existence of at least two modes of concept learning, only one of which involves the formulation of freely expressible rules. However, whether the other mode corresponds to implicit learning depends on the specificity of the resulting knowledge, and so the uncertainties highlighted in the last section with respect to the specificity carry over to the interpretation of the results in this section. The evidence presented here is suggestive and

motivates the need for further research in the manner outlined at the end of the review of evidence for the first characteristic of implicit learning. Specifically, the Hayes' (1987) Resident's task could be extended to include a more complete measure of transfer knowledge. If the specificity of transfer is maintained, then the distinct properties of implicit learning could be more unambiguously investigated.

Theoretical issues

This section focusses on the manner in which implicit learning and implicit knowledge are best conceptualized. This thesis has proceeded by defining implicit knowledge as a distinct "non-manipulable" type of knowledge that can be indexed by two characteristics. The first characteristic is that implicit knowledge can be elicited only by specific tasks. The second characteristic is that implicit rather than explicit knowledge has different properties of storage and retrieval. These two characteristics are discussed in turn.

This thesis emphasizes the specificity of transfer of implicit knowledge. Of course, people fail to apply much of their explicit knowledge in situations where it would be useful (for a set of recent reviews on learning transfer, see Cormier & Hagman, 1987). What is taken to be peculiar to implicit knowledge is that transfer fails even when the subject is currently aware of the exact mapping between corresponding elements of the original and transfer tasks. In implicit learning paradigms, the transfer task has often been a logical or likely component of the learning task (e.g., the VPA transfer measure of the Resident's task, Hayes, 1987; the predictive knowledge questionaires of the control tasks, Berry & Broadbent,

1984), and so the corresponding elements of the two tasks are identical. Sometimes the transfer task has been simply a different surface embodiment of the same underlying structure (Berry & Broadbent, 1988; Berry, in preparation), and subjects may be informed of the mapping between corresponding task elements (Berry & Broadbent, 1988) or the mapping may be a previously overlearned one (e.g, typing versus speaking, Berry, in preparation).

Evidence of transfer specificity may be regarded as lying along a continuum. On the one end, there is failure to transfer to free recall. While it is interesting if free recall fails to elicit classification knowledge, such failure is not strong evidence for a distinct "non-manipulable" type of knowledge. There are two reasons why free recall might fail to elicit knowledge that we would not want to regard as "non-manipulable" or as constituting a distinct data base. First, there may be a lot of knowledge to be retrieved, and knowledge not free recalled in a given period of time may be recalled when the subject is given another attempt (Erdelyi & Becker, 1974). Second, in free recall (but not a forced choice task like classification performance) the subject has the option of not responding, and low confidence knowledge may not be elicited. Failure to elicit knowledge by free recall would provide evidence of "implicit" learning at an everyday level of explanation but only weak evidence for a distinct knowledge base. Towards the other end of the continuum of evidence of transfer specificity, there is failure to transfer under increasingly rigorous conditions, using forced choice measures. This provides stronger evidence for a distinct knowledge type.

This thesis will not attempt a precise operational definition of implicit knowledge, but it is to be expected that any

future definition of implicit knowledge will include some notion of how the knowledge is difficult to elicit - that is, of specificity of transfer of the knowledge. Such a definition may state exactly which type of task will elicit the knowledge and which will not. Unfortunately, our empirical knowledge is not yet at an adequate level to allow such a definition to be formulated. The most useful strategy at the present time may be to simply explore the specificity of transfer of knowledge acquired in different domains. The pattern of transfer in domains which produce knowledge that is very difficult to elicit may suggest a more useful and complete definition of implicit knowledge.

This first characteristic of implicit learning - the specificity of transfer - captures part of what is meant by calling implicit knowledge "unconscious" (e.g., Hayes, 1987; Lewicki, 1986; Reber, 1988), but leaves the researcher with a clear research programme: Can the pattern of transfer define a conceptual task that elicits the information with a high degree of specificity? On the other hand, defining implicit knowledge as unconscious places the burden on the researcher of showing that no conscious measure elicits the knowledge. Whereas "conscious measure" may have some clear referents, it does not with respect to structured measures of knowledge (compare Dulany, Carlson, and Dewey, 1985, and Reber et al., 1985). and, in these cases, there may be no noumenal fact of the matter. In many cases, consciousness may have to be legislated rather than fought over (see e.g. Allport, 1989; Putnam, 1981).

The second characteristic is that implicit as compared to explicit knowledge should be of a fundamentally different type. Evidence suggesting different principles of storage and retrieval would be evidence for different knowledge types.

This characteristic also appears to capture in a neutral way part of what has been meant by implicit learning (e.g., Berry & Broadbent, 1988; Hayes, 1987; Lewicki, 1986; Reber, 1988). Previous research has often taken inspiration from our notions of "unconscious" to investigate in what way implicit knowledge is of a different type; for example, that its development is unaffected by secondary tasks (Hayes & Broadbent, 1988), psychiatric impairment (Abrams & Reber, 1988), or attentional manipulations (Kellogg, 1982; Lewicki, 1986). Again, this second characteristic is not dependent on having a clear theory of the unconscious; the only hope is that implicit will be different from explicit knowledge in some interesting way. Thus, the two characteristics outlined here capture much of what is meant by implicit knowledge in the literature, while leaving open how it should best be theoretically addressed. It is to this issue that we now turn.

Three constructs with which implicit knowledge could be compared are procedural versus declarative knowledge (Anderson, 1983; Cohen, 1984; Ryle, 1949), automatization (e.g., Shiffrin & Schnieder, 1977), and modularity (see Marr, 1982; or, for a more extreme view, Fodor, 1984). Procedural knowledge refers to "knowing how" and declarative knowledge to "knowing that" (Ryle, 1949). The distinction has been embodied in a number of psychological theories; for example, in J. R. Anderson's (1983) ACT* theory, procedural knowledge is represented in the form of productions and declarative knowledge in the form of a semantic network (see also Cohen, 1984; Tulving, 1985). One way of relating implicit knowledge to the procedural-declarative distinction is to say that a task is implicitly acquired if procedural knowledge developed in the absence of, or prior to, declarative knowledge. The problem with this

approach is that the procedural-declarative distinction is insufficiently empircally specifiable to add anything to our existing definition of implicit knowledge. All actions by a subject on any task must rely on procedural knowledge. And any action of a subject involving making a symbolic choice can be interpreted as asserting a proposition. So any measure on any task that involves the content (rather than speed or quality) of a subject's action will simultaneously measure procedural and declarative knowledge. Declarative knowledge could be taken to mean some other measure that could be used to account for performance. But the issue then becomes identically the issue of specificity of transfer; defining the conceptual task that the subject can perform and showing that it does not transfer to other measures.

Consider, for example, a task used by Willingham, Nissen, and Bullimer (1989; for similar results with a similar task, see Stadler, 1989). In a sequential reaction time (SRT) task, subjects pressed one of four buttons immediately below one of four lights, depending on which light was on. When a 10-trial repeating sequence was used, reaction times decreased relative to a random sequence condition, and this was taken as a measure of procedural knowledge. In a subsequent classification task, subjects had to indicate which light would come on at each point in the sequence. This classification task was taken to be a measure of explicit declarative knowledge, although there is no a priori reason for doing so, as classification can also be taken as a measure of procedural knowledge (e.g., Reber & Allen, 1978, p. 203). The point is that some subjects can be good on the SRT task but not on the classification task, so that there is specificity of transfer. That specificity of transfer is important is also shown by a lack of

transfer for these subjects between different versions of the SRT task when surface perceptual or motor characteristics of the task were altered. Subjects good on the classification task, on the other hand, did show transfer on the SRT task when its surface characteristics were altered. Thus, the specificity of transfer, rather than the procedural-declarative distinction per se, helps to define two different types of knowledge bases, the implicit and the explicit.

Implicit learning is in some respects similar to automatization. Automatization is a process by which the repeated application of flexible procedures to a given situation results in the procedures being elicited in an efficient but inflexible way whenever the situation is presented (see e.g., Schneider, Dumais, & Shiffrin, 1984). J. R. Anderson (1983) has described automatization in terms of the ACT* production system. According to ACT*, all knowledge originally comes in declarative form, and is interpreted by general flexible procedures. Through practice on a task, knowledge compilation (automatization) leads to domain specific productions that may fire off without reference to declarative knowledge. A key distinction between automatization and implicit learning is that implicit learning appears to result in inflexible (i.e., implicit) knowledge predominantly at the early stages of the learning curve (see e.g., Hayes, 1987; Lewicki, 1986), and not at the point of overlearning. Thus, most studies arguing for implicit learning, have done so when the learning is still very incomplete - this is true of the Broadbent, Lewicki, and Reber studies. Further, both Stanley et al. (1989) and Sanderson (1989) found that performance knowledge on the control tasks did not transfer to verbal knowledge early in practice, but did after extended practice.

This is the opposite pattern of results expected on the basis of automatization.

Fodor's (1983) concept of modular knowledge is also in some ways similar to that of implicit knowledge, in that a module has its own proprietary database. However, the emphasis in the case of modules is that information available to other cognitive processes is often not available to the module, whereas in the case of implicit knowledge the emphasis has been that the implicit knowledge is not available to other cognitive processes; though these may be different sides of the same coin. Other properties of modules, such as their genetic basis, are in general not shared by implicit knowledge, which is, of course, learned; but Fodor (1985, p. 39) has suggested that "Mother Nature, having tried peripheral modular mechanisms and found them good, [may have] then contrived, via the novice-expert shift, to simulate some of the effects of modularity at the level of central systems." Fodor had automatization in mind, but implicit learning could also do the job.

One key point of debate in the implicit learing literature has been whether implicit knowledge is best represented in an abstract way (e.g., Reber, 1989; Reber & Allen, 1978) or in terms of the storage and deployment of exemplars (e.g., Brooks, 1978; Ericsson & Simon, 1984, p. 114). Brooks (1978) generated strings from two different grammars and used them in a paired associate learning paradigm in which strings were paired with English words. The strings of the two grammars could be distinguished in a nonobvious way: Whether they were paired with a word describing a New World or Old World entity. Subjects subsequently informed of the distinction could classify nongrammatical strings and grammatical strings of the two grammars; subjects not informed could

not. The results provide evidence that a strategy based on analogy to stored exemplars can lead to above chance categorization. Of course, a subject could follow such a strategy without knowing the rules of the grammars. Reber and Allen (1978) presented subjects either with a paired associate learning task (each exemplar was paired with an English word) or a task that simply involved observing exemplars. They observed several differences in the way subjects subsequently categorized grammatical and nongrammatical exemplars and argued that subjects can use analogy to stored exemplars but there still remains some other mechanism that implicitly abstracts the rules of the grammar. The results do suggest two different strategies, but do not rule out an exemplar model of both. When subjects are induced (say, by paired associate learning) to have a "fairly concrete memorial space consisting of specific items and parts of items (Reber & Allen, 1978, p. 2)" that can be recalled, subjects may be able to classify on the basis of analogy in such a way that the main stages of the process, and the associated information, are available in a Think Aloud protocol. Such knowledge may be transferable to a range of superficially dissimilar tasks (especially if the subject is told of a correspondence between elements of the stored exemplars and elements of a new task) and thus could be regarded as explicit (but nonanalytic). On the other hand, even if the exemplars could not be distinctly recalled, they may still be stored and used to classify. For example, according to the models of Estes (1986), Hintzman (1986), and Medin and Schaffer (1978), information from the stored exemplars combines to give a sense of familiarity, a recognition judgement, or a classification judgement, without necessarily allowing the subject to report on intermediate stages or on which

exemplars were used. If the exemplars were perceptually encoded, then the knowledge may not transfer to perceptually dissimilar stimuli, and may be regarded as implicit.

McAndrews and Moscovitch (1985) sought to determine whether rule-based or exemplar-based information was a more important determinant of classification performance in artificial grammar learning. Grammaticality and similarity to studied exemplars were manipulated independently. Similarity was measured (inversely) by the smallest number of differences in letter positions to any studied exemplar. Grammaticality and similarity were found to account for a roughly equal amount of variance in classification perfomance. McAndrews and Moscovitch (1985) concluded that there was evidence for the abstraction of rule-based information. However, while the manipulations McAndrews and Moscovitch (1985) made are plausible in the absence of any particular exemplar model, it remains quite possible that an exemplar model could account for the effects of both similarity and grammaticality. The grammaticality effect may arise because each letter position of a grammatical item is likely to be the same as the letter position of a large subset of stored exemplars (though nonidentical subsets for different letter positions), though some letter positions of nongrammatical items may be different to the letter positions of any stored exemplars. The issue is best resolved by actually running simulations of different models; whether exemplar models like those of Estes (1986) or Hintzman (1986) can account for the pattern of classification performance under simple observational learning conditions is an open question.

Mathews et al. (1989) argued that according to typical exemplar models initial variability in the knowledge representations

across subjects would be due to the variability in the goodness with which exemplars are encoded. After extensive experience with the same exemplars, the knowledge representations of subjects should become very similar. Mathews et al. measured the convergence in subjects representations according to their similarity in types of trigrams mentioned in the subjects' free reports of the rules that they used in classification. After every 10 trials, subjects gave free reports for a total of 800 trials. Mathews et al. report that the similarity across subjects was low and did not increase with practice, in apparant contradiction to exemplar models. However, if free report of the rules is based on a separate knowledge base to classification performance, or a biased and selective access to the same knowledge base (see page 22), it is far from clear that the lack of convergence found in free recall applies to classification performance knowledge. In fact, Mathews et al. report that the types of trigrams appearing in recalled exemplars are significantly different to those appearing in the free report of the rules, implying that if different measures access the same knowledge base, they do so in selective ways. Further, as free report for rules was repeatedly assessed, subjects may to some extent access previous verbalizations in preference to the classification knowledge base (see e.g. Bekerian & Bowers, 1984; Broadbent & Broadbent, 1977), accentuating idiosyncracies between subjects. In sum, exemplar models are still highly plausible candidates for explaining classification performance in artificial grammar learning.

Broadbent et al. (1986) also considered a "situation-matching" strategy as a possible explanation for performance on the control tasks. In a similar analysis to that of

Brooks (1978), they suggested that a subject could construct a "look-up table" which would determine the appropriate action by matching the current situation to the most similar of the entries already in the table. This suggestion was formally instantiated in a model by Cleeremans (reported in Marescaux, Luc, & Karnas, 1989[1]). The model built up a look-up table relating situations to responses (levels of controlling variables). If the current situation was matched in the table, then the associated response was made; otherwise the response was random. If a response lead to the target output, then the response was entered in the table. Marescaux et al. attempted to test several predictions of the model.

The first prediction was that the knowledge should transfer to a questionaire that excluded the dynamic aspect of the task. This prediction was confirmed; subjects could answer questions requesting responses to particular situations (in order to achieve a target they had learned to reach). The second prediction was knowledge obtained with one target should not transfer to another; this prediction was supported with both performance and questionaire measures (for previous consistent results, see Berry & Broadbent, 1987; McGeorge & Burton, 1989, Experiment 1; for contrary results, see McGeorge & Burton, 1989, Experiment 2; Sanderson, 1989; the reasons for this discrepancy are not clear, but may relate to the amount of explicit knowledge available to the subject). The third prediction was that there should be consistency of response to the same situation. Subjects were given situations that they had encountered in the task and in which the response chosen had been followed by the target. The response chosen by subjects on these

[1]The information presented here is based on an English outline of the paper provided by Donald Broadbent.

questions matched their previous response only 50% of the time; the model would predict 100% concordance. The fourth prediction was that initially subjects should respond at random; this prediction was not confirmed.

Further results by Stanley et al. (1989) also suggest the importance of a look-up table; they found that having subjects memorize correct solutions to specific situations improved later control task performance (as did a number of other manipulations, discussed below). The findings of Marescaux et al. (1989) and of Stanley et al. are encouraging and suggest the feasibility of developing further models based on simple condition-action associations. Barto (e.g., Barto & Sutton, 1981; Barto, Anderson, & Sutton, 1982) has suggested a Connectionist learning rule for creating a look-up table. Such a model of the control systems would probably produce similar predictions to Cleeremans'. Because of the gradual "tuning" of Connectionist models, a Connectionist model may be able to predict the less than 100% concordance that was obtained by Marescaux et al..

Berry and Broadbent (1988) encompassed the notion of a look-up table within a view of learning that included two main modes: Unselective (U mode) and selective (S mode) learning. In U mode, the person observes the task variables unselectively and attempts to store all the contingencies between them in the form of condition-action links (implicit knowledge). Berry and Broadbent suggested that "It may well be hard to report so many links (p. 253)", but do not elaborate on why this should be. One explanation could be in terms of encoding specificity: If a lot of context is encoded, a lot may have to be reinstated to elicit the response. In S mode only a few variables are selected and only the contingencies

between these key variables are observed (explicit knowledge). Hayes (1987) further proposed that S mode, unlike U mode, relied on abstract working memory (Broadbent, 1984), and that U mode relied on the automatic registration of cooccurrence frequency. However, the key feature of implicit knowledge, its specificity of transfer, is left unexplained by such a frequency acount. Why, for example, should the frequency knowledge be applied to classification but not to Hayes' VPA measure? Again, one explanation may be in terms of the number of variables associated with the response which need to be reinstated to elicit the response. A Connectionist approach might provide one way of modelling the links between a number of variables such that specificity of transfer emerges. Rumelhart and McClelland (1986) suggested that lawful behaviour and judgements may be produced by a network in which rules are not explicitly represented but emerge from the way that interacting units are connected; thus, subjects behaving according to such a network might find it hard to justify their decisions according to stated rules, or to respond appropriately to elements of the problem presented in isolation. The relevance of using models that attempt to simultaneously satisfy a number of constraints is suggested by results of Stanley et al. (1989). They found that control task performance was improved to an equal extent by a number of different types of training; namely, memory training on specific condition-action exemplars; training on a single rule guaranteed to give perfect performance; training based on instructions from an expert subject; and training based on a number of guidelines and hints. As suggested by Stanley et al., equal improvement by different methods may come about because, despite the training, subjects do not rely on a single rule or type of constraint;

instead, learning consists of tuning and balancing a number of constraints.

Connectionist modelling would also provide one way of addressing another issue: Hayes (1987) suggested that implicit learning involved the registration of cooccurrence frequency but not contingency (which involves the frequency of one feature occurring in the absence of another). This distinction could easily be incorporated into a connectionist model by the way that the absence of a feature is coded.

This discussion has highlighted several theoretical issues that could be addressed by computational modelling. Specifically, whether exemplar models (for example, of Estes, 1986, and Hintzman, 1986) can account for artificial grammar learning; whether a Connectionist look-up table can account for performance on the control tasks; and whether implicit learning is sensitive to cooccurrence or contingency.

Logic and Aims of Thesis

This thesis develops both an experimental programme to address the issues raised by the review of the empirical literature, and a programme of computational modelling to address the issues raised by the discussion of theoretical approaches to implicit learning.

The review of existing work on implicit concept formation highlighted the need for further research to test the specificity of the knowledge underlying performance, and to determine its distinctive characteristics. Accordingly, the first major aim of this thesis was to develop concept formation tasks to address this need. The Residents task of Hayes (1987a) uniquely provided a case

where one structured measure (the VPA measure) underpredicted performance. The VPA measure did not test for all possible knowledge, but could easily be extended. Thus, the experimental part of this thesis starts by employing the Residents task and an extended VPA measure; successive experiments use a modified paradigm to test the specificity of performance knowledge. The distinctive characteristics of the knowledge underlying performance are investigated by comparing the effect of a secondary task manipulation on several knowledge measures, including performance.

The discussion of theoretical issues highlighted the need for a computational modelling approach to implicit learning. Exemplar and connectionist models seemed particularly promising; and several types of both sorts of model can be constructed. The modelling focusses on the artificial grammar learning and control tasks because a wealth of existing data already exists for how subjects perform these tasks. Thus, the second major aim of this thesis was to determine what class of models could learn the grammar and control tasks, and to provide competitive support for one class of model in accounting for how subjects perform these tasks.

Chapter Two

Experiments with the residents task

Introduction

The review of previous concept formation studies presented in Chapter One highlighted the current paucity of evidence for implicit concept formation. According to the logic used in Chapter One, the knowledge underlying classification performance is implicit only if it can be regarded as constituting a distinct database. Failure of classification performance to transfer to another test that putatively tests for the same knowledge provides evidence for implicit concept formation, but it is only strong evidence if the transfer and classification tasks are of similar sensitivities. Some previous studies found retrospective free recall of the knowledge underlying classification insufficient to account for classification performance (e.g., Rommetveit, 1960; Wilson, 1975), but free recall, especially in a limited time period, is known to be an insensitive measure of the subject's knowledge (e.g., Erdelyi & Becker, 1974; Tulving, 1984). Only one study, by Hayes (1987), found a failure of classification performance to transfer to another structured knowledge measure. Initially, because of the importance of this finding, the set of results found by Hayes is described. Then two experiments conducted for this thesis are reported that attempt to confirm and extend the results of Hayes.

In the Residents Task used by Hayes (1987), subjects classified nine "computer people" as living in one of three towns. Each computer person was described by a list of phrases, and for each person there was one unique phrase that distinguished it from all the others. In the one experiment that showed the failure of

transfer (Hayes, 1987; Experiment 5, Group U), each person was described by three irrelevant phrases as well as the unique phrase. To minimize the salience of the unique phrases, the combinations of shared phrases were maximally dissimilar across computer people. On each trial, subjects were told if they were right or wrong. Subjects classified to a criterion of two thirds of the computer people correct two thirds of the time, and they were then given a transfer task which Hayes' (1987) called the "VPA" measure (for Verbal Production of Associates). For this task, subjects were asked to recognize the unique phrases, and then produce the associated town for each phrase considered unique. Hayes (1987) found that the VPA measure was significantly less than the classification performance on the last block.

Hayes (1987) manipulated four variables that he speculated would influence implicit learning: Salience, level of performance, type of instructions, and type of feedback. VPA and classification performance were equivalent (1) if the salience of the relevant information was increased by using a single irrelevant phrase for each computer person (Hayes, 1987; Experiment 5, Group S), or grouping three irrelevant phrases into blocks that recurred in different computer people (Hayes, 1987; Experiment 7); (2) if subjects classified until they reached a high level of performance (one perfect pass through all nine computer people; Hayes, 1987; Experiment Six); (3) if subjects were given an analytical instructional set by being informed of the existence of the unique phrases (Hayes, 1987; Experiment 8); or (4) if subjects were given complete feedback by being told what the correct town was on each trial (Hayes, 1987; Experiment 9).

The manipulations Hayes (1987) made may be highly

informative of the nature of implicit concept formation. However, Hayes' interpretation of the results depends on only one case where classification performance did not appear to transfer to the VPA measure (Hayes, 1987; Experiment 5, Group U). It is therefore important to determine the robustness of this result. Also, as indicated in Chapter One, the transfer test used by Hayes fails to tap a highly plausible source of knowledge that could be used for classification. In (Hayes, 1987; Experiment 5, Group U), the combination of shared phrases was unique for each computer person. If subjects had remembered pairs or triplets of shared phrases, they could have classified perfectly without knowledge of the unique phrases. In fact, when this source of information was not available to subjects because the combinations were not unique (Hayes, 1987; Experiment 7), knowledge of the unique phrases, as assessed by the VPA measure, was sufficient to account for performance. It is therefore important to assess transfer knowledge of the combinations of shared phrases under the conditions of Hayes (1987; Experiment 5, Group U). Experiment One addressed these issues.

EXPERIMENT ONE:

The Residents Task

Introduction

The aim of Experiment One was to investigate the finding of implicit learning in the Residents Task (Hayes, 1987) using a more complete transfer task than had been used previously. Experiment One replicated the procedure of Hayes (1987; Experiment 5, Group U) and also attempted to test knowledge of the unique combinations of shared sentences, using a transfer test that followed as closely as possible the test of the unique sentences used by Hayes (1987). For the transfer test of unique combinations, subjects were asked to recognize combinations of shared sentences as unique or not, and to indicate the associated town for those combinations recognized as unique. Only triplets were tested to keep transfer testing short.

Method

Design. There was one between subjects factor, Test Order. Subjects received the test of knowledge of the unique sentences either before or after the test of knowledge of triplets.

Subjects. The subjects were 12 paid volunteers aged between 18 and 35 from the Oxford University subject panel.

Materials and Apparatus. The materials and apparatus were the same as those used by Hayes (1987; Experiment 5, Group U). Specifically, the stimuli were 9 computer people described by four sentences. One sentence was unique to a particular person, and each of the other sentences was shared by two other people who resided in each of the other two towns (see Table 1). The stimuli were presented on a colour monitor by a Sinclair ZX spectrum.

Table 1. Residents Task.

Town 1	Town 2	Town 3
1. U1 S1 S4 S7	4. U4 S1 S5 S6	7. U7 S1 S2 S9
2. U2 S2 S5 S8	5. U5 S2 S3 S7	8. U8 S3 S4 S5
3. U3 S3 S6 S9	6. U6 S4 S8 S9	9. U9 S6 S7 S8

Unique sentences:	Shared sentences:
U1 = LIVES IN THE NORTH	S1 = COLLECTS STAMPS
U2 = HAS A SISTER	S2 = ENJOYS FISHING
U3 = HAS A WELSH MOTHER	S3 = IS RIGHT HANDED
U4 = PLAYS THE GUITAR	S4 = ENJOYS READING
U5 = OWNS A GREEN CAR	S5 = HAS BLUE EYES
U6 = LIKES CHINESE FOOD	S6 = LISTENS TO RADIO 2
U7 = WEARS A RING	S7 = HAS SHORT HAIR
U8 = BANKS AT NATWEST	S8 = OWNS A BICYCLE
U9 = GOES TO CHURCH	S9 = PLAYS SPORT

A trial consisted of a single presentation of a computer person followed by a response from the subject, and a block consisted of a pass through all nine computer people. For each block, the order of computer people was randomly changed, and so was the order of sentences for each computer person. The computer stopped when subjects classified at least six computer people correctly at least twice over the previous three blocks, irrespective of the performance on any one block. The number of blocks to reach this criterion, the number of people correctly classified on each block, and the reaction time to respond to each person on the last block were automatically printed by the spectrum.

For the recognition test of the unique sentences, the nine unique sentences and the nine shared sentences were typed on cards

and presented one at a time in a preset random order. This differs from the procedure used by Hayes (1987), who presented the recognition stimuli aurally. The current procedure was used to lessen the working memory load on subjects when being tested for the unique combinations. For the recognition test of the unique combinations, the nine combinations used in the task, as well as nine other combinations of the shared sentences, were typed on cards. For the new combinations, each of the shared sentences was used three times, once in the top, middle, and bottom positions of the card. The new combinations were constructed to be as different as possible to the actual combinations. The combinations were presented one at a time in a preset random order.

Procedure. The nature of the task was described to subjects, and subjects performed the task until they reached the performance criterion. They were then given the transfer tests. Subjects were assigned on an alternating basis to being tested for the unique sentences first or the unique combinations first. Testing for the unique sentences followed the procedure of Hayes (1987; Experiment 5). Specifically, subjects were informed that there were unique sentences, and were asked to recall the unique sentences associated with each town. Subjects were then given a recognition test of the unique sentences; if a particular sentence was judged to be unique, subjects were asked to name the town the sentence was associated with. Similarly, testing for the unique combinations involved first subjects judging whether a particular presented combination was unique to a particular computer person; if the combination was judged to be unique, the subject then stated the town with which he thought it was associated.

Results

Following Hayes (1987), the measures of explicit knowledge of the unique sentences were VR (the number of unique sentences correctly recalled), VRU (the number of unique sentences correctly recognized), and VPA (the number of unique sentences that were recognized and for which the correct town was stated). In addition, there were the following measures of explicit knowledge: VRC (the number of unique combinations correctly recognized), VPAC (the number of unique combinations that were recognized and for which the correct town was stated), and CA (the number of computer people that could be correctly classified on the basis of either the VPA or the VPAC measure). Classification performance measures were the number of blocks to criterion, average reaction time on the last block, and CE (the number of computer people correctly classified on the last block).

Transfer knowledge. Table 2 presents the means obtained in this experiment, and also the means obtained by Hayes (1987; Experiment 5, Group U), for the transfer test measures. It can be seen that the present values for VR and VPA are somewhat higher than those obtained by Hayes. A one-way ANOVA with Test Order as the factor revealed no effect of test order on any of the above measures, all $Fs < 1$. Hence this factor will not be considered further.

Table 2. Transfer test measures.

Measure	Mean	Mean (Hayes 1987)
VR	4.00 (1.35)	2.36
VRU	7.50 (1.24)	6.91
VPA	5.25 (1.14)	3.46
VRC	4.58 (1.78)	----
VPAC	1.33 (1.16)	----
CA	6.00 (1.41)	----

Note Standard deviations appear in parentheses.

Classification performance. Table 3 presents the means obtained in this experiment, and also the means obtained by Hayes (1987; Experiment 5, Group U) for the classification measures. The means across the two experiments are quite comparable.

Table 3. Classification performance.

Measure	Mean	Mean (Hayes 1987)
blocks	8.25 (7.38)	8.73
RT	1.89 (0.43)	2.21
CE	5.83 (1.34)	6.00

Note Standard deviations appear in parentheses.

Comparison of classification and transfer performance. The crucial dissociation found by Hayes (1987; Experiment 5) was between VPA and CE, and also, less importantly, between VR and CE. Comparing VR and CE, $t(11) = 3.19$, $p < .01$, replicating Hayes. However, for VPA and CE, $t(11) = 1.29$, $p > .1$. A more complete measure of transfer knowledge than VPA is CA, which measures the combined knowledge of unique sentences and combinations. Comparing CA and CE, $t(11) = 0.35$, $p > .1$.

Discussion

The aim of Experiment One was to (a) determine the replicability of the findings of Hayes (1987; Experiment 5, Group U), and (b) determine the extent of the transfer knowledge of both unique sentences and combinations. Unfortunately, the key finding of Hayes, of a dissociation between classification performance and transfer knowledge of the unique sentences, was not replicated. Transfer knowledge of the unique combinations was not great (see table 2); but there was no need for it to be as classification performance could be accounted for in terms of transfer knowledge of the unique sentences alone. In fact, subjects often spontaneously stated that they ignored the shared sentences once they had spotted the unique one. Thus, there is no need to postulate implicit learning during the performance of the task.

There is one procedural difference between Experiment One and Hayes (1987; Experiment 5): In Experiment One, the recognition stimuli were presented visually and not aurally, and therefore the conditions at encoding and retrieval matched more closely in Experiment One than they did in Hayes' study. Thus, the lack of transfer found by Hayes may have been an effect of encoding specificity. However, the results of Hayes (1987; Experiment 7) are not consistent with this interpretation: In Experiments 5 and 7 the match between conditions during encoding and retrieval were the same, but only in Experiment 5 did retrieval fail.

In order to investigate the discrepancy in results between Experiment One and Hayes (1987; Experiment 5, Group U), an analysis was undertaken of the expected values of classification and VPA performance for given levels of knowledge of the unique sentences.

Expected values on the VPA measure were simply calculated by assuming that the subject knew a certain number, n, of unique sentences and would always get those right; for the remaining (9 - n) sentences, it was assumed that the subject guessed with a probability of 1/6 of being right (a probability of 1/2 of correctly saying "unique", and a further probability of 1/3 of saying the correct town). In order to determine expected values on the classification task, computer simulations were run, each of 100 subjects. The simulations produced correct classification responses to a certain number, n, of people, and random responses to the remaining (9 - n) people. The mean number of blocks to criterion in Hayes (1987; Experiment 5) was nine for Group U, and 14 for Group S; let 20 be an estimate of the maximum number of blocks taken by any subject. For each simulation of a subject, responses were produced until criterion or the twentieth block was reached, whichever came first. Empirical simulation rather than formal analysis was conducted for the classification task because of the complexity of the stopping rule used by Hayes (1987).

The probability of reaching criterion in 20 or less blocks with purely random classification responses was estimated by a simulation as .13. This was substantially increased if it was assumed that subjects knew the unique sentences of only a few people. In fact, subjects' score (3.5) on the VPA measure in Hayes (1987; Experiment 5, Group U) suggests that subjects knew the unique sentences of two or three people on average. Assuming subjects knew two people and otherwise responded randomly, the expected VPA was 3.17. A computer simulation indicated that with this level of knowledge 66% of subjects would reach criterion in 20 blocks and the average classification score (including those who did not reach

criterion) would be 5.3. If subjects knew three people and otherwise responded randomly, the expected VPA was 4.00. A simulation indicated that 87% would reach criterion by chance alone and the average classification score would be 5.7: This is not very different from the classification score of 6.0 obtained by Hayes (1987; Experiment 5, Group U). Thus, for the level of knowledge indicated by the VPA score, the classification score of Hayes (1987; Experiment 5, Group U) is very close to the expected value based on assumed perfect transfer of underlying knowledge between the two tasks.

Note that the expected difference in VPA and classificaton performance was 2.1 if subjects knew two people and 1.7 if they knew three people. In general, the more people it was assumed that the subjects knew, the less was the discrepancy between the expected VPA measure and the expected classification performance on the last block. If subjects knew four people, then the expected VPA was 4.8 and the expected classification performance was 6.3. If subjects knew five people then the corresponding figures were 5.7 and 6.6. The expected differences are thus 1.5 and 0.9, respectively. The decline in expected discrepancy between VPA and classification with increasing underlying knowledge is consistent with the results of Experiment One. Subjects scored 5.3 on the VPA measure and 5.8 on classification in Experiment One. Subjects' score on the VPA measure in Experiment One suggests that the subjects knew the unique sentences of four or five people on average; compare this with the two or three people probably known by the subjects in Hayes (1987; Experiment 5, Group U). Thus, the results of both Hayes (1987; Experiment 5, Group U) and Experiment One are consistent with a single database, consisting mainly of knowledge of the relationships

between unique sentences and towns, underlying performance on both the classification and transfer tasks.

Three issues have been highlighted by these results with the Residents Task. First, it appears that the procedure used in the Residents Task did not induce implicit learning: Future research should attempt to attempt to change the conditions under which the stimuli are presented to the subject in order to minimize the possibility of explicit learning. Second, the above analysis suggests that future experimental paradigms used to investigate implicit concept formation should equalize baseline probabilities on the classification and transfer tasks to aid clear interpretation of the results. And third, an attempt should be made to ensure the transfer task taps possible knowledge in a reasonably complete way. Experiment Two addressed these issues.

EXPERIMENT TWO:

Modifying the Residents Task

Introduction

The aim of Experiment Two was to investigate implicit learning in a modified version of the Residents Task used by Hayes (1987). Three basic modifications were made to the task. These modifications are described in turn, and then the way in which they addressed the problems of the Residents Task are indicated.

The first change made to the Residents Task was that the underlying rule did not relate a unique sentence to a person but a unique sentence to a town. The structure of the materials is shown below in Table 4. Unlike with the residents tasks used by Hayes (1987). a conjunctive rule, involving either pairs or triplets of sentences, will not result in perfect classification performance.

Table 4. Category structure for Experiment Two.

	Town 1	Town 2	Town 3
Person:	1 ADEF	5 BDEF	9 CDEF
	2 ADEG	6 BDEG	10 CDEG
	3 ADFG	7 BDFG	11 CDFG
	4 AEFG	8 BEFG	12 CEFG

A=COLLECTS STAMPS; B=ENJOYS FISHING; C=IS RIGHT HANDED; D=ENJOYS READING; E=HAS BLUE EYES; F=HAS SHORT HAIR; G=OWNS A BICYCLE

The change in stimulus structure required a change in the transfer test; there was no longer a unique sentence for each person but for each town. Thus, in Experiment Two, transfer knowledge was assessed by asking subjects to underline the important sentence(s) for each town, and to indicate the rule connecting the sentence(s)

to the town. This transfer test, like the VPA measure, tests for the subjects ability to apply their knowledge to elements of an exemplar in isolation.

This transfer test is similar to the one used by Dulany et al. (1984) for Reber's grammar learning paradigm; they asked subjects' to underline the grammatical or nongrammatical part of test exemplars. Although Dulany et al. regarded this test as a measure of explicit knowledge, Reber, Allen, and Regan (1984) argued that an implicit sense of "something wrong" with a particular part of an exemplar could guide the subject to underline the correct part of the exemplar judged to be nongrammatical; and such a sense of something wrong could of course be used for the classification task. The application of the measure was different in Experiment Two of this thesis in that it did not rely on the subject sensing "something wrong" with a particular part of an exemplar; rather the subject had to formulate general rules for all exemplars. The subject had to decide on the relevance of elements of exemplars presented in isolation, and not in the context of a particular exemplar. This may or may not provide a measure of explicit knowledge; the important point is that it tests a possible form of specificity of classification knowledge (does classification knowledge only apply to whole exemplars?)

The final change was that concept learning was incidental. Pilot testing indicated that the task could be easily solved explicitly under intentional conditions. Further, previous demonstrations of implicit learning have involved misdirection of some sort: In the studies by Lewicki (e.g., 1986), Reber (e.g., 1976), and Rommetveit (1960), learning was under incidental conditions; in the dynamic control tasks used by Berry and Broadbent

(1988) and Hayes (1987), the crucial variable occurred on the previous trial. Perhaps the failure of the Residents Task to elicit implicit learning was due partly to the intentional learning conditions used. In Experiment Two, a cover task (counting 'E's, 'H's, or 'S's) exposed subjects to the people and the towns for a fixed number of blocks. Subjects were then unexpectedly asked to classify people into towns for two blocks and without feedback.

This design addressed two problems with the Residents Task outlined in the last chapter: The problem of unequal baseline probabilities on the performance and transfer tasks, and the problem of assessing knowledge in a reasonably complete way. Using the current design the baseline probabilities on the performance and transfer tasks can be equalized, or set to bias an interpretation against implicit learning. To show this, the baseline probabilities of performing the classification and transfer tasks by chance alone were determined; these probabilities are shown in Table 5. They are:

A. The probability of classifying X or more people correctly on average over two blocks. This was calculated using the binomial theorem, and considering all possible combinations. This is the actual performance measure used in Experiment Two.

B. The probability of classifying X or more people correctly using rules elicited from the transfer task, assuming the task cannot elicit knowledge from the database used for classification. To determine these probabilities, a computer simulation was run of 10,000 subjects. The following assumptions were made: 1. Rules could only consist of one, two, or three sentences for each town. Consistently, in the actual experiment subjects never considered more than three sentences for a town. 2. Each rule length occurred

with equal probability independently for each town. 3. Each rule could be conjunctive or disjunctive with equal probability independently for each town. 4. If two or three rules clashed in assigning a person to a town, the person was assigned to the two or three possible towns with equal probability. 4. If no rule assigned a person to a town, the person was assigned to each town with probability 1/3.

Table 5. Probability of classifying X or more people correctly by chance alone on the classification (A) and transfer (B) tasks

X	A	B
5	.25	.29
6	$7.9*10^{-2}$	.12
7	$1.0*10^{-2}$	$3.5*10^{-2}$
8	$8.7*10^{-4}$	$1.5*10^{-2}$
9	$4.0*10^{-4}$	$4.2*10^{-3}$
10	$7.6*10^{-7}$	$6.0*10^{-4}$
11	$4.1*10^{-9}$	$6.0*10^{-4}$
12	$4.0*10^{-12}$	$1.0*10^{-4}$

For each X, it can be seen that the probability of reaching that performance level on the transfer test (B) is considerably higher than on the classification task (A), often by several orders of magnitude.

Turning now to the second problem with the Residents Task used in Experiment One, the transfer test used in Experiment Two should allow, within the limits of the subject's memory, a complete assessment of the subject's rules involving the descriptive sentences. The rules elicited from the subject were used to predict

an expected classification performance, and this predicted value was compared to the actual value.

Method

Subjects. The subjects were 12 paid volunteers aged between 18 and 35 from the Oxford University subject panel. No subject had participated in Experiment One.

Materials and Apparatus. As shown in Table 4, the stimuli were 12 computer people described by four sentences, living in one of three towns. The stimuli were presented on a colour monitor by a Sinclair ZX Spectrum. For both the cover task and the classification task, a trial consisted of a single presentation of a computer person followed by a response from the subject, and a block consisted of a pass through all 12 computer people. For each block, the order of computer people was randomly changed, and so was the order of sentences for each computer person. For the classification task, the Spectrum automatically printed the number of people correctly classified on each block and the reaction time to respond to each person on the last block.

Procedure. All subjects received first 25 blocks of the cover task, followed by two blocks of the classification task, and finally the test for explicit knowledge. For the cover task, a trial began by the Spectrum displaying the four sentences describing a person, which were on the screen for six seconds. As soon as the sentences appeared the subject pressed the space bar; this resulted in the town the person lived in appearing under the sentences. The subject was to count the number of 'E's, 'H's, or 'S's in the sentences, depending on whether the person lived in Town One, Two, or Three, respectively. The letters were chosen so as to be roughly equally

distributed in the sentences across the three towns. The letter counted was dependent on the town so that the town would be a salient property of the person. The 25 blocks lasted 45 minutes. For the classification task, the subject was told to remember or guess the town associated with each person; no feedback was given. For the transfer test, the subject was shown the 7 sentences and asked to underline the sentence or sentences important for deciding whether a person lives in the town, and to indicate the rule relating the sentence or sentences to the town. This was repeated for each town.

Results

The dependent measures were the average number of people correctly classified across the two blocks, the predicted number of people correctly classified based on the rules subjects gave, and the reaction time for correct and incorrect classifications (seconds). Table 6 presents the means for these variables.

Table 6. Classification and transfer performance.

Measure	Mean
Average Classification Performance	4.13 (2.08)
Predicted Performance	4.58 (1.51)
Reaction Time Correct	3.07 (0.69)
Reaction Time Incorrect	3.27 (0.78)

Note: Standard deviations appear in parentheses

Actual and predicted correct classifications. The predicted number of classifications was calculated by applying the rules the subject gave to each computer person in turn. If no rule applied, the probability of correctly classifying the person was set as 1/3. If

more than one rule applied, any priority indicated by the subject was followed; if no priority was indicated, the person was assigned to the possible towns with equal probability.

There was no significant difference between the predicted and actual number of correct classifications, $t(11)<1$. The Spearman's correlation between the predicted and the average actual number of people correctly classified was +.52, $p>.05$.

Reaction Time. There was no significant difference between the correct and incorrect reaction times, $t(11)=1.07$.

Discussion

The aim of Experiment Two was to investigate implicit learning with a modified version of Hayes' (1987) Residents Task. There were two problems with the design of the Residents Task: The baseline probabilities on the classification and explicit knowledge tests were unequal, artifactually biasing an interpretation in terms of implicit learning; and the transfer test tapped a small proportion of the subjects' possible knowledge. Experiment Two addressed these problems by changing the nature of the rule relating the people to the towns and by employing a modified transfer test. Additionally, learning in Experiment Two was under incidental rather than intentional conditions, consistent with previous procedures that have shown implicit learning (e.g., Lewicki, 1986; Reber, 1976; and Rommetveit, 1960).

Implicit learning was not demonstrated with the current set of procedures; there was no difference between the number of people the subject could classify and the number of people the subjects' rules could classify. Lewicki (1986) argued that reaction time rather than classification performance was a more sensitive measure

of implicit learning. In the current task, it may be that a response produced by a rule or a guess may be facilitated by implicitly acquired knowledge if the response was correct or inhibited if the response was incorrect; however, there was no significant difference between the reaction time for correct and incorrect responses.

General discussion

This chapter reported two experiments that investigated the possibility of implicit learning in simple concept formation tasks. Experiment One was a direct extension of procedures used by Hayes (1987). The results arising from this experiment demonstrated that the transfer task used by Hayes (1987) did tap the same level of knowledge as classification performance, and provided no evidence for implicit learning. Experiment Two modified both the classification and transfer tasks and found no evidence of learning.

These negative findings suggest that if implicit learning is to be demonstrated in a concept learning paradigm a careful analysis needs to be made of the conditions under which it would be plausibly found. Chapter Three attempts such an analysis.

Chapter Three

A new paradigm for investigating implicit concept formation

Introduction

Chapter Three addresses some of the possible reasons for why implicit learning did not emerge in Experiment Two. Initially, the role of the cover task, of the display characteristics of a task, and of the nature of the rule to-be-learned in inducing implicit learning are considered. Such an analysis suggests specific procedures that may overcome the difficulties of Experiment Two. Then three experiments are reported that incorporate these procedures.

An incidental learning paradigm was introduced in Experiment Two based on the use of incidental learning in previous investigations of implicit learning (Lewicki, 1986; Reber, 1976; Rommetveit, 1960). The only criterion used for the cover task in Experiment Two was that it should interfere with the subject forming correct hypotheses. However, Ward (1988) warned concept formation researchers against treating incidental learning as a monolithic entity; it has long been known that different types of cover tasks have different consequences for learning (Craik and Tulving, 1975). The cover task in Experiment Two, viz. counting the number of letters, may have done more than simply inhibit explicit learning; specifically, it may have biased subjects to process at a level below that necessary to distinguish the categories.

Anderson (1987) argued that all learning involves the contents of working memory; this might apply equally to implicit as to explicit learning. Relatedly, Bowers (1984) urged researchers to distinguish nonconscious influences that are explicitly

uncomprehended rather than unnoticed. Thus, in order for implicit learning to occur it may be necessary for subjects to represent the crucial information in working memory at the correct level. Two sets of results by Lewicki (1986) seem to contradict this statement. First, Lewicki (1986; Experiment 3.1) has obtained implicit learning involving subliminal stimuli. However, the procedure for establishing subliminality was not detailed sufficiently to assess its adequacy (Holender, 1986), and it appears that individual thresholds were not determined for each subject. Second, Lewicki (1986; Experiment 4.9) also obtained learning involving reportedly unattended stimuli. Specifically, subjects formed associations between visually presented objects and threat words on the unattended channel of a dichotic shadowing task. However, the greater the subject's trait anxiety the greater the learning evinced, suggesting the effect depended on the threat words being primed. Further, during presentation of the threat words shadowing performance on the attended channel was disrupted, consistent with the hypothesis that the threat words were attended to, if only briefly.

A further result by McGeorge & Burton (in press; Experiment Three) deserves comment. Subjects counted the number of horizontal lines in 30 sets of four digits. Each set of digits contained at least one "3". Subjects later classified as "old" more new sets containing, rather than not containing, a "3". Although in this case a low level cover task allowed significant learning, McGeorge and Burton's concept task as compared to Experiment Two involved a simpler relationship with simpler material. There may have been sufficient registration of digits in working memory for learning to proceed.

Experiments Three, Four, and Five used procedures to ensure as far as possible the presence of the relevant information in working memory; specifically, these experiments used cover tasks that involved memorizing sentences rather than scanning for letters. The details of the cover tasks will be described later with each experiment.

Characteristics of the experiment other than the cover task will also influence the way in which stimuli are processed. Hammond, Hamm, Grassia, and Pearson (1987) suggested several task characteristics that are intuition rather than analysis inducing; for example, they considered that the display of a large, rather than small, number of continuous, rather than discrete, variables induced intuitive rather than analytic cognition. Hammond et al. (1987) also recommended distinguishing the surface and depth characteristics of a task. Surface refers to the way in which task variables are displayed to the subject and depth refers to the covert relationships between the variables, i.e. the nature of the rule that is to be learned by the subject. Several characteristics of the task used in Experiment Two would tend to induce analytic rather than intuitive cognition. In terms of surface characteristics, variables were discrete rather than continuous, and in terms of depth characteristics, each town was not indicated by multiple cues; only a single sentence for each town was crucial.

Implicit and explicit learning can be hypothetically linked to the intuition or analysis inducing characteristics of a task. Surface characteristics of a task are probably important in determining the mode of learning (Berry and Broadbent, 1988) in which subjects predominantly approach a task. If the majority of the surface characteristics of a task are intuition inducing, then

subjects will probably learn in an implicit rather than explicit mode. However, the final outcome might depend on the congruence between surface and depth characteristics. If the underlying rule is simple, the subject may notice it by spectating on his own performance (Hayes, 1987), and the resulting explicit knowledge may not allow the experimenter to detect the underlying implicit knowledge. Or the rule may not be amenable to implicit learning at all, perhaps by not involving a family resemblance among correct choices (Mathews, Buss, Chinn, & Stanley, 1988). On this view, where there is congruency between intuition inducing surface and depth characteristics, implicit learning is most likely to be demonstrated. For example, the control tasks involve mainly intuition inducing surface characteristics: Displays are brief, variables are continuous, and performance can be perceptually measured when the information is displayed graphically, all characteristics that Hammond et al. (1987) regard as intuition inducing. With an appropriate rule, such as a lagged relationship, implicit learning is observed. These cases may be contrasted to an experimental task in which the majority of surface characteristics are analysis inducing. Subjects would then be more likely to approach the task in an hypothesis-testing mode and implicit learning would be difficult to demonstrate, regardless of depth characteristics.

The surface characteristics of Experiment Two were changed in Experiments Three, Four, and Five by using continuous rather than discrete categories. Specifically, people were not assigned to towns but to particular incomes; each of the 12 people had a unique income between 5000 and 10,500 pounds. This may have three desirable effects. First, the simple presence of a continuous

variable may influence the way subjects approach the task (Hammond et al, 1987). Second, because people lie on a continuum of incomes, the underlying structure of three discrete categories (high, medium, and low income) may not be obvious to subjects. It would thus be harder to formulate accurate hypotheses. Third, in assigning incomes rather than towns to people, the subject may be partially right, rather than simply right or wrong. Thus the average income assigned to people of different categories might be a more sensitive test of subjects' knowledge than the number of people assigned to the correct category.

Another surface characteristic of Experiment Two was changed, namely the simultaneous display of exemplar and category information. The transfer task employed in Experiment Two asked subjects to name the relevant aspects of the exemplars for each category. Thus, implicit learning would have been demonstrated by showing that given exemplar information subjects could name the correct category (classification performance), but given the category subjects could not name the relevant aspect of the exemplars (transfer task performance). That is, the association between the category and the relevant aspect of the exemplar would have been unidirectional. This is analogous to the control tasks, where subjects can produce the correct action for a given situation, but cannot predict the situation that results from a given action (Berry & Broadbent, 1984). This asymmetry will depend on exemplar and category information being differentially processed, and may be facilitated by a temporal asymmetry in which category information is presented only after the exemplar has been presented. Experiments Three, Four, and Five incorporated this temporal relationship.

Depth characteristics were also manipulated in Experiments

Three, Four, and Five. For these experiments, two different tasks were used. One task used the same stimuli and the same stimulus structure as shown in Table 4, Chapter Two. That is, each income category was determined by a single crucial sentence (the single sentence task). The second task used a stimulus structure that might be more appropriate for implicit learning; this structure is shown in Table 7. Each income category was determined by a pair of sentences; each sentence alone was only partially predictive (the double sentence task).

Table 7. Stimulus structure for double sentence task.

Income:	Category 1	Category 2	Category 3
Person:	1 ABCE	5 CDAE	9 EFAC
	2 ABCF	6 CDAF	10 EFAD
	3 ABDE	7 CDBE	11 EFBC
	4 ABDF	8 CDBF	12 EFBD

A=COLLECTS STAMPS; B=ENJOYS FISHING; C=IS RIGHT HANDED; D=ENJOYS READING; E=HAS BLUE EYES; F=LISTENS TO RADIO 2

In summary, Experiments Three, Four, and Five, unlike Experiment Two, incorporated the following features: Procedures were used to ensure as far as possible that the crucial information was represented in working memory (the cover tasks involved the subjects memorizing sentences); a continuous rather than discrete category was used (incomes rather than towns); category information always succeeded exemplar information; and category assignment could depend on either a pair of sentences or a single sentence. Experiment Three differed from Experiments Four and Five in terms of the cover task used; Experiment Five was a tighter test of a result

found in Experiment Four.

EXPERIMENT THREE:

Partial report

Introduction

Experiment Three used partial report as a cover task. A trial consisted of the following sequence: Four sentences were displayed followed by a blank screen, and then an income appeared. Depending on the value of the income, the subject was to report the first two, middle two, or last two sentences. For the single sentence task, the crucial sentence was never a member of the pair of sentences to be reported. For the double sentence task, both crucial sentences never formed the pair of sentences to be reported. This served two purposes. First, at the time the income is displayed the salience of the crucial information would not be enhanced by having to be reported. Hopefully, this would decrease the probability of the subject explicitly noticing a connection between the sentences and the income. Second, the temporal asymmetry between category information and the crucial exemplar information was maintained. The predicted category always occurred after the predictive sentences, but never vice versa.

As in Experiment Two, the cover task was followed by two blocks of the classification task, and then transfer testing. For Experiment Three, two transfer tasks were used: "Free recall", in which subjects were shown the sentences and asked to describe exactly how they performed the classification task; and the same structured transfer task used in Experiment Two. The baseline probabilities for the classification task are shown in column A of Table 8, and for the structured transfer task, in columns B (for the single sentence task) and C (for the double sentence task). Columns

A and B are the same as for Experiment Two. To determine the probabilities for column C, i.e. the double sentence rule, exactly the same procedure was followed as for the single sentence rule. That is, a computer simulation was run of 10,000 subjects. The following assumptions were made: 1. Rules could only consist of one, two, or three sentences for each town. 2. Each rule length occurred with equal probability independently for each town. 3. Each rule could be conjunctive or disjunctive with equal probability independently for each town. 4. If two or three rules clashed in assigning a person to a town, the person was assigned to the two or three possible towns with equal probability. 5. If no rule assigned a person to a town, the person was assigned to each town with probability 1/3.

Table 8. Probability of classifying X or more people correctly by chance alone on the classification (A) and transfer (B,C) tasks

X	A	B	C
5	.25	.29	.28
6	$7.9*10^{-2}$	.12	$7.7*10^{-2}$
7	$1.0*10^{-2}$	$3.5*10^{-2}$	$1.7*10^{-2}$
8	$8.7*10^{-4}$	$1.5*10^{-2}$	$3.7*10^{-3}$
9	$4.0*10^{-4}$	$4.2*10^{-3}$	$8.0*10^{-4}$
10	$7.6*10^{-7}$	$6.0*10^{-4}$	$1.0*10^{-4}$
11	$4.1*10^{-9}$	$6.0*10^{-4}$	$<1.0*10^{-4}$
12	$4.0*10^{-12}$	$1.0*10^{-4}$	$<1.0*10^{-4}$

Method

Design. There was one between subjects factor, Task Type. Half the subjects were assigned to the single sentence task and half the

subjects to the double sentence task.

Subjects. The subjects were 24 paid volunteers aged between 18 and 45 from the Oxford University subject panel. No subject had participated in Experiments One or Two.

Materials and Apparatus. For both single sentence and double sentence tasks, the stimuli were 12 computer people described by four sentences and falling into three categories (see Tables 4 and 7, respectively). There are six different ways of ordering the three categories in terms of high, medium, and low income; each way was used for both Task Types with an equal number of subjects. Thus, the sentences were counterbalanced across income levels. People in the low income category had incomes of 5000, 5500, 6000, and 6500 pounds; people in the medium income category had incomes of 7000, 7500, 8000, and 8500 pounds; and people in the high income category had incomes of 9000, 9500, 10,000, and 10,500 pounds. Within an income category, the ranking of people in terms of incomes was the same as the ranking of people shown in Tables 4 and 7. The stimuli were presented on a colour monitor of a Sinclair ZX spectrum.

For the cover task, a trial consisted of a presentation of four sentences describing a person for six seconds, followed by a blank screen for four seconds, and then a statement indicating the income of the person. After the subject made his response, he pressed a button to clear the income and start the next trial. A block consisted of a pass through all 12 computer people. For the single sentence task, the appropriate crucial sentence for each income category, "A", "B", or "C", occurred randomly in either position that was not to be reported; for example, for the low income category (say, category 1 in Table 4) the crucial sentence

(that is, sentence "A") occurred randomly in either of the bottom two positions (for the low income, the subject was to report the top two sentences). The remaining sentences describing the person were randomly assigned to the remaining positions. For the double sentence task, one of the appropriate crucial sentences for each income category ("A", "C", or "E") occurred randomly in either position that was not to be reported. The order of computer people was randomly changed on each block.

During the classification performance phase, a trial consisted of a presentation of a computer person for six seconds followed by a response from the subject, and a block consisted of a pass through all 12 people. There was no feedback. For each block, the order of computer people was randomly changed, and so was the order of sentences for each computer person. The Spectrum automatically printed the number of people correctly classified on each block, and the reaction time and the specific incomes suggested by the subject on the last block.

Procedure. All subjects received first 10 blocks of the cover task, followed by two blocks of the classification task, and finally the transfer task. For the cover task, subjects were told that it was a memory task and instructed to report the first two sentences if the income was between 5000 and 6500 pounds, the middle two sentences if the income was between 7000 and 7500 pounds, and the last two sentences if the income was between 9000 and 10,500 pounds. They were not informed about the remaining tasks. The 10 blocks lasted 45 minutes; thus, subjects were exposed to the stimuli for the same duration as in Experiment Two. For the classification task, subjects were asked to remember or guess the income associated with each person; no feedback was given. For the free recall test,

subjects were shown the seven (single sentence task) or six (double sentence task) sentences and asked to write down how they assigned incomes to people, if they used any rules involving the sentences, and if they remembered any particular people. They were then asked in the structured task to underline the sentence or sentences important for deciding whether a person belonged to a income category, and to indicate the rule relating the sentence or sentences to income category; this was repeated for each income category.

Results

The dependent measures were the number of people correctly classified across the two blocks, the predicted number of people correctly classified (based separately on the free recall and structured tasks), and reaction time for correct and incorrect classifications (seconds). Table 9 presents the means for these variables.

Table 9. Classification and transfer task performance.

Measure	Task: Single	Double
Classification Performance	4.04 (0.75)	4.00 (1.02)
Predicted Performance:		
Free recall	4.06 (0.78)	4.14 (1.40)
Structured test	3.41 (1.63)	3.76 (1.01)
Reaction Time Correct	11.03 (3.34)	12.20 (3.16)
Reaction Time Incorrect	11.05 (3.26)	11.92 (2.68)

Note: Standard deviations appear in parentheses

The mean classification performance and predicted performance expected by chance alone is 4.00. None of the

classification or predicted performance measures differed significantly from 4.00, ps>.10.

As expected from the means for classification performance, a 2 X 2 (Performance Type (Actual versus Predicted) by Task Type) mixed model analysis of variance indicated no significant effects, whether predicted performance was based on free recall (all Fs < 1) or the structured test (Fs of 1.73, 0.22, and .35 for the main effect of Performance Type, for the main effect of Task Type, and for the interaction, respectively). Classification performance was subjected to a further analysis. For each subject, the mean income assigned to people from each category was calculated, and the order of the means compared to the order high income category > medium income category > low income category. Four out of 24 subjects had the means in the correct order; that is, exactly the amount expected by chance alone. For reaction time, a 2 X 2 (Correct versus Incorrect by Task Type) mixed model analysis of variance indicated no significant effects, all Fs < 1.

Discussion

Experiment Three provided no evidence for learning, either implicit or explicit. These results contrast in spirit with the view that implicit learning

> "[takes] place quite naturally and simply in any subject who devote[s] sufficient attention to a structured stimulus environment (Reber & Lewis, 1977, p. 333)".

Rather, the results suggest that for learning to take place, the subject should be active in some suitable way. Similar results in analogous domains have been found by Schacter (e.g., Schacter & Graf, 1986) and Hartman, Knopman, and Nissen (1989). Schacter has

found that repetition priming of new word-to-word associations in fragment completion (Schacter & Graf, 1986; Graf & Schacter, 1989) and free association (Schacter & McGlynn, 1989) depends on initial elaborative processing of the word pairs, and not simply exposure to them (see Chapter Seven for a discussion of possible relationships between "implicit memory", as defined by Schacter, 1987, and "implicit learning" as defined in this thesis).

Similar results were also obtained by Hartman et al. (1989). They obtained procedural learning in a task that required categorizing presented words; the speed to categorize decreased when the sequence of categories was repeated in a fixed order. This procedural knowledge may be implicit as the knowledge did not transfer to a task that required predicting the next category to appear. The important point is that procedural learning only appeared with tasks that Hartman et al. (1989) argued required some effort (like categorization); simple word naming did not lead to procedural learning.

Reber has previously used exemplar memory as a cover task (e.g., Reber, 1967), and memorizing exemplars may be a suitably active task for inducing concept learning. Reber has also used the seemingly passive task in which subjects are simply asked to observe exemplars (e.g., Reber & Allen, 1978) and still obtained learning in a concept formation task. However, in the "observation" task, the effort of subjects to determine what will be required of them cannot be ignored; even in this situation subjects will probably attempt to memorize exemplars, and thus be engaged in a suitably active way with the stimulus material. The cover task given in Experiment Three may have been sufficiently convincing to subjects that they did not attempt to commit the information to long term memory.

EXPERIMENT FOUR:

Paired associate learning

Introduction

The aim of Experiment Four was to investigate implicit learning using a cover task that was more "active" than the partial report task used in Experiment Three. Initially, the task is briefly described, and then the possibly relevant ways in which a subject performing the task would be "active" rather than "passive" are considered.

For the cover task in Experiment Four, subjects were simply asked to memorize which income went with which person. The absence of a clear category structure should provide misdirection for subjects if they attempt to formulate hypotheses, in a similar way that the lagged relationships do in the control tasks (Berry & Broadbent, 1988; Hayes, 1987). A trial consisted of the following sequence: A person was displayed, the subject typed in the income he thought the person might have, and then the actual income appeared. The subject continued until a certain criterion had been reached (8 or more people correctly classified), or time had run out (10 or 15 blocks, depending on the learning rule). Subjects were not informed of the criterion.

Whereas in Experiment Three subjects were merely exposed to the relevant information while performing a task that required no more than registering the information in working memory, in Experiment Four subjects were more actively involved in forming links between the exemplars and the category. The task used in Experiment Four satisfies several ways in which activity might be relevant, depending on whether one adopts a PDP, ACT* (J. R.

Anderson, 1983), "procedures of mind" (Kolers & Roediger, 1984), or reinforcement learning (Barto, Sutton, & Brouwer, 1981) perspective on implicit learning:

1. Implicit concept formation might result from the operation of a holographic or PDP-type long term memory system (see Chapter Five for specific models). Implicit learning would then depend on the use of strategies that transfer exemplar and category information from working memory to long term memory. However, the PDP viewpoint does not entail that subjects have a clearly differentiated memory of the exemplars. Many PDP systems can only accurately retrieve individual exemplars if all the stored exemplars are coded with orthogonal sets of features (e.g., J. A. Anderson, 1983). In Experiment Four, the exemplars did not have orthogonal sets of features and so the weight changes associated with successive exemplars could result in the correct generalizations about income while destroying the ability of the system to retrieve any particular exemplar. It should be noted that the hypothesized sufficiency of attempted exemplar storage in leading to implicit learning contrasts with the finding that learning exemplars does not automatically lead to category learning (Medin, Dewey, & Murphy, 1983).

2. Implicit learning might require the activation of the goal to respond with the correct category during the learning phase of an implicit concept formation task. For example, according to ACT* theory (J. R. Anderson, 1983), productions fire only if the goal in their conditions matches the goal in working memory. Also according to ACT* theory, a production can form without an intermediate declarative (explicit) stage only by being generalized or discriminated from existing productions. Thus, the production

useful during the testing phase of the experiment inherits its goals from the productions employed during the learning phase. If no production in the learning phase involves the goal to respond with the correct category, then no production will be generalized or discriminated that will be useful to the subject in the testing phase, and no implicit learning will be observed. In Experiment Four, it is plausible that if a subject could not remember the exact income he would have as a subgoal to respond with an income in the right neighbourhood. The cover task in Experiment Three required no relevant goals or subgoals.

3. Implicit learning might require a close matching of the mental operations required in the learning and testing phases. According to Kolers and Roediger (1984), the mind is best seen in terms of domain specific procedures or skills. Because of the specificity of mental procedures, there is no reason to expect that knowledge acquired in a partial report task will transfer to a task, such as categorization, requiring quite different mental procedures. In Experiment Four, there was no separate learning and testing phases and so the problem does not arise.

4. Implicit learning might require the subject, in the context of an exemplar, to choose a category for which he subsequently receives reinforcement, positive or negative depending on context (see Chapter Six for specific models). The actions of subjects in Experiment Four, but not Experiment Three, could be viewed in this way. Implicit learning would then be the development of a look-up table according to some reinforcement learning principle (e.g., Barto, Sutton, & Brouwer, 1981). Assuming Barto reinforcement learning in Experiment Four, with reinforcement a continuous function of the difference between suggested and actual income, an

income in the correct rather than incorrect category would on average produce less negative reinforcement and thus would be subsequently favoured in the context of the exemplar.

Thus, the memory task used in Experiment Four satisfied several ways in which activity might be relevant to the emergence of implicit learning.

Because subjects classified until they reached a certain level, the baseline probabilities for this procedure favours the artifactual demonstration of implicit learning. The probability of achieving given levels of predicted performance on the structured transfer task by chance alone are as given in columns B and C of Table 8. However, the probability of reaching the classification performance criterion (eight or more out of 12 people classified correctly at least once in 10 or 15 blocks) is considerably higher than the corresponding entry in column A of Table 8 (which refers to the average performance over two blocks). With the stopping rule used in Experiment Four, if subjects learn rapidly then the effect of chance influences on performance will be minimized; if they learn slowly then the effects of chance influences may predominate over learning. In this sense, Experiment Four can be regarded as exploratory. Initially, it seemed desirable to probe the possibility that the current task characteristics might elicit implicit learning.

Method

Design. There was one between subjects factor, Task Type. Half the subjects were assigned to the single sentence task and half to the double sentence task.

Subjects. The subjects were 24 paid volunteers aged between 18 and

45 from the Oxford University subject panel. No subject had participated in Experiments One, Two, or Three.

Materials and Apparatus. The same materials were used as for Experiment Three. A trial consisted of the presentation of a computer person for six seconds, followed by a blank screen while the subject typed in his response, and then a display of the actual income of the person. A block consisted of a pass through all 12 people. For each block the order of computer people was randomly changed, and so was the order of sentences for each person. The Spectrum automatically printed the number of people correctly classified on each block, and the reaction time and the specific incomes suggested by the subject on the last block.

Procedure. Subjects were told that they were to perform a memory task. They were not informed of the existence of categories or of the transfer tasks. For the single sentence task, subjects continued until they had classified eight or more people correctly in a block, or until 10 blocks had elapsed. For the double sentence task, subjects continued until they had classified eight or more people correctly, or until 15 blocks had elapsed; the upper limit had to be increased in this case because of the greater difficulty of the double compared to single sentence task. The same transfer tests were used as Experiment Three; specifically, subjects first answered in an open ended way how they went about doing the task, what rules they used, and what people they remembered (the free recall task). The category structure was then shown to subjects and they were asked to underline the important sentences for each category (the structured task).

Results and discussion

The dependent measures were the number of blocks, the number of people correctly classified, the predicted number of people correctly classified (based separately on the free recall and structured tasks), and reaction time for correct and incorrect classifications. Table 10 presents the means for these variables.

Table 10. Classification and transfer task performance.

Measure	Task: Single	Double
Number of blocks	5.58 (3.03)	10.50 (4.83)
Classification Performance	8.17 (2.25)	7.08 (1.68)
Predicted Performance		
Free recall	8.26 (2.37)	4.63 (1.08)
Structured	8.10 (2.93)	4.57 (1.05)
Reaction Time Correct	10.75 (2.79)	11.35 (3.72)
Reaction Time Incorrect	11.01 (4.34)	11.40 (2.84)

Note: Standard deviations appear in parentheses

To calculate predicted performance for the free recall task, both particular people and general rules recalled were used to derive a predicted performance. If a subject did correctly recall one or more people then, to be conservative, the probability of correctly classifying the person was taken to be 1.0, regardless of the probability predicted by the subject's rules. Two subjects failed to reach criterion in 10 blocks on the single sentence task, and five subjects failed to reach criterion in 15 blocks on the double sentence task. Considering predicted performance based on the structured test, a 2 X 2 (Performance Type by Task Type) mixed model analysis of variance on classification performance indicated significant main effects of Performance Type, $F(1,22)=10.24$, $p<.005$, Task Type, $F(1,22)=9.36$, $p<.01$, and a significant interaction

effect, $F(1,22)=9.21$, $p<.01$. The interaction was analyzed further with t-tests. The difference between actual and predicted performance was not significant for the single sentence task, $t<1$, but was for the double sentence task, $t(11)=5.29$, $p=.0005$. This is potentially an important finding, but it should be interpreted cautiously because of the problem of unequal baseline probabilites, mentioned earlier. Exactly the same pattern of results is obtained if predicted performance is based on free recall. A 2 X 2 (Response (correct vs incorrect) by Task Type) mixed model analysis of variance on reaction time indicated no significant effects, all Fs < 1.

The Spearmans correlations between actual and predicted performance were significant for the single sentence task: $r=.61$, $p<.05$, for predicted performance based on the structured test, and $r=.60$, $p<.05$, for predicted performance based on free recall; but not for the double sentence task: $r=.39$ and $-.05$, respectively, $ps>.1$. The lack of correlation on the double sentence task would be consistent either with classification performance being based on a distinct implicit knowledge base or with classification performance being increased by chance factors unrelated to the subjects' underlying knowledge.

Before going on to consider Experiment Five, where the influence of chance factors on classification performance was investigated, another possible reason for the superiority of classification over transfer performance on the double sentence task should be dealt with. Specifically, whereas the structured transfer task requested category level rules, the classification task could be performed purely on the basis of exemplar level information. On the free recall task, subjects recalled only .75

computer people on average (this is different to the predicted performance based on free recall given in Table 10, where rules were used as well), but perhaps they would be able to recall many more incomes if they were prompted with complete descriptions of people. If the subject only used category level information to assign incomes to people, then the probability that a person will be assigned the correct specific income given that the person was assigned to the correct category is 1/4. Thus, considering the people that were correctly classified but not correctly recalled, the expected number of people assigned to the correct specific income can be calculated on the null hypothesis that the subject only has category level information for computer people not recalled; this is (number of people correctly classified minus those correctly recalled)/4. Call this figure the expected correct income assignment; the average was 1.58 (standard deviation 0.48). The number of people actually assigned to the correct specific income without being correctly recalled was 2.17 (standard deviation 1.27). The difference - that is, 0.6 people - is not enough to account for the two person difference between actual and predicted classification performance. Further, a t-test indicated that the difference between expected and actual correct income assignment was not significant, $t(11)=1.76$, $p>.1$. That is, it does not appear that subjects could accurately recall the income associated with particular people.

To summarize, Experiment Four found a discrepancy between predicted and actual performance on the double sentence rule. However, Experiment Four used a stopping rule on the classification performance task that would allow chance fluctuations to be important under some conditions. Specifically, subjects were

stopped if they reached criterion at any point in 15 blocks (for the double sentence rule), rather than being tested for a fixed number of trials. Thus, the results are consistent with implicit category level information or chance fluctuations increasing classification performance above that predicted. Experiment Five attempted to determine which factor was responsible.

EXPERIMENT FIVE:

Paired associate learning on the double-sentence task

Introduction

The aim of Experiment Five was to replicate, using appropriate baseline probabilities, the finding in Experiment Four of a difference between classification and transfer task performance. In Experiment Four, the chance probability of doing well on the classification performance task was greater than on the transfer tasks, and this difference in probabilities may have artifactually produced the difference between classification performance and transfer knowledge. To address this problem, Experiment Five differed from Experiment Four in that subjects received a testing phase, consisting of two blocks of classification performance, after completing the learning phase. The same transfer tests were used as in Experiment Four. The baseline probabilities for various outcomes of the classification testing phase and the transfer tasks for Experiment Five are shown in Table 8, columns A and C, respectively. All probabilities are low and comparable for the two columns.

Method

Subjects. The subjects were 18 paid volunteers aged between 18 and 45 from the Oxford University subject panel. No subject had participated in any of the previous experiments.

Materials and Apparatus. The same apparatus was employed as for Experiment Four; only the materials for the Double Sentence task were used.

Procedure. Subjects performed on the Double Sentence task until

they had classified eight or more people correctly, or until 15 blocks had elapsed. Next, in the testing phase, subjects performed for an additional two blocks without feedback on the double sentence task. Finally, subjects wrote down how they did the task (free recall task) and underlined the crucial sentences for each income category (structured transfer task). Any ambiguity in the subjects' rules was clarified with further questioning.

Results and discussion

The dependent measures were the number of people correctly classified on the last block of the learning phase (the classification performance measure of Experiment Four), the number of people correctly classified across the two blocks of the testing phase, and the predicted number of people correctly classified (based separately on the free recall and structured tests). Table 11 presents means for these variables.

Table 11. Classification and transfer task performance.

Measure	
Classification Performance	
Last block of learning phase	6.83 (1.86)
Average for testing phase	5.03 (1.63)
Predicted Performance	
Free recall	4.83 (1.42)
Structured task	4.81 (2.28)

Note. Standard deviations appear in parentheses.

The means for predicted performance on the free recall and structured tasks and classification performance on the last block of

the learning phase were comparable to those obtained in Experiment Four (4.63, 4.57, and 7.08, respectively). *t*-tests indicated that the key result of Experiment Four was replicated, that is, predicted performance was significantly lower than classification performance on the last block of the learning phase, $t(17)=4.01$, $p<.001$, for free recall, and $t(17)=3.19$, $p<.05$, for the structured task. Also in accord with the results of Experiment Four, the Spearman's correlation between predicted performance and classification performance on the last block of the learning phase was nonsignificant, .18 and .21 for the free recall and structured tasks respectively, $ps>.1$. However, there was a significant drop in performance from the last block of the learning phase to the testing phase, $t(17)=3.93$, $p<.005$, indicating that the measure of performance in the learning phase had probably been inflated by chance fluctuations in subjects' performance. There was no significant difference between predicted performance and performance in the testing phase, $ts<1$. The power of this test for detecting a population one-person difference is an ample .83.

Table 12 shows the classification performance and predicted performance separately for subjects who did or did not reach criterion. Nine people failed to reach criterion in 15 blocks. A 2 X 2 (Classification Performance in the testing phase vs Predicted Performance by Above vs Below criterion in the learning phase) mixed model analysis of variance indicated no significant effects, all $Fs<1$ whether predicted performance was based on the free recall or structured tasks.

Table 12. Performance for subjects below and above criterion.

Measure	Below criterion	Above criterion
Classification Performance		
Average for testing phase	4.67 (1.27)	5.39 (1.93)
Predicted Performance		
Free recall	4.43 (1.54)	5.18 (2.89)
Structured task	4.57 (0.92)	5.09 (1.80)

Note. Standard deviations appear in parentheses.

The correlation between predicted performance and performance in the testing phase was .65 for free recall and .67 for the structured task, $ps < .05$.

General discussion

The aim of Experiments Three, Four, and Five was to explore the possibility of implicit learning in a modified version of the concept formation task used in Experiment Two. Experiments Three, Four, and Five differed from Experiment Two in terms of both surface and depth characteristics. In terms of surface characteristics, Experiments Three, Four, and Five used a continuous rather than discrete response category (income rather than towns) and category information always succeeded exemplar information. In terms of depth characteristics, category assignment could depend on either a pair of sentences or a single sentence. Additionally, Experiments Three, Four, and Five used procedures that emphasized processing information to a level that could be used to distinguish the

categories. Experiment Three differed from Experiments Four and Five in terms of the cover tasks used: In Experiment Three, the subjects performed a partial report task in which the income indicated which of the person's characteristics to report; in Experiments Four and Five, the subjects performed a paired associate memory task which involved linking each person with an income.

In Experiment Three no learning was observed. In Experiments Four and Five, subjects learned, but not implicitly. The contrast of the results of Experiment Three as compared to Experiments Four and Five suggests that the emergence of concept learning under incidental conditions depends on an appropriately active subject (see also Flint, 1979); what sort of activity is appropriate is a matter of conjecture. It might depend on the use of strategies to ensure the long term encoding of information; the activaton of appropriate goals; the practice of a highly specific skill; or the giving of a response with context-contingent reinforcement. Unfortunately, the results do not shed any light on what sort of activity is necessary for distinctively implicit learning.

Experiments Three, Four, and Five provided no evidence that subjects can learn a concept formation task implicitly. In terms of previous concept formation studies, these results are consistent with those of Carlson and Dulany (1985) but contrast with those of Reber (e.g. 1976); see Chapter One for a review. Experiments Three, Four, and Five and Carlson and Dulany (1985) may have differed from Reber (1976) in that the former failed to establish the conditions necessary for implicit concept formation to occur. Alternatively, the discrepancy in results may be due to the more powerful method of eliciting transfer knowledge in the

former studies, where rules were elicited in a variety of ways and used to derive a predicted classification performance. It should be noted that in Experiments Three, Four, and Five free recall alone did account for performance; but Reber (e.g., Reber & Lewis, 1977) has argued that free recall is insufficient to account for artificial grammar learning. However, as discussed in Chapter One, predicted performance based on free recall and actual performance in artificial grammar learning have not been compared in a systematic way, and so the issue is still open.

Because new concept formation tasks could be generated indefinitely without acheiving the necessary conditions for implicit concept formation, the most efficient way of resolving this issue would probably be to apply the more powerful transfer tests to Reber's artificial grammar learning task, for which it is claimed the necessary conditions have already been established (e.g., Reber, 1989). That is, rules could be elicited from subjects by free recall and structured tests and used to derive a predicted classification performance for the artificial grammar learning task. In this way, the extent to which the knowledge underlying classification performance will transfer to other tasks can be determined. If there is found to be a high specificity of transfer then the differences between artificial grammar learning and the tasks used in Experiments One to Five could be explored; if there is found to be considerable transfer then the notion of implicit concept formation would be undermined.

Chapter Four

Artificial grammar learning

Introduction

The aim of Chapter Four was to explore implicit concept formation using a concept formation task already established in the implicit learning literature; namely, Reber's artificial grammar learning task. Initially, the reasons for adopting this task are discussed, and then three experiments using artificial grammar learning are reported.

Experiments One to Five jointly indicated a close correspondence between classification performance on various concept formation tasks and transfer knowledge. In Experiments One to Five various manipulations were made that were plausibly relevant in inducing implicit learning: Learning was intentional or incidental; single or multiple cues indicated the correct category; relevant information was or was not required in working memory for the cover task; and subjects were or were not informed about category structure. However, there was no manipulation that could detectably impair the transfer of classification knowledge to free recall and other measures. Perhaps some further small change would induce implicit learning; for example, increasing the intuition inducing surface characteristics by describing computer people with continuous dimensions rather than discrete features. Thus, one strategy for the thesis would be to carry on changing the tasks developed so far in the hope that implicit learning would eventually emerge. A more efficient strategy might be to address one of the tasks that provided the initial inspiration for implicit concept formation, using the knowledge elicitation techniques employed in

Experiments One to Five. In this way, a task could be used where it is claimed the necessary conditions for implicit concept formation have already been established.

Various investigators have shown that Reber's grammar learning task (Reber, 1967) possesses many characteristics associated with implicit learning. In this task, subjects are exposed to strings of letters generated by a finite-state grammar, and are later asked to classify new strings as grammatical or not. Recently, it has been claimed that learning is not interfered with by a concurrent memory demanding task (Hayes, 1987), by incidental conditions (Mathews et al., 1989), or by psychiatric disorder (Abrams & Reber, 1988). Further, Reber (e.g., 1967, 1989) has claimed that with the artificial grammar learning task subjects acquire insufficient explicit knowledge to account for classification performance. Thus, artificial grammar learning would provide a good paradigm for rigorously testing the notion of implicit concept formation.

Reber's task differs from the tasks used in Experiments One to Five in at least two ways: First, in the Reber task, explicit and other transfer knowledge has not been systematically tested (see Chapter One). Second, in the Reber task, the subject is typically only exposed to positive exemplars during learning, whereas in most concept formation tasks, including those in Experiments One to Five, the subject is exposed to both positive and negative exemplars during learning. The first difference might result in implicit concept formation being claimed (in artificial grammar learning) when it is not there, and the second difference might result in implicit learning not being detected (in Experiments One to Five) because the task conditions were not suitable for invoking it.

Experiment Six employed the artificial grammar learning task and investigated the influence of both these differences.

EXPERIMENT SIX:

Learning an artificial grammar

Introduction

The aim of Experiment Six was, first, to determine the extent to which artificial grammar learning transferred to free recall and appropriate structured knowledge measures; and, second, to determine the influence of exposure to both positive and negative exemplars, rather than only positive exemplars, on implicit learning.

An important question for artificial grammar learning is what would be an appropriate transfer test. Dulany et al. (1984) provided one structured measure to which there was perfect transfer; this result helped to define the conceptual task of subjects as recognizing the well-formedness of complete strings or part strings embedded within complete strings (see Chapter One). Thus, in order to test the specificity of this conceptual task, it would be important to employ transfer tasks of a different nature. A post-performance free recall of the strategies and rules used by the subject would constitute a different task and also would elicit explicit knowledge. But such a procedure could quite plausibly be regarded as insensitive and incomplete (e.g., Brewer, 1974; Erdelyi & Becker, 1974; Tulving, 1983). For example, in the Conditioning Without Awareness literature, it has been found that the subject may show knowledge of the experimental contingencies when asked specific questions even when such knowledge is not shown by free recall (for a review, see Brewer, 1974). Whereas free recall gives the subject the option of not responding, forced choice questions do not. In the case of artificial grammar learning, the experimental

contingencies relate to which letters can occur at each position in a sequence. Thus, Experiment Six used a test that specifically asked subjects which letters could occur after different stems, varying in length from 0 letters upwards (the test of Sequential Letter Dependencies, or SLD test). By asking the subject to formulate general rules with reference to the presented constituents of exemplars, and out of the context of a particular exemplar, the SLD test is analogous to the transfer tests used in Experiments One to Five. Because the SLD test asks subjects to judge the well formedness of part strings not embedded within whole strings (exemplars), it tests the transfer of classification knowledge to a task different to the conceptual task defined above.

Because the SLD test is analogous to the transfer test used in Experiments One to Five, it is important to see if it can account for performance on artificial grammar learning. If it can, the seeming inconsistency between the results of Experiments One to Five and the results previously obtained with artificial grammar learning would be removed. If it can not, evidence for implicit concept formation would be obtained and the differences between artificial grammar learning and the tasks used in Experiments One to Five could be explored to delineate the task conditions that invoke implicit learning.

In Experiments Three, Four, and Five, the rules elicited from free recall alone were able to predict classification performance. Although free recall has been applied to artificial grammar learning before (e.g., Reber & Allen, 1978; Mathews et al., 1989), it has not been shown that accurately applying the set of recalled rules to each exemplar in turn underpredicts actual classification performance. Thus, Experiment Six applied both the

SLD test and free recall to artificial grammar learning.

In Experiments One to Five, subjects were exposed to exemplars as well as nonexemplars of each category. In artificial grammar learning, subjects have typically been exposed only to positive exemplars, but, in one exception, Brooks (1978) had subjects learn two grammars simultaneously. Subjects associated particular grammatical exemplars with English words, belonging to one of two classes. The grammar from which the exemplar was generated could be determined by the class of word with which it was associated; and indeed subjects could later discriminate exemplars from the two grammars using this cue. Reber and Allen (1978) argued that the Brooks technique induced a learning strategy that inhibited the normal implicit abstraction process. An important aspect of the Brooks technique is its emphasis on learning specific exemplars, but it remains plausible that providing a distinction between two categories may help induce a strategy that inhibits implicit learning. Thus, Experiment Six employed two groups: One saw only grammatical exemplars and the other saw both grammatical and nongrammatical exemplars, distinguished by being presented in different colors. In terms of depth characteristics (Hammond et al., 1987), the second group is exposed to illegitimate associations, and so is different to the groups in Experiments One to Five. In terms of surface characteristics, the second group is exposed to two different types of exemplar, and so is similar to the groups in Experiments One to Five. It may be this surface characteristic, the presence of contrast, that elicits explicit, and inhibits implicit, learning.

In summary, Experiment Six employed two groups that differed at the learning stage: One group saw only grammatical exemplars and

the other group saw both grammatical and nongrammatical exemplars. There were three sets of dependent variables: Classification performance, knowledge elicited in free recall, and knowledge elicited by the sequential letter dependency test.

Method

Subjects. The subjects were 40 paid volunteers aged between 18 and 35 from the Oxford University subject panel.

Materials and Apparatus. The grammar used was the one used by Dulany et al. (1984), Hayes (1987), and Reber and Allen (1978); see Figure 1. The 20 grammatical acquisition exemplars and the 50 grammatical and nongrammatical test exemplars were the ones used by Dulany et al. (1984); see Figure 2. Twenty nongrammatical acquisition exemplars were created, also shown in Figure 2. Five were taken from the nongrammatical test exemplars, and the remaining 15 were made by substituting an inappropriate for an appropriate letter in an otherwise grammatical string. The position of violation covered letter positions one to six over the 15 exemplars.

Figure 1. The finite-state grammer used in Experiment Six.

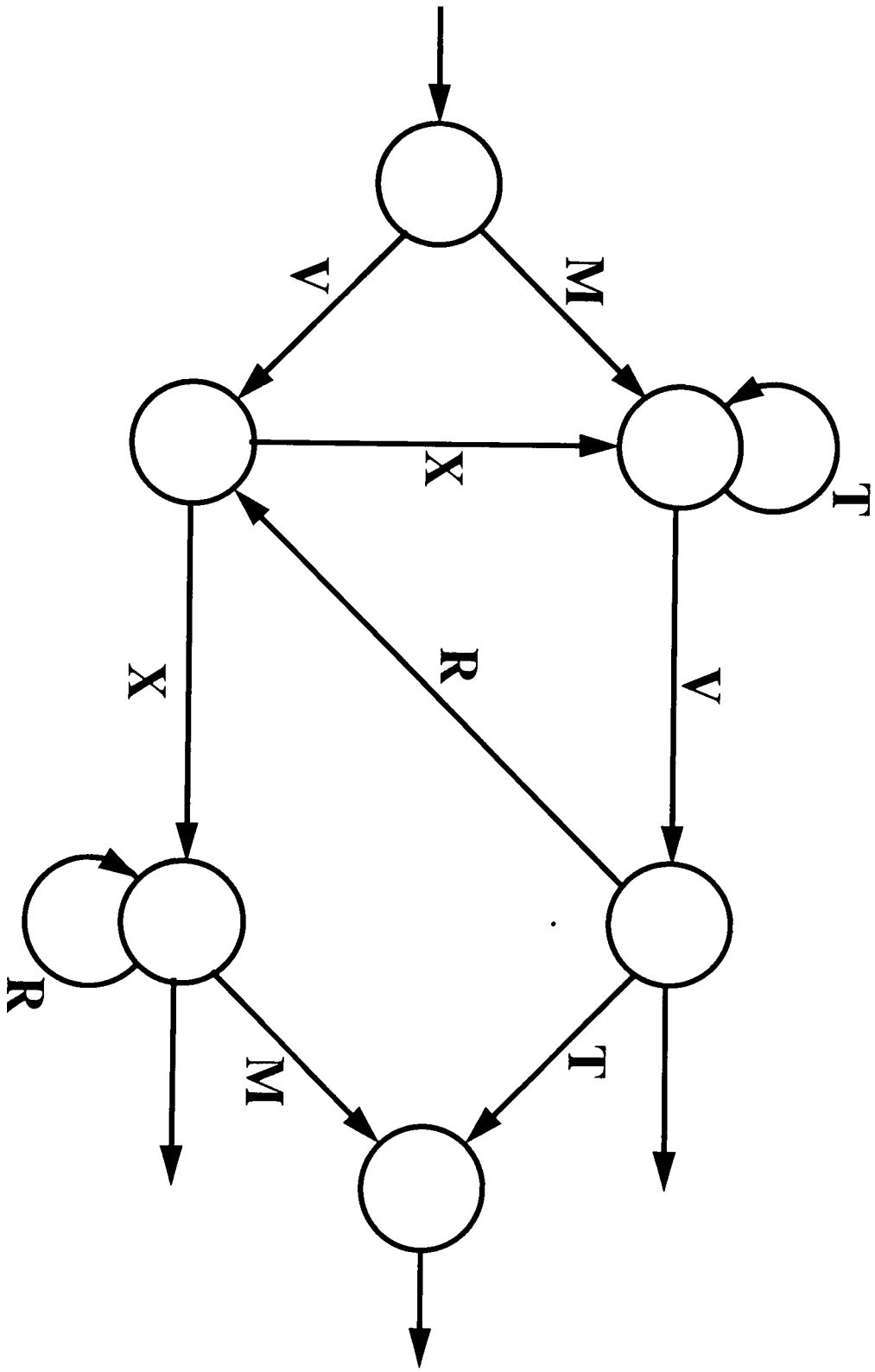

Figure 2.
Exemplars presented in acquisition and test periods.

Acquisition		Test	
Grammatical	Nongrammatical	Grammatical	Nongrammatical
MTTTTV	MTTXTV	VXTTTV	VXRRT
MTTVT	VXTMTV	MTTTV	VXX
MTV	MTVTT	MTTVRX	VXRVM
MTVRX	MTXTV	MVRXVT	XVRXRR
MTVRXM	MTX	MTVRXV	XTTTTV
MVRX	MTTVRT	MTVRXR	MTVV
MVRXRR	VRVT	MVRXM	MMVRX
MVRXTV	VXVTXV	VXVRXR	MVRTR
MVRXV	VTVRX	MTTTVT	MTRVRX
MVRXVT	TXVT	VXRM	TTVT
VXM	RVRXVT	MVT	MTTVTR
VXRR	VXTMVT	MTVT	TVTTXV
VXRRM	VXT	MTTV	RVT
VXRRRR	MMRX	MVRXR	MXVT
VXTTVT	MVRMRR	VXRRR	VRRRM
VXTVRX	MVRTR	VXTV	XRVXV
VXTVT	TTVT	VXR	VVXRM
VXVRX	VXRRT	VXVT	VXRT
VXVRXV	VXX	MTV	MTRV
VXVT	MXVRXM	VXRRRM	VXMRXV
		VXTTV	MTM
		VXV	TXRRM
		VXVRX	MXVRXM
		VXVRXV	MTVRTR
		MVRXRM	RRRXV

During the acquisition phase, each exemplar was displayed on a colour monitor by a Sinclair ZX Spectrum for five seconds, and the total set of exemplars was presented six times in a different random order each time. Randomization was constrained to avoid making the grammar salient. Dulany et al. (1984) presented each exemplar for 10 seconds, and the total set of exemplars was presented three times. In this experiment, individual display.time was traded against number of presentations in an attempt to encourage implicit rule formation or exemplar learning rather than explicit exemplar learning. For the positive exemplars only group (POS), only the 20 grammatical acquisition exemplars were displayed. For the positive and negative exemplars group (POSNEG), all 40 acquisition exemplars were displayed. The grammatical items were displayed in black and

the nongrammatical exemplars in red; grammatical and nongrammatical exemplars alternated.

During the classification phase, each exemplar was displayed in black until the subject pressed "1" to indicate grammatical or "0" to indicate nongrammatical. The 50 test exemplars were repeated once in a different random order. The Spectrum automatically printed the response for each exemplar for each presentation, as well as summary totals.

For the test of sequential letter dependencies (SLD), all possible grammatical stems were generated of length zero to five, with the constraint that the possible exemplars based on the stem could be no more than six letters long and it must be possible for at least one letter to follow the stem. This produced a total of 32 stems, including the "null" stem. The stems were ordered such that previous stems did not contain later stems; see Figure 3 in the Appendix for this chapter. Each stem was displayed in black by the Spectrum. The subject's verbal responses were noted by the experimenter, who pressed a button to display the next stem.

Procedure. For the learning phase, both groups received the following instructions, taken from Dulany et al. (1984) (the variation for the POSNEG group is indicated in brackets):

"This is a simple memory experiment. You will see items made from the letters M, R, T, V, and X. The items will run from 3 to 6 letters in length. You will see a set of 20(40) items. Your task is to learn and remember as much as possible about all 20(40) items."

For the classification phase, the POS subjects were informed that the order of letters in each item was determined by a complex

set of rules; the POSNEG subjects were informed that the order of letters in the black items followed a complex set of rules, but that the order in the red items broke those rules in some way. All subjects were then told that they would now see some more items, only half of which followed the rules, and they were to decide which items followed the rules.

After classifying all the items, subjects were given transfer tests, first with free recall and second with the SLD test. In the free recall test, subjects were asked to indicate how they decided whether an item followed the rules or not, any strategies they used, and any rules they thought the (black) items followed, even if they were not confident as to the correctness of the rules. Subjects were also asked to indicate any specific exemplars they could recall. Subjects were urged to be as complete as possible. Subjects' responses were recorded on tape. In the SLD test, subjects were shown stems, and the experimenter probed with possible next letters (M, V, X, R, and T). To each letter the subject said "yes" or "no" and gave a confidence rating on a five point scale, where "1" indicated a guess and "5" indicated certainty.

Results

Classification performance

The proportions of items judged correctly by groups POS and POSNEG were .65 (SE=.02) and .60 (SE=.01) respectively. The groups differed significantly, $t(38)=2.35$, $p<.05$. These proportions are comparable to the overall proportion correct obtained by Dulany, Carlson, and Dewey (1984) (.63 for the Implicit-Sequential group).

For the two presentations of each exemplar, the mean proportions of judgements that were correct-correct (CC),

error-correct (EC), correct-error (CE), error-error (EE), and the average of the two mixed cases (AV) are displayed in Table 13.

Table 13.

Consistency of judgments

	Group		Dulany et al.
	POS	POSNEG	Implicit-Sequ.
Judgement Type			
CC	.50 (.08)	.44 (.08)	.47
EC	.14 (.06)	.17 (.06)	.14
CE	.16 (.05)	.16 (.06)	.18
AV	.15 (.04)	.16 (.04)	.16
EE	.21 (.08)	.23 (.06)	.21

Note Standard deviations appear in parentheses.

A 2 X 2 (Group (POS vs POSNEG) by Error Type (EE vs AV)) mixed model analysis of variance indicated significant main effects of Group, $F(1,38)=4.35$, $p<.05$, and of Error Type, $F(1,38)=18.74$, $p<.001$. That is, group POSNEG rather than POS made a greater number of both error types. Also, subjects made more error-error than mixed error types, as found by Dulany et al. (1984). Further analysis indicated that the probability of correct classification on the two successive presentations of each exemplar were stochastically dependent; that is, the probability of correct on the second presentation was greater given the subject was correct rather than incorrect on the first presentation, $CC/(CC+CE)=.75$ as compared to $EC/(EC+EE)=.42$, $t(39)=12.67$, $p<.0001$. If probability correct on the first and second presentations of any exemplar are regarded as continuous variables, then the strength of the dependence between

probability correct on successive presentations is indicated by the tetrachoric correlation, r_t, between correct versus incorrect on the first presentation and correct versus incorrect on the second[1]. The tetrachoric correlation calculated on the average data of all subjects was .55; the tetrachoric correlation calculated separately for each subject and then averaged was .52. The latter correlation is a more valid measure of the dependence between probability correct on successive presentations: It does not include effects due to the presence of generally good subjects and generally bad subjects.

Nongrammatical exemplars were classified according to the point at which the nongrammaticality occurred; specifically, an exemplar could be nongrammatical in all letters (PA, three exemplars), at the first letter only (P1, five exemplars), the second letter only (P2, five exemplars), the third letter only (P3, five exemplars), the fourth letter only (P4, four exemplars), the fifth letter only (P5, two exemplars), or the sixth letter only (P6, one exemplar). The mean proportions correct are shown in Table 14. See the Appendix for a breakdown and analysis by groups.

[1]The tetrachoric (rather than phi) correlation is useful in this case because it will later allow a fair comparison with Spearman's correlations calculated on continuous data (see McNemar, 1969, for a discussion of the comparability of tetrachoric, phi, Spearman, and Pearson correlations).

Table 14.

Classification performance and position of violation for nongrammatical items

Position of violation:

PA	P1	P2	P3	P4	P5	P6
.71	.65	.57	.65	.55	.53	.64
(.24)	(.20)	(.20)	(.26)	(.20)	(.29)	(.39)

Note Standard deviations appear in parentheses.

t tests indicated that the overall mean for all exemplar types except for P4, $t=1.58$, and P5, $t=0.95$, differed significantly from chance, $ps<.05$.

Test of Sequential Letter Dependencies

The proportions of correct responses to the SLD test by groups POS and POSNEG were .64 (SE=.02) and .60 (SE=.01) respectively. The difference between the groups was marginally significant, $t(38)=1.88$, $p=.07$. Questions were classified according to the letter position probed; specifically, the questions could relate to the first position (Q1, in this case, the questions involved only one stem), the second position (Q2, two stems), the third position (Q3, three stems), the fourth position (Q4, six stems), the fifth position (Q5, eight stems), or the sixth position (Q6, 11 stems). The mean proportions correct are shown in Table 15. See the Appendix for a breakdown and analysis by groups.

Table 15.

Proportion correct and position probed on SLD test

Position probed:

Q1	Q2	Q3	Q4	Q5	Q6
.68	.68	.72	.63	.61	.59
(.22)	(.15)	(.11)	(.10)	(.09)	(.09)

Note Standard deviations appear in parentheses.

t tests indicated that proportions correct on all question types were significantly above chance, ps<.005. Because a proportion correct of 1.00 required saying "yes" about 40% of the time, the proportion correct for each question type was compared to the expected chance proportion correct given the overall response bias of the subject and the particular response bias required by the question. This did not change the above results.

Performance and SLD

The Spearman's within-groups correlation across subjects between classification performance and proportion of questions correct on the SLD test was .29, p<.05 (a within-groups correlation was used in order to detect any association independent of that already demonstrated by the between groups effects).

Is the level of knowledge elicited by the SLD test sufficient to account for the level of classification performance achieved by subjects? The data was analyzed in two ways to answer this question.

First, d' was calculated for each subject for both classification performance and SLD performance. Table 16 shows the mean values for groups POS and POSNEG. If both the classification and SLD tasks elicited equivalent levels of knowledge, then the d's

for the two tasks should be similar. A 2 X 2 (Group (POS vs POSNEG) by Task (classification vs SLD)) mixed model analysis of variance indicated only a significant effect of group, $F(1,38)=9.22$, $p<.005$. The F for Task was 3.88, which was marginally significant, $p=.06$. That is, subjects tended to have greater d's for the SLD rather than classification task. The F for the interaction was 0.27. The Spearman's within-groups correlation between the d's for the classification and SLD tasks was .29, $p<.05$.

Table 16.

Sensitivities for the classification and SLD tasks.

Task:	POS	POSNEG
Classification	0.72 (0.32)	0.47 (0.28)
SLD	0.79 (0.25)	0.60 (0.24)

Note Standard deviations appear in parentheses.

The second way used to determine if the classification and SLD tasks elicited equivalent levels of knowledge was to apply a transformation to the SLD responses to yield a predicted classification performance. What transformation should be applied? One desideratum of such a transformation would be that there exists evidence that the transformation represents how subjects actually processed a potentially common knowledge base underlying classification and SLD performance. Significant correlations of predicted and actual classification across strings within each subject, and of overall predicted and actual classification across subjects, would help provide evidence for the psychological validity of the transformation. If such a transformation produced a predicted classification performance nonsignificantly different to actual classification performance, there would be evidence that the

SLD test can access the knowledge base underlying classification performance with the same degree of sensitivity as classification performance. The psychological validity of the transformation would be further confirmed if the different average levels of classification performance for POS and POSNEG were matched by identical differences in the predicted classification performances based on the transformation.

Slovic and Lichtenstein (1971) reviewed a large number of studies related to how subjects combine information from a number of sources to make a judgement. A linear (additive) model did a remarkably good job of predicting human judgement over a diverse range of tasks, even when there was evidence that subjects employed some nonlinear strategies: Nonlinear models accounted for minute, even if significant, amounts of additional variance over the linear models. Thus, if subjects are employing SLD-type information for the classification task, a linear combination of SLD information may fruitfully provide a first approximation to a predicted classification performance based on SLD knowledge. Indeed, Hammond et al. (1987) regarded a linear combination of information especially likely in "intuitive" judgements.

What would a linear model concretely mean for a subject faced with classifying artificial grammar strings? One strategy for the subject would be to determine for each exemplar the degree to which each successive letter is legitimate or illegitimate. The more letters that are regarded as legitimate, and the greater the certainty with which they are regarded as legitimate, the more likely the subject may be to answer "grammatical". Additionally, the more letters that are regarded as illegitimate, and the greater the certainty with which they are regarded as illegitimate, the less

likely the subject may be to answer "grammatical". One way to do the classification task, therefore, would be to sum for each exemplar the number of legitimate associations weighted (positively) by a confidence, and the number of illegitimate associations weighted (negatively) by a confidence, and classify all exemplars below a criterion as nongrammatical and those above as grammatical.

Accordingly, for each exemplar for each subject, a weighted sum was made for "legitimate" and "illegitimate" successive letters, where legitimacy was determined by the subject's responses on the SLD test and the weights were the subject's confidence ratings. Legitimate letters added to the sum, and illegitimate letters subtracted from it. A problem arises in applying this procedure to nongrammatical items: Letters subsequent to the position of violation may be important in guiding the subject's grammaticality judgements, but there are no nongrammatical stems in the SLD test. Thus, a fudge was introduced; namely, the letter immediately after the position of violation did not enter into the weighted sums, but thereafter the stem was treated as if it was the corresponding grammatical one. This procedure, by increasing the number of supposed legitimate associations above the actual number, produces a conservative estimate of the discriminating power of the subject's rules. The procedure could be unambiguously applied to 19 of the nongrammatical items; the remaining six exemplars included those that were nongrammatical in all letter positions and the exemplar produced by an inappropriate addition rather than substitution. Thus, these six exemplars were not considered in the subsequent analyses. In order to equalize the number of grammatical and nongrammatical exemplars considered, six grammatical examplars, randomly selected, were discounted from further analysis. These

were the last six exemplars shown in Figure 2. If the number of grammatical and nongrammatical exemplars considered had remained unequal, any systematic bias to respond "grammatical" would have produced above 50% classification performance, even in the absence of valid knowledge of the grammar.

The weighted sum for legitimate and illegitimate associations for each exemplar was divided by the number of associations in the exemplar (i.e., the number of letters minus one for grammatical exemplars, and minus two for nongrammatical exemplars), to produce an average legitimacy minus illegitimacy (SUM) for each exemplar for each subject. For each subject, the correlation between SUM and the tendency to respond grammatical was computed over exemplars. The correlations were small but consistently positive. They were transformed to Fisher's z to normalize the distribution; the mean value (transformed back to a correlation) was .17. Of course, large correlations could be expected only if guesswork did not play a part in either the performance or the SLD test; equally clearly, the small size of the correlations will require consideration. A t-test over subjects indicated that the correlations differed significantly from 0; $t(39)=5.29$, $p<.0001$. A problem arises in interpreting this result in that if the SLD test and classification performance tap independent knowledge bases both of which are to some extent accurate, then both knowledge bases will tend to produce a "grammatical" response when the item is grammatical, and a correlation may emerge between the SLD test and classification performance computed over exemplars. If the knowledge bases are independent, then the correlation should reduce to zero if it is calculated separately for grammatical and nongrammatical exemplars

for each subject. On the other hand, if there is a single knowledge base, then separately calculated correlations should be reduced from the correlation over all exemplars together (since, given a common database, both the SLD test and classification performance would be correct for the same reasons) but to a value greater than zero (assuming the common database is not perfectly accurate). Thus, the correlation between SUM and the tendency to respond "grammatical" was calculated separately for grammatical and nongrammatical exemplars for each subject, and the values z transformed. The mean value (transformed back to a correlation) was .14 for grammatical exemplars and .10 for nongramatical exemplars; these two values did not differ significantly from each other, $t(39)=0.87$. The average of the grammatical and nongrammatical correlations (.12) was significantly lower than the correlation over all exemplars, $t(39)=3.40$, $p<.005$. Importantly, the average of the grammatical and nongrammatical correlations was significantly different from 0, $t(39)=3.45$, $p<.005$.

To calculate a predicted performance for each subject (call it PPSUM), an exemplar was classified as grammatical if the SUM for that exemplar was greater than the mean SUM (for that subject), and as nongrammatical if the SUM was lower than the mean SUM. The means of PPSUM for POS and POSNEG were .63 (SE=.02) and .59 (SE=.02). A 2 X 2 (Group (POS vs POSNEG) by Performance Type (PPSUM vs performance)) mixed model analysis of variance indicated only a significant effect of Group, $F(1,38) = 6.35$, $p<.05$. That is, both PPSUM and classification performance were lower for POSNEG rather than POS. The F for Performance Type was 0.91, and the F for the interaction was 0.06.

The Spearman's within-groups correlation between PPSUM and

performance was -.06. This correlation, which is over subjects, should be distinguished, first, from the correlation between SUM an tendency to respond "grammatical", which was over exemplars, and, second, from the two previous correlations over subjects between the classification and SLD tasks : The correlation between classification performance and proportion of correct SLD responses; and the correlation between d' for classification and d' for SLD. The correlation over exemplars and both these other two correlations over subjects were significant. The absence of correlation between classification performance and PPSUM might have been because PPSUM used only 37% of all the subject's responses to the SLD test, and so was a less reliable measure of the subject's knowledge than measures based on all the subject's responses. To test the plausibility of this explanation, d' was calculated for only those SLD responses that were used to derive PPSUM. If the low correlation between PPSUM and classification performance was due to the relatively few responses on which PPSUM was based, then the correlation between d' for those SLD responses and d' for classification should also be low. In fact, the Spearman's within-groups correlation between d' for just those SLD responses used to calculate PPSUM and d' for classification was -.06. The mean d' on these SLD questions for groups POS and POSNEG was 0.86 (0.09) and 0.58 (0.06), respectively.

SUM was recalculated firstly considering only high confidence rules (given a rating of "4" or "5"), and secondly ignoring the confidence ratings. In the first case the above pattern of results was repeated with perhaps a slightly better fit to actual performance; in the second case the fit to actual performance was somewhat worse.

Another transformation was applied to the SLD responses to

obtain a predicted classification performance. Six was added to each of the subject's SLD responses to yield a positive number between one (definitely nongrammatical) and 11 (definitely grammatical). For each test exemplar, the geometric mean was found of the corresponding SLD responses (ranging from one to 11). Call this mean the PRODUCT for the exemplar. This multiplicative transformation might be regarded as more plausible than the additive one used for PPSUM: A single letter which the subject regarded as nongrammatical would influence PRODUCT more strongly than SUM. Logically, a single nongrammatical letter makes the exemplar nongrammatical. The mean PRODUCT was found for all the test exemplars used in deriving PPSUM, and exemplars were classified as "grammatical" if they were above mean PRODUCT, and "nongrammatical" if they were below: This procedure produced a predicted classification performance called PPPROD. The mean PPPROD for groups POS and POSNEG was .63 (.02) and .58 (.02), respectively. A 2 X 2 (Group (POS vs POSNEG) by Performance Type (PPPROD vs performance)) mixed model analysis of variance indicated only a significant effect of Group, $F(1,38) = 8.51$, $p<.01$. That is, both PPPROD and classification performance were lower for POSNEG rather than POS. The F for Performance Type was 0.86, and the F for the interaction was 0.01.

The Spearman's within-groups correlation between PPPROD and performance was -.01. The Spearman's within-groups correlation between PPPROD and PPSUM was .76. In summary, the multiplicative and additive transformations, yielding PPPROD and PPSUM, respectively, produced virtually identical results.

Free recall and performance

The rules elicited by the free recall test were used to

classify each exemplar in turn, to produce a predicted performance (PPFR) for each subject. Only rules that could be clearly applied were used; for example, "See if sounding the string out makes a word" was not used, as no strings exactly specified a word. If no rule applied to an exemplar, it was assigned a probability correct of .5. The mean values for PPFR for groups POS and POSNEG were .55 (SE=.01) and .52 (SE=.01), respectively. A 2 X 2 (Group (POS vs POSNEG) by Test (PPFR vs Performance)) mixed model analysis of variance indicated significant main effects of Group, $F(1,38)=10.70$, $p<.005$, and of Test, $F(1,38)=50.02$, $p<.001$. That is, group POS was better than POSNEG on both tests. Also, the rules elicited from free recall underpredicted actual performance. The within-groups Spearman correlation between PPFR and performance was .03, between PPFR and number of correct SLD responses was -.03, and between PPFR and PPSUM was -.02. The correlations did not differ significantly between groups.

Control subjects

Six control subjects were run only on the SLD test to determine baseline performance. The proportion of correct responses was .50 (SE=.03), not different from chance. The proportions for question types Q1 to Q6 were (with standard errors) .37 (.03), .45 (.01), .49 (.02), .53 (.03), .51 (.04), and .50 (.03), respectively. Only Q1 differed significantly from chance, with subjects systematically getting the questions wrong. The mean d', PPSUM, and PPPROD was .06 (SE=.04), .53 (SE=.02), and .53 (SE=.03), respectively. None of these measures were significantly different from chance.

Discussion

The aim of Experiment Six was to investigate implicit learning in the Reber task using procedures similar to those used in Experiments One to Five. Accordingly, in Experiment Six subjects were probed with two tests of knowledge, free recall and the SLD test. Further, Experiment Six introduced a group exposed to both grammatical and nongrammatical exemplars during learning (group POSNEG) in addition to the standard group exposed to only grammatical exemplars (group POS). The results of Experiment Six addressed two main questions: First, can performance be accounted for by either test of knowledge? Second, is the type or quantity of learning affected by the presence of nongrammatical exemplars?

Three results pointed to a correspondence between classification performance and ability to answer the SLD test. First, there was a significant correlation between classification performance and correct responses on the SLD test. Second, subjects' d's for the classification and SLD tasks were not significantly different. In fact, d's were marginally higher for the SLD rather than classification task; future research should investigate the reliability of this effect. And third, there was a close match between classification performance and predicted performance based on answers to the SLD test. To calculate a predicted performance, a model was constructed in which subjects classified on the basis of a weighted sum of the "legitimacy" and "illegitimacy" of each exemplar (the SUM for that exemplar), as determined from the answers to the SLD test. According to the model, subjects responded "grammatical" or "nongrammatical" depending on whether the SUM for the exemplar fell above or below the mean SUM. Within each subject, the SUM for each exemplar was found to predict the tendency of the subject to respond

"grammatical", even when the actual grammatical status of the exemplar was partialed out. This result suggests that the model at least partly reflects how subjects applied their knowledge to the classification task. Further, there was no significant difference between classification performance and predicted classification performance; the drop in classification performance of POSNEG as compared to POS was matched by an identical drop in predicted classification performance. Thus, there is evidence that the SLD test can access the knowledge underlying classification performance with about the same degree of sensitivity as classification performance.

However, two results seem inconsistent with the conclusion that the SLD test and classification performance accessed the same knowledge base: The lack of (within group) correlation between actual and predicted classification performance across subjects, and the small size of the correlation between SUM and the tendency to respond grammatical within each subject (.17, on average). In fact, the first result, the lack of correlation across subjects between predicted and actual classification performance, is very suprising given very close matching of average values across groups of different levels of knowledge (POS and POSNEG). This situation is explicable if different cues (SLD stems and complete strings) accessed the same knowledge base, but did so with a highly variable efficiency. Thus, a stable estimate of the knowledge elicited by a task can only be obtained by averaging over a number of subjects. The small correlation between different tests within each subject can therefore only be detected by very powerful tests (powerful enough to detect a correlation of approximately .17). Also, the slightly greater correlation with classification performance of

number of correct SLD responses rather than of PPSUM would be expected: PPSUM was calculated using only 37% of the total number of SLD responses, and therefore provided a less reliable measure of the subject's knowledge than total number of correct SLD responses. This explanation is consistent with the finding that when d' for the SLD task was calculated using only the questions employed in deriving PPSUM, rather than all the SLD questions, the correlation between d' on the SLD and classification tasks was reduced to zero. In sum, the evidence for the psychological validity of the additive transformation used to derive the predicted classification performance is provided mainly by the close matching of average values of predicted and actual classification performance across groups. Further research is needed to determine the robustness of this result.

The variability in efficiency with which different cues access the knowledge base in Experiment Six can be contrasted, first, with the case of identical cues in Experiment Six and, second, with the case of different cues in the previous experiments. First, in Experiment Six, identical cues accessed the knowledge base with a much less variable efficiency than different cues. This is indexed by the correlation between probability correct on successive tests of the same exemplar on the classification task, r_t=.55 (also note the correspondence Dulany et al. (1984) found between their scoring task and classification performance based on identical exemplars, r=.83). The difference between this r_t (.55) and the r_s between PPSUM and classification (-.06) was significant, z=3.68, p<.0005. Second, for findings reported in Chapter Three (Experiments Four and Five) but not for Experiment Six, the performance on different tests was correlated. Considering those

tasks where learning was obtained and classification performance was not inflated by chance factors: For Experiment Four, single sentence task, the correlations between the knowledge measures varied between .60 and .89; in Experiment Five, correlations varied between .56 and .67. This difference in dependence between different knowledge measures across the experiments could be explained in terms of different knowledge types underlying performance in the different experiments. In fact, this second finding in conjunction with the first is closely analogous to data that was obtained by Hayman & Tulving (1989) in a different domain and which was used by them to argue for the existence of different memory systems. While both the first and second findings obtained here can perhaps be more simply explained in terms of a single memory system obeying established principles, any potential evidence for distinct implicit and explicit knowledge systems needs to be considered carefully. The "different systems" argument of Hayman and Tulving (1989) will first be considered, and then the "single system" counter-argument as applied to the current data will be discussed.

The difference in the variability with which same and different cues access a common knowledge base in the current paradigm is closely analogous to the difference in variability that same and different cues access a common knowledge base in the primed fragment completion paradigm used by Hayman and Tulving (1989). In the fragment completion paradigm subjects first study a list of words (e.g., AARDVARK), and are then shown fragments of words (e.g., -AR-VA--), and the subjects are asked to complete the fragment with the first word that comes to mind. The fragments of words that have been studied are more likely to be completed. Hayman and Tulving

(1989) found that successive tests of the same fragment were highly stochastically dependent, but successive tests of different fragments of the same word were virtually stochastically independent. They suggested that these results could be explained if fragment completion relied on a "traceless" procedural memory system. When the fragments were used to cue recall, successive tests were stochastically dependent, regardless of whether the fragments were same or different. Thus, they speculated that cued recall relied on a different type of memory system.

Hayman and Tulving's (1989) results with fragment completion are analogous to those obtained in Experiment Six (in that successive tests were correlated to a relatively small degree), and their results with cued recall are analogous to those obtained in Experiments Four and Five (in that successive tests were correlated to a relatively large degree). Hayman and Tulving's (1989) concern was to provide evidence for a distinction between memory systems based on the use or not of episodic traces. This is not the concern of this thesis, but this analogous difference in properties obtained in Experiments Four, Five, and Six could be used as evidence for a distinction between implicit and explicit types of knowledge, and also allow some connections to be drawn between the experimental domains investigated in this thesis and those investigated by Hayman and Tulving (1989). In terms of evidence for implicit and explicit knowledge types, the variability in the knowledge acquired in Experiment Six may be seen as a form of specificity of transfer, with consistent transfer occurring only for identical cues; hence, this knowledge might be regarded as implicit. The knowledge acquired in Experiments Four and Five, on the other hand, may be seen as relatively integrated and consistently applied in different

situations; hence, this knowledge may be regarded as explicit. The task used in Experiment Six appears to rely on a different type of learning than that employed in Experiments Three to Five for a further reason. Specifically, in Experiment Six, free recall was insufficient to access the knowledge base with the same degree of efficiency as classification performance; in Experiments Three to Five, free recall was sufficient. This failure of transfer is also suggestive evidence for an implicit knowledge base underlying classification performance.

A simpler explanation of the characteristics of Experiments Four, Five, and Six than that in terms of different knowledge types is that the same type of knowledge is involved in all the experiments, in that the same general principles of storage and retrieval apply, but that simply more information is stored in the case of Experiment Six than in Experiments Four and Five. The greater amount of information involved in Experiment Six is indicated by the length of the SLD test as compared to the structured test used in Experiments Four and Five; or by the number of exemplars used in Experiment Six as compared to Experiments Four and Five. Many theories of memory, involving a single system, would predict that the more facts there are connected to a topic, the less reliably any one of them would be retrieved; see, e.g., Shiffrin (1970), or, more recently, J. R. Anderson (1983). Because of the amount of information stored in the case of Experiment Six, but not Experiments Four and Five, it is relatively unlikely that different cues will retrieve exactly the same information; hence, different tests of the same knowledge do not correlate to a substantial degree. Although different information is retrieved by different cues, it will, on the average, be just as accurate; hence, the close

matching of PPSUM and classification performance.

Why should identical cues (specifically, whole exemplars) retrieve information more consistently than different cues (specifically, part exemplars as compared to whole exemplars) in terms of a single type of knowledge? This question is important because the linear transformation used to derive PPSUM appears to provide a model of classification performance, but it does not predict the consistency obtained with successive classifications of the same exemplar. One explanation is that the right-to-left context of a letter in an exemplar provides a cue in addition to the left-to-right context used to derive PPSUM; this additional cue helps to constrain retrieval and maintain consistency. It should be noted that this additional cue does not aid the retrieval of any more accurate information; it simply results in greater consistency in retrieving the same information. Another explanation is that the first classification provides a learning trial for the second; a similar phenomenon has been found by Broadbent and Broadbent (1977) for the recall of a word after it has been presented for recognition. Of course, there was no feedback during classification in Experiment Six, so incorrect responses would be reinforced as much as correct responses. Consistently, in the artificial grammar learning task, Mathews et al. (1989) found that classification without feedback (and without initial learning of the correct mapping for the letters used) resulted in delayed learning when feedback was later introduced.

The difference in free recall between Experiments Four and Five, on the one hand, and Experiment Six, on the other, can also be simply explained in terms of a single type of knowledge. Note that classification involves recognition, or not, of the presented

associations. Given that the set of associations to be recalled in Experiment Six is considerably larger than in Experiments Four and Five, the difference between recall and classification performance in Experiment Six but not Experiments Four and Five may be simply because the difference between recall and recognition diminishes as the set of items to be remembered is made smaller (Davis, Sutherland, & Judd, 1961). In sum, although the findings with respect to the variability with which cues access the knowledge base, and the inadequacy with which free recall accesses the knowledge base in Experiment Six, are suggestive evidence of different knowledge types, they can be simply re-explained in terms of a single knowledge type. From the point of view of the thesis, however, it is important that actual cases of implicit learning are not prematurely rejected. It is clear that with the suggestive evidence available the hypothesis of implicit learning on the artificial grammar learning task should be tentatively proposed, and further evidence sought.

In general, implicit knowledge of a domain can be characterized by the "conceptual task" that specifically elicits it. That is, by considering those tasks that empirically do elicit the knowledge, and those that do not, it should be possible to abstract a "conceptual task" that defines what it is that the implicit knowledge allows the subject to do. If the knowledge base underlying classification in artificial grammar learning is to be regarded as implicit, what is the conceptual task that specifically elicits the knowledge? Because the knowledge base can be adequately accessed by the SLD test (on the average), the conceptual task needs to encompass the SLD test. Thus, the conceptual task could be redefined to be the recognition of well-formedness of exemplars or

elements of exemplars (in isolation or not). That subjects can recognize the well-formedness of elements of exemplars in isolation is shown most strikingly by the high levels of performance on the SLD test for stems of length zero and one (see Table 14), where elements are maximally isolated. It is also apparant from Table 14 that increasing the length of the stem does not benefit SLD performance.

The presence of nongrammatical exemplars interfered with both performance and free recall. Subjects reported that they found it hard to remember what appeared in black and what in red, and so they may have confused correct and incorrect information. Did this interfere with both implicit and explicit learning? If free recall is taken as a measure of explicit learning, and performance as a measure of implicit learning, then the presence of contrast did not appear to affect implicit or explicit learning differentially: The group by test type (performance versus free recall) interaction was not significant. Thus, Experiment Six did not provide evidence that the presence of nonexemplars inhibited implicit learning in Experiments One to Five. However, informing subjects of the relevance of the distinction between black and red before rather than after learning may allow subjects to distinguish correct and incorrect information explicitly and so inhibit implicit learning.

In summary, the findings of Experiments One to Five were confirmed in that classification knowledge did transfer to recognizing the relevance of elements of exemplars not presented in the context of particular exemplars. Can the now rather broad conceptual task defining subjects' abilities in artificial grammar learning be regarded as involving implicit knowledge? The results provided suggestive evidence in that there was evidence for a

different type of knowledge in Experiment Six as compared to Experiments Four and Five (the variability with which different cues accessed the knowledge base was different in Experiment Six as compared to Experiments Four and Five), and there was some evidence for the specificity of transfer (the knowledge did not transfer to free recall in Experiment Six, but it did in Experiments Three to Five; further, the variability mentioned above may be seen as a form of specificity of transfer). However, failing to transfer to free recall is only weak evidence of specificity, as free recall is often regarded as an insensitive test (e.g., Nelson, 1978; Tulving, 1983). The argument that SLD and classification, on the one hand, and free recall, on the other, rely on distinct knowledge bases clearly needs to be strengthened. If they do tap distinct knowledge bases, then it should be possible to influence free recall but not classification and the SLD test, or vice versa. Experiment Seven probed this possibility.

EXPERIMENT SEVEN:

Effect of a concurrent task on artificial grammar learning

Introduction

The aim of Experiment Seven was to investigate the possibility of distinct implicit and explicit types of knowledge associated with artificial grammar learning by using a dual task methodology. Initially, the reasons for adopting such a methodology are described, and then appropriate measures of performance on the additional task are discussed.

Experiment Six produced some evidence that a different type of knowledge was formed with artificial grammar learning than that formed with the tasks used in Experiments Three to Five. First, in contrast to Experiments Three to Five, the classification knowledge did not transfer to an immediate free recall test; this result suggests that the knowledge underlying classification performance with artificial grammars might be implicit. Second, in contrast to Experiments Four and Five, the structured transfer test (the SLD task) did not correlate with classification performance to any substantial extent. There was stochastic dependence between successive tests of the knowledge base only when identical cues were used (cf. Hayman & Tulving, 1989). This last result can be seen as a form of transfer specificity, with consistent transfer dependent on identical cueing. It was suggested in Experiment Six that the SLD and classification tasks might tap the same implicit knowledge base with roughly equal sensitivity (because the SLD test predicted average classification performance), whereas free recall might tap only the explicit aspect of the subjects' knowledge. If implicit and explicit knowledge are of a different type, then it

should be possible to influence one but not the other with a suitable manipulation. This argument forms the basis of Experiment Seven.

A suitable manipulation is suggested by the results of Hayes (1987; and as reported in Broadbent, 1989); he found that random number generation (RNG) interfered with artificial grammar learning when subjects were given explicit search instructions, but did not interfere under standard memory instructions. One interesting interpretation of these results is that RNG interferes with the formation of explicit but not implicit knowledge; so under standard memory instructions, RNG might interfere with tests of explicit but not implicit knowledge. In Experiment Seven subjects were instructed to generate random numbers while they were memorizing grammatical exemplars. The aim was to see if this differentially interfered with subsequent free recall rather than with performance or responding on the SLD test.

In order to investigate the relation between RNG and artificial grammar learning, the sequences of random numbers generated by subjects were also subjected to analysis. To determine if different subjects assigned different priorities to the RNG and grammar learning tasks, a measure of randomness is required that is sensitive to the priority assigned to the RNG task. Subjects may not possess an accurate conception of randomness, and so a greater priority assigned to the RNG task might not lead to more random sequences. As far as Experiment Seven is concerned, the direction of the effect is not important. In fact, Graham and Evans (1977) reported that under dual as compared to single task conditions there was a decrease in randomness as indexed by an unspecified measure of first order dependency. Truijens, Trumbo, and Wagenaar (1976),

however, reported that the presence of a secondary task (pursuit tracking) increased randomness for the second to fourth order analyses when Wagenaar's (1970) phi measure of nonrandomness was used. Phi varies between -1 and +1, where 0 indicates randomness. For an n^{th} order analysis, phi is positive if there are too many repetitions between the current response and the response n places back, and negative if there are too many alternations. In general, subjects produce too many alternations (i.e. phi is negative) at least up to n=6 (Treisman & Faulkner, 1987; Wagenaar, 1970); Truijens et al. found that the presence of a secondary task reduced the number of alternations.

Phi seems to provide a performance measure of the RNG task that behaves in a known way as subjects allocate attention either to the RNG task or another one. Thus, analyses will be reported using first, second, and third order phi (phi1, phi2, and phi3). Additionally, analyses will be reported using first order entropy (H1, i.e. zero order dependency)[1]. It was not feasible to calculate higher order analyses of entropy, as such analyses are possible only if there are at least six observations per cell (Miller, 1955); this only obtained in Experiment Seven for first order entropy. The phi measure was specifically designed to evade the necessity for long response sequences, as the number of cells is effectively reduced to two for all category sizes and orders of analysis (see Wagenaar, 1970).

[1] For the calculation of phi, see Wagenaar (1970), and of first order entropy, see Attneave (1959).

Method

Subjects. The subjects were 12 paid volunteers from the Oxford University subject panel. No subject had participated in Experiment Six.

Materials and Apparatus. These were the same as for group POS of Experiment Six.

Procedure. Initially subjects were given a minutes practice on the RNG task. A metronome was set to give a click every two seconds; subjects were told to give a digit between 0 and 9 every time they heard a click, and to make the sequence of digits as random as possible. It was pointed out that each digit should on average occur equally often and be equally likely to follow any digit. Subjects were discouraged from repeating well learned sequences like 12345 or telephone numbers. It was suggested that they could imagine a hat containing 10 pieces of paper, one for each digit; every click they could imagine drawing a piece of paper out of the hat, reading it, and then replacing it. After a minutes practice on the RNG task, subjects were given the same instructions as for group POS of Experiment Six and they were also asked to generate random numbers. They were told that their priority was to make sure that they generated a digit for each click and to ensure that the sequence of digits was random.

After the learning phase, the procedure was identical to that used for group POS: Subjects classified grammatical and nongrammatical exemplars, were asked to give a complete account of how they did this, and finally were administered the SLD test.

Results

The proportion of exemplars correctly classified was .56 (SE=.02). This is significantly lower than the proportion classified by group POS (.65, SE=.02), $t(30)=3.28$, $p<.005$. The proportion of correct responses on the SLD test was .53 (SE=.02). This is also significantly lower than the proportion obtained by group POS (.64, SE=.02), $t(30)=4.04$, $p<.001$. The mean d' for the classification and SLD tasks was .29 (SE=.07) and .23 (SE=.06), respectively. They were not significantly different, $t<1$. The responses to the SLD test were used to derive a predicted classification performance, PPSUM, in the same way as for Experiment Six. The mean PPSUM was .59 (SE=.03). A 2 X 2 (Performance Type (predicted versus actual) by Group (Experiment Seven versus POS)) mixed model analysis of variance indicated only a significant effect of Group, $F(1,30)=5.65$, $p<.05$. That is, both predicted and actual performance was lower for Experiment Seven than for POS. For Experiment Seven, the Spearman's correlation between performance and proportion of responses correct on the SLD test was .35, between d' on the classification and SLD tasks was .31, and between performance and PPSUM, .49 (for $p<.05$ and 12 subjects, the critical values for r_s are .50, 1 tailed, and .59, 2 tailed).

The predicted performance based on free recall, PPFR, was .52 (SE=.01). A 2 X 2 (Performance Type (predicted versus actual) by Group (Experiment Seven versus POS)) mixed model analysis of variance indicated significant main effects of Group, $F(1,30)=12.46$, $p<.005$, and Performance Type, $F(1,30)=29.94$, $p<.001$, and a significant interaction, $F(1,30)=4.42$, $p<.05$. That is, both performance and PPFR ($t(30)=2.08$, $p<.05$) were lower for Experiment

Seven than for POS. Although PPFR was lower than performance for both groups (for Experiment Seven, $t(11)=2.71$, $p<.05$), the difference was less for Experiment Seven than for POS. The correlation between PPFR and performance was .57, and between PPFR and PPSUM, .49.

The means and correlations with performance for the measures of randomness under practice and dual task conditions are shown in Table 17. In the practice condition, each randomness measure is calculated from the 30 practice responses given by the subject. In the dual task condition, each randomness measure is calculated from 300 responses per subject. The correlations are the Spearman's correlations between each measure of randomness and classification performance on the grammar learning task. Note that in the practice condition, the correlation is with later artificial grammar learning; in the dual task condition, the correlation is with concurrent artificial grammar learning (for $p<.05$ and 12 subjects, the critical values for r_s are .50, 1 tailed, and .59, 2 tailed).

Table 17.

Performance on random number generation task.

	H1[1]	Phi1	Phi2	Phi3
Dual Task				
Mean	3.251	-.103	-.086	-.072
SD	.062	.008	.020	.026
r_s	.43	-.057	-.604	-.258
Practice				
Mean	3.144	-.109	-.098	-.109
SD	.161	.010	.022	.098
r_s	.45	.294	-.550	-.252

Discussion

The aim of Experiment Seven was to provide evidence for distinct implicit and explicit knowledge types in artificial grammar learning by using a concurrent RNG task. Based on Hayes (1987; and in Broadbent, 1989), it was expected that the RNG task would interfere with tests of explicit knowledge, like free recall, but not potential tests of implicit knowledge, like performance and SLD. In fact, the RNG task interfered with all three tests of knowledge.

Experiment Seven failed to replicate a key finding of Hayes (1987, Experiment 10); that is, the finding of a lack of interference of RNG on performance under standard memory instructions. There is one key procedural difference between

[1] The maximum H1, for a sequence with complete first order randomness, is $\log_2 10 = 3.32$. To provide an unbiased estimate of the true first order entropy, the H1s calculated for each subject, and presented in Table 1, should be increased by .216 in the practice condition, and .022 in the dual task condition (see Miller, 1955, for this "Miller-Madow correction").

Experiment Seven and Hayes that might account for these seeming incongruent findings. Hayes did not explicitly inform subjects of the priority to be given to the two tasks; in Experiment Seven, subjects were told to give the RNG task priority. Hayes might have subtly communicated different priorities to the implicitly and explicitly instructed groups. Alternatively, given ambiguous instructions, subjects may have adjusted priorities according to the perceived difficulty of the grammar learning task. When asked to test hypotheses under dual task conditions, subjects may have focussed on only one or two attributes of each string to comply with the experimenter's instructions. When asked to memorize seemingly arbitrary sets of letters, subjects may have attempted to take in all attributes of each string to comply with the experimenter's instructions. According to this view, only the explicitly instructed subjects fully engaged in the RNG task, and this interfered with all types of learning.

Do the present results shed light on the possibility of a dual task trade-off? If different subjects adopted different priorities, then the results of Truijens et al. (1976) suggest that there should be a positive correlation between phi and performance. In fact, there is a significant negative correlation between second order phi and performance. However, there are reliable individual differences in performance on the RNG task (Wolitzky & Spence, 1968), and these differences are related to other personality traits (Graham & Evans, 1977), and thus, possibly to performance in some types of learning tasks. These differences would be confounded with possible differences in priorities in determining the correlation between performance and phi under dual task conditions. The fact that practice phi correlates with later performance to the same

extent as dual task phi correlates with concurrent performance suggests that there is no extra source of covariance (brought about by priority differences between subjects) influencing the latter correlation; that is, subjects may have differed little in their priority levels and followed instructions to give maximum priority to the RNG task. But the current results are not strong enough to demonstrate the presence or absence of a dual task trade-off.

Thus, future research needs to determine the influence on artificial grammar learning and RNG of systematically manipulating priorities for implicitly and explicitly instructed subjects. This would be important in interpreting the results of Experiment Seven and Hayes (1987). If the measures of artificial grammar learning behave differently with priority changes for implicitly and explicitly instructed subjects, then the notion of distinct learning modes and knowledge types for artificial grammar learning may receive support; if they behave similarly, then the evidence for the notions presented in Hayes (1987; and in Broadbent, 1989) would be undermined.

EXPERIMENT EIGHT:

Modes of learning and types of knowledge

Introduction

The aim of Experiment Eight was to determine the influence of priority manipulation on artificial grammar learning under dual task conditions, and to determine the difference between single and dual task conditions. Initially, the reasons for investigating this issue are discussed, and then the logic of the methodology used is described.

Experiment Seven attempted to isolate distinct implicit and explicit knowledge types acquired during artificial grammar learning by employing a technique used by Hayes (1987; and in Broadbent, 1989). Hayes found that concurrent RNG interfered with the classification performance of explicitly but not implicitly instructed subjects. One interpretation of these results is that concurrent RNG interferes with the formation of explicit but not implicit knowledge. However, Experiment Seven found that concurrent RNG interfered with both free recall, an index of explicit knowledge, and classification performance and SLD responding, potential indices of implicit knowledge, under implicit instructions. There are two aspects to these results of Experiment Seven: They did not provide evidence for different knowledge types; but they also did not replicate Hayes' original result of an implicit learning mode that was resistant to dual task interference.

One key procedural difference between Experiment Seven and Hayes (1987; and as reported in Broadbent, 1989) was that in Experiment Seven subjects were clearly told to give the RNG task priority; Hayes did not report any priority instructions, nor did he

report any measures of RNG performance. It is possible that Hayes' subjects altered priorities in the implicit and explicit conditions as a function of different perceived task demands.

If subjects' priorities changed with implicit and explicit instructions, then the meaning of the difference in classification performance between implicitly and explicitly instructed subjects is, of course, ambiguous. It may be the case that implicitly and explicitly instructed subjects do learn in different modes. A proper test of this hypothesis with RNG requires that instructions and priorities are orthogonally manipulated. If priority has an effect on classification performance then results of Hayes (1987; in Broadbent, 1989) contain a potential artifact; but, even so, if the effect of priority is different for implicitly and explicitly instructed subjects, there is still good evidence for separate implicit and explicit learning modes. If priority has no effect on classification performance then the absence of dual task interference in the implicit rather than explicit conditions in Hayes' study cannot be based on a priority artifact. In this case, the comparison between single and dual task subjects allows a procedural replication both of Hayes and of Experiment Seven, to determine the replicability of both sets of results.

In addition to assessing the possibility of different learning modes, Experiment Eight also allows an assessment of the possibility of different knowledge types. If effects of priority or dual versus single task are greater for free recall than classification performance and SLD, there would be evidence for different implicit and explicit knowledge types. This is in principle separate to the question of different learning modes, but it would be expected on the current framework (that links learning

modes and knowledge types) that the evidence for different knowledge types would parallel that for different learning modes.

The effect of priority manipulations on the performance of two concurrent tasks defines a Performance Operating Characteristic (POC) for the two tasks. The concept of a POC was introduced by Norman and Bobrow (1975, 1976), and elaborated by Navon and Gopher (1979; 1980), as a means of determining whether two tasks require overlapping resources. Norman and Bobrow argued that if at least one of the tasks affected the other when priorities were changed, then it can be concluded that they share a resource. The logic of inferring resource limitations from POCs was questioned by Navon (1984). Navon distinguished alterants from resources as commodities: Commodities come in units, each of which is used exclusively by one process at a time; alterants are other things that do not have these properties but still affect performance (e.g., the activation of a memory node). Essentially, Navon argued that dual task interference may result not from the scarcity of any mental commodity, but from alterants (that is, cross talk) which may be specific to every task pair. Fitzgerald, Tattersall, and Broadbent (1988) pointed out that if both commodities and alterants are treated as resources, the POC can still be legitimately used to infer shared resources. Also, cross talk may be taken to be more likely between tasks requiring similar processes or representations. Thus, differential dual task interference for explicitly as compared to implicitly instructed subjects provides evidence that these subjects used different processes or representations to learn the grammar.

Method

Design. Experiment Eight employed a 2 X 3 (Instructions - implicit versus explicit - by Condition - single task versus dual task with grammar high priority versus dual task with grammar low priority) between subjects design, with an equal number of subjects in each of the six cells.

Subjects. The subjects were 60 paid volunteers aged between 18 and 45 from the Oxford University subject panel.

Materials and Apparatus. These were the same as those used in Experiment Six (group POS) and Experiment Seven.

Procedure. All subjects were exposed to the exemplars with the same displays used in Experiment Six (group POS), and then performed the classification, free recall, and SLD tasks in that order. Half the subjects performed five minutes of RNG alone before being exposed to the exemplars and half the subjects performed five minutes of RNG alone after the SLD task. The instructions for RNG were identical to those used in Experiment Seven. No further RNG was performed by the single task subjects. The remaining subjects performed RNG while being exposed to the exemplars, and were given priority instructions as to which task to emphasize. They were told to concentrate on the primary task, to perform it as best they could, and not to let the secondary task interfere; they were to attend to the secondary task to the extent that they were performing the primary task as well as they could.

Implicit subjects were given the same instructions used in Experiments Six and Seven; they were simply asked to memorize the exemplars. Explicit subjects were additionally asked to search for rules. The instructions were taken verbatim from Dulany et al.

(1984):

"The order of letters in each item of the set you are about to see is determined by a rather complex set of rules. The rules allow only certain letters to follow other letters. Since the task involves the memorization of a large number of complex strings of letters, it will be to your advantage if you can figure out what the rules are, which letters may follow other letters, and which ones may not. Such knowledge will certainly help you in learning and remembering the items."

Results

Classification performance.

Table 18 presents the proportions of items correctly classified for the six different groups. A 2 X 3 (Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) analysis of variance indicated only an effect of Condition, $F(2,54)=5.08$, $p<.01$. The F for Instruction was 1.67 and the F for the interaction was 1.02. The effect for condition was further analyzed by means of two orthogonal contrasts. The first contrast tested for a dual versus single task effect by comparing the single task groups with the average of the low and high priority groups, $F(1,54)=10.00$, $p<.01$. That is, dual rather than single task conditions interfered with classification performance. The second contrast compared the high and low priority groups, $F(1,54) < 1$. To summarize these important results, performing under dual rather than single task conditions interfered with classification performance, but there was no effect of priority, and no interaction of Condition with Instructions.

Table 18.

Classification performance.

	Condition:		
	Single task	High priority	Low priority
Instructions:			
Implicit	.69 (.10)	.66 (.06)	.62 (.05)
Explicit	.69 (.09)	.60 (.07)	.61 (.06)

Note Standard deviations appear in parentheses.

The mean d' for classification is shown in Table 19. The same pattern can be seen in Table 19 as in Table 18, and a 2 X 3 (Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) analysis of variance yielded similar results. The analysis indicated only an effect of Condition, $F(2,54)=5.69$, $p<.01$. The F for Instruction was 1.45 and the F for the interaction was 1.69.

Table 19.

d' for classification performance.

	Condition:		
	Single task	High priority	Low priority
Instructions:			
Implicit	.94 (.55)	.75 (.32)	.56 (.25)
Explicit	.91 (.50)	.45 (.36)	.51 (.31)

Note Standard deviations appear in parentheses.

The mean proportions of judgements that were incorrectly classified on both presentations (EE) or on only one (AV) are displayed in Table 20. A 2 X 2 X 3 (Error Type (EE versus AV) by

Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) mixed model analysis of variance indicated significant main effects of Error Type, $F(1,54)=22.80$, $p<.0001$, and of Condition, $F(2,54)=3.92$, $p<.05$. That is, subjects made more error-error than mixed error types, as found in Experiment Six, and by Dulany et al. (1984). The effect of condition was not analyzed further. Analysis indicated that the probability of correct classification on the second presentation was significantly greater when the subject was correct rather than incorrect on the first presentation (.77 compared to .43), $t(59)=14.42$, $p<.0001$. The strength of this stochastic dependence is indicated by $r_t=.53$, both for the average data and for the average r_t calculated for each subject separately.

Table 20.

Consistency of judgements

		Condition:		
		Single task	High priority	Low priority
Instructions:				
Implicit	EE	.16 (.08)	.19 (.07)	.21 (.06)
	AV	.15 (.03)	.15 (.06)	.16 (.06)
Explicit	EE	.18 (.10)	.25 (.08)	.24 (.06)
	AV	.14 (.04)	.20 (.04)	.15 (.04)

Note Standard deviations appear in parentheses.

The proportions correct for nongrammatical exemplars classified according to the position of violation are shown in Table 21. t tests indicated that all types differed significantly from chance, $ps<.001$, except for P5, $t=1.49$, and P6, $t=1.79$. See the

Appendix for a complete breakdown and analysis by groups.

Table 21.

Classification performance and position of violation for nongrammatical items

Position of violation:

PA	P1	P2	P3	P4	P5	P6
.71	.68	.60	.69	.64	.55	.41
(.24)	(.20)	(.20)	(.17)	(.20)	(.26)	(.39)

Note Standard deviations appear in parentheses.

Test of Sequential Letter Dependencies

The proportions of correct responses to the SLD test for the six groups are shown in Table 22. A 2 X 3 (Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) analysis of variance indicated only an effect of Condition, $F(2,54)=12.83$, $p<.0001$. The F for Instruction was 0.05 and the F for the interaction was 0.46. As for classification performance, the effect for condition was further analyzed by means of two orthogonal contrasts. The contrast for a dual versus single task effect was highly significant, $F(1,54)=25.40$, $p<.0001$. That is, dual rather than single task conditions interfered with SLD responding. The contrast for priority was nonsignificant, $F(1,54)=.15$. In summary, the results mirrored those for classification performance.

Table 22.

SLD test.

Condition:			
	Single task	High priority	Low priority
Instructions:			
Implicit	.65 (.10)	.58 (.08)	.57 (.05)
Explicit	.67 (.08)	.56 (.06)	.55 (.06)

Note Standard deviations appear in parentheses.

The proportions correct for SLD questions classified according to the position probed are shown in Table 23. See the Appendix for a complete breakdown and analysis by groups. *t* tests indicated that all question types differed significantly from chance, *p*s<.0005.

Table 23.

Proportion correct and position probed on SLD test

Position probed:					
Q1	Q2	Q3	Q4	Q5	Q6
.64	.61	.69	.59	.58	.58
(.20)	(.15)	(.12)	(.12)	(.10)	(.09)

Note Standard deviations appear in parentheses.

Classification performance and SLD

The Spearman's within-groups correlation across subjects between classification performance and proportion of questions correct on the SLD test was .35, *p*<.01.

Table 24 shows the mean d' for the SLD task for the six groups. The d's for the SLD and classification tasks were very similar (compare Tables 19 and 24). A 2 X 2 X 3 (Task (SLD versus

classification) by Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) mixed model analysis of variance indicated only a significant effect of Condition, $F(2,54)=14.74$, $p=.000$. There was also a marginal effect for Task by Condition, $F(2,54)=2.79$, $p=.07$. The Spearman's within-groups correlation between d' for the SLD and classification tasks was .44, $p<.01$.

Table 24.

SLD test.

	Condition:		
	Single task	High priority	Low priority
Instructions:			
Implicit	0.93 (0.41)	0.46 (0.28)	0.46 (0.24)
Explicit	1.05 (0.38)	0.36 (0.18)	0.48 (0.27)

Note Standard deviations appear in parentheses.

The SLD responses were transformed as in Experiment Six to produce a SUM for each exemplar for each subject. For each subject a correlation was computed over exemplars between SUM and tendency to respond "grammatical". These correlations were transformed according to Fisher's z. The mean value, converted back to a correlation, was .30. This is significantly different from zero, $t(59)=9.70$, $p<.0001$. The average correlation computed separately for grammatical and nongrammatical items was .23, significantly lower than the correlation over all exemplars at once, $t(59)=4.70$, $p<.0001$, but still significantly greater than zero, $t(59)=7.97$, $p<.0001$. As argued in Experiment Six, these significant results provide evidence both for the the psychological validity of the transformation used to derive SUM, and for the existence of at least

partially overlapping knowledge bases for the SLD and classification tasks.

The mean values for PPSUM for the six groups are displayed in Table 25. A 2 X 2 X 3 (Performance Type (PPSUM versus actual classification performance) by Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) mixed model analysis of variance indicated only a significant effect of Condition, $F(2,54)=7.81$, $p=.001$. There were also marginal effects for Instruction, $F(1,54)=3.72$, $p=.059$, and Performance Type, $F(1,54)=2.85$, $p=.097$. The effect for Condition was further analyzed by means of the two orthogonal contrasts. The contrast for a dual versus single task effect was highly significant, $F(1,54)=19.75$, $p<.001$. That is, dual rather than single task responding interfered with both knowledge measures. The contrast for priority was nonsignificant, $F(1,54)=.22$. The Spearman's within-groups correlation between PPSUM and classification performance was .24, $.05 < p < .10$. It is interesting to compare this correlation for implicit and explicit groups, on the grounds that if explicit rather than implicit knowledge is more consistently applied (see Experiment Six), then the correlations between different knowledge measures should be higher for an explicit rather than implicit knowledge base. In fact, the correlation was identically .23 for implicit and explicit groups.

Table 25.

PPSUM

	Condition:		
	Single task	High priority	Low priority
Instructions:			
Implicit	.71 (.10)	.62 (.08)	.60 (.08)
Explicit	.63 (.10)	.57 (.05)	.61 (.10)

Note Standard deviations appear in parentheses.

Classification performance and free recall

The mean values of PPFR for the six groups are displayed in Table 26. A 2 X 2 X 3 (Test Type (PPFR versus actual classification performance) by Instruction (implicit versus explicit) by Condition (single task versus grammar learning high priority versus grammar learning low priority)) mixed model analysis of variance indicated significant effects of Test Type, $F(1,54)=160.83$, $p<.0001$, and of Condition, $F(2,54)=7.81$, $p=.001$. That is, PPFR substantially underpredicted actual classification performance. The effect of condition was analyzed by orthogonal contrasts. The contrast for single versus dual task was highly significant, $F(1,54)=13.25$, $p<.005$. That is, dual rather than single task conditions interfered with both knowledge measures. The contrast for priority was nonsignificant, $F(1,54)=.02$. The Spearman's within-groups correlation between PPFR and classification performance was .32, $p<.05$, and between PPFR and PPSUM was .50, $p<.01$. For implicit subjects only, these correlations were .26 and .51, respectively; for explicit subjects only, they were .35 and .45, respectively. Neither difference between the groups approaches significance, $ps>.10$.

Table 26.

PPFR

	Condition:		
	Single task	High priority	Low priority
Instructions:			
Implicit	.57 (.06)	.53 (.02)	.51 (.02)
Explicit	.56 (.07)	.51 (.02)	.54 (.04)

Note Standard deviations appear in parentheses.

Random number generation

The measures of randomness used in Experiment Seven were again used in Experiment Eight; that is, first order entropy (H1), and first, second, and third order phi (phi1, phi2, and phi3). H1 was corrected for each subject according to the number of digits produced by that subject to provide an unbiased estimate of the true first order entropy (the Miller-Madow correction, as described in Miller, 1955). In addition to these measures, several others were used, recommended by Baddeley (1966), in an attempt to cover as many aspects as possible of the subjects' responses. These measures were the number of arithmetic sequences (NAS), the proportion of digrams repeated at least once (PRD), and the number of digits produced (ND). NAS refers to the number of occasions that successive digits differed by exactly one (divided by the number of digrams, to give a proportion); for example, if a five was followed by a six, or an eight by a seven. The means of these measures for the six groups under single task conditions are shown in Table 27. The means for the four dual task groups under dual task conditions are shown in Table 28. The average numbers of digits generated under single and dual task conditions were 141 and 269, respectively. Waganaar

(1970) indicated that phi is slightly biased in a way that depends on the number of digits generated, though an exact correction was not available. Thus, so that a fair comparison could be made between dual and single task conditions, the phi under dual task conditions was calculated for each subject using the first ND digits produced under single task conditions by that subject.

Table 27.

RNG under single task conditions

	Measure:						
Group:	H1	Phi1	Phi2	Phi3	NAS	PRD	ND
Implicit							
Single	3.310	-.103	-.099	-.080	.173	.822	138
	(.065)	(.007)	(.010)	(.013)	(.041)	(.072)	(9)
High Pri	3.277	-.096	-.088	-.074	.211	.813	143
	(.097)	(.023)	(.013)	(.020)	(.076)	(.068)	(2)
Low Pri	3.329	-.091	-.089	-.079	.194	.800	143
	(.016)	(.022)	(.020)	(.016)	(.071)	(.039)	(2)
Explicit							
Single	3.320	-.099	-.090	-.082	.187	.809	142
	(.047)	(.015)	(.021)	(.019)	(.057)	(.054)	(3)
High Pri	3.328	-.082	-.073	-.078	.214	.781	139
	(.058)	(.036)	(.021)	(.014)	(.066)	(.058)	(7)
Low Pri	3.331	-.096	-.094	-.083	.213	.801	140
	(.037)	(.020)	(.011)	(.015)	(.067)	(.040)	(17)

Note "Single" refers to those subjects who performed artificial grammar learning under single task conditions; "High Pri" refers to those subjects who gave artificial grammar learning high priority; "Low Pri" refers to those subjects who gave artificial grammar learning low priority.

Standard deviations appear in parentheses.

Table 28.

RNG under dual task conditions

	Measure:						
Group:	H1	Phi1	Phi2	Phi3	NAS	PRD	ND
Implicit							
High Pri	3.202	-.097	-.071	-.039	.280	.920	259
	(.110)	(.020)	(.020)	(.034)	(.083)	(.050)	(52)
Low Pri	3.274	-.097	-.075	-.067	.251	.935	287
	(.055)	(.015)	(.024)	(.025)	(.065)	(.018)	(5)
Explicit							
High Pri	3.272	-.072	-.070	-.063	.248	.925	252
	(.049)	(.048)	(.027)	(.026)	(.048)	(.026)	(38)
Low Pri	3.289	-.095	-.085	-.061	.287	.930	280
	(.045)	(.014)	(.019)	(.015)	(.093)	(.015)	(24)

Note "High Pri" refers to those subjects who gave artificial grammar learning high priority; "Low Pri" refers to those subjects who gave artificial grammar learning low priority.
Standard deviations appear in parentheses.

In order to control partially Type I error-rate, multivariate analyses were first conducted with the measures of randomness; only effects that were multivariately significant or marginally significant were considered further. Two multivariate analyses were conducted. First, a Hotellings between single and dual task conditions with six measures of randomness (ND was not used, as a different number of digits was requested in the two conditions); and, second, a MANOVA using the seven measures under dual task conditions only, with factors of Priority (high and low), Instruction (implicit versus explicit), and the Priority by

Instruction interaction.

The Hotellings T^2 comparing single and dual task conditions was 363.81, $F(6,31)=52.21$, $p<.0001$. Four of the six measures were univariately significant: Phi2, $F(1,36)=13.53$, $p<.001$; Phi3, $F(1,36)=22.32$, $p<.0001$; NAS, $F(1,36)=33.54$, $p<.0001$; and PRD, $F(1,36)=188.65$, $p<.0001$. The two nonsignificant measures were H1, $F(1,36)=2.55$, and Phi1, $F(1,36)=0.07$. In sum, subjects under dual rather than single task conditions produced relatively more repetitions than alternations at the second and third order of dependency, produced more arithmetic sequences, and used the same digrams more often.

Considering now only data collected under dual task conditions, the Hotellings T^2 for Priority was 17.03, $F(7,30)=2.03$, $p=.084$. Two measures were univariately significant: H1, $F(1,36)=4.06$, $p=.051$; and ND, $F(1,36)=6.54$, $p<.05$. The nonsignificant measures were: Phi1, $F(1,36)=1.66$; Phi2, $F(1,36)=1.52$; Phi3, $F(1,36)=2.49$; NAS, $F(1,36)=0.05$; and PRD, $F(1,36)=1.18$. That is, when subjects were asked to give the RNG task greatest priority, they produced more digits and distributed them over the 10 response categories more evenly. The Hotellings T^2 for Instructions was 14.18, $F(7,30)=1.69$ (and there were no univariately significant results), and the Hotellings T^2 for the interaction Priority by Instructions was 14.86, $F(7,30)=1.77$ (and there were no univariately significant results).

Grammar learning and RNG

So far the analyses of the effect of priority on artificial grammar learning and RNG task performance have been conducted on each task separately. A trade-off between the two tasks might be best detected by an analysis that takes into account both tasks

simultaneously. A correlation between classification and RNG performance under dual task conditions would in part reflect any mutual interference between the two tasks, but it would also reflect any pre-existing personality or ability factors that influenced both tasks. However, such factors (except for a general time-sharing ability) would also influence performance of either of the tasks singly. Thus, if single task RNG performance were to be partialled out of the correlation between the two tasks under dual task conditions, factors affecting both tasks singly should be accounted for. Any trade-off between the two tasks might then appear in the partial correlation between grammar learning and RNG performance. Table 29 presents the partial correlations between measures of randomness and measures of grammar learning, both under dual task conditions, with the measure of randomness under single task conditions partialled out of all correlations using the same measure under dual task conditions. Out of 21 correlations, one achieved significance at the .05 level.

Table 29.

Partial correlations between grammar learning and RNG measures.

	RNG:						
Grammar:	H1	Phi1	Phi2	Phi3	NAS	PRD	ND
Classification	-.11	-.08	.21	.09	-.10	-.05	-.11
No. of correct SLD responses	-.03	.05	.29	-.01	-.04	.04	.03
PPFR	-.16	.00	.21	.24	-.05	-.26	-.31[1]

[1] $p < .05$.

Discussion

The aim of Experiment Eight was to investigate the possibility of different modes of learning and different types of knowledge in artificial grammar learning by systematically exploring the influence of concurrent random number generation. Experiment Eight provided data relevant to the influence of different task priorities under dual task conditions, and to the influence of performing under dual versus single task conditions. Experiment Eight also allowed an attempted replication of the important findings of Experiments Six (and Seven); this last issue will be dealt with first.

As in Experiment Six, the results of Experiment Eight indicated a correspondence between classification performance and ability to answer the SLD test. There was a significant correlation between classification performance and correct responses on the SLD test. The d's for the two tasks were similar. Further, there was a close match between classification performance and predicted performance based on answers to the SLD test, using a linear transformation. The psychological validity of the transformation was supported by the significant correlation across exemplars between the SUM for each exemplar and the tendency of the subject to respond "grammatical", and, most importantly, by the close matching of average PPSUM and classification performance across groups with different levels of classification performance. The close matching of classification performance and PPSUM across groups provides evidence that the SLD test accessed the knowledge base underlying classification performance with the same sensitivity as classification performance. As Experiment Eight used a different

manipulation than Experiment Six to influence classification performance, the close matching of classification performance and PPSUM in Experiment Eight considerably strengthens the case for both the psychological validity of the linear transformation and the sensitivity of the SLD test to the classification knowledge base.

As in Experiment Six, Experiment Eight provided suggestive evidence that the knowledge acquired during artificial grammar learning is of a different kind to that acquired in Experiments Three to Five. First, the knowledge underlying classification performance only inadequately transferred to an immediate free recall test. Second, the correlations between different tests of the knowledge base were considerably lower than those observed in Experiments Four and Five, and provided evidence for the variability in efficiency with which different cues access the knowledge. The knowledge consistently transferred only for identical cues. As noted in Experiment Six, however, both these sets of findings can be explained in terms of a single type of knowledge. In order to investigate the possibility of different knowledge types further, Experiment Eight explored the effect of a dual task on the different knowledge measures.

No measure of artificial grammar learning - neither classification performance, SLD responding, nor free recall - showed an effect of priority under dual task conditions. This was the case even though the priority manipulation was effective in changing subjects' RNG performance; specifically, when the RNG task was given high rather than low priority, subjects produced more digits and distributed them over the 10 response categories more evenly. The effect of priority on only one of the tasks was confirmed by the absence of significant correlations between dual task grammar

learning and RNG performance with single task RNG performance partialled out; a trade-off would be detected by this measure only if both tasks were affected simultaneously. Thus, the apparant decrease in resources applied to RNG by subjects asked to give RNG low rather than high priority was not matched by an increase in resources effectively applied to artificial grammar learning. One possibility is that although subjects attempted to apply more resources to artificial grammar learning, they simply did not know how to do so effectively. This possibility is consistent with the finding, discussed below, that explicit rather than implicit instructions had no impact on subjects' performance: Knowing that there are rules to be found, and attempting to find them, did not benefit effective acquisition of the rules.

In terms of interpreting the results of Hayes (1987; and in Broadbent, 1989), the lack of effect of priority on artificial grammar learning, and the lack of an interaction between priority and implicit versus explicit instructions imply that Hayes' results were not subject to a priority artifact. That is, the detriment to artificial grammar learning produced by RNG under explicit but not implicit instructions in Hayes' study could not have been due to subjects in the different instructional conditions assigning artificial grammar learning and RNG different priorities. However, this null result for implicit subjects obtained by Hayes was not replicated by Experiment Seven. Experiment Eight allowed a more powerful assessment than was available to Hayes of the effect of dual task conditions on implicit and explicit subjects. In fact, consistent with Experiment Seven, but not with Hayes' study, in Experiment Eight, all measures of artificial grammar learning were impaired by dual as opposed to single task conditions, under both

implicit and explicit instructions.

The data of Experiment Eight indicate that this discrepancy between Experiments Seven and Eight, on the one hand, and Hayes (1987; and in Broadbent, 1989), on the other, cannot reasonably be put down to a priority artifact in Hayes' study, but they do suggest the possible source of the discrepancy. The means from Experiment Eight can be used to estimate the population difference in classification performance between single and dual task conditions; this estimate is seven exemplars. Using the MS_e from Experiment Eight (the MS_e used by Hayes was not available), the procedure of Hayes (1987) can be analyzed for its power. Hayes (1987) ran three groups: explicit-single task, implicit-single task, and implicit-dual task, with 10 subjects in each group. The one-way ANOVA calculated by Hayes on these three groups has a power of .3; that is, it is more likely than not that Hayes' procedure will fail to detect the effect of the dual task on implicitly instructed subjects. A problem with low power designs is, of course, that a true population effect may be detected in only some groups. One may speculate that when Hayes (reported in Broadbent, 1989) later ran the explicit-dual task group, this simply happened to be one of the cases where the effect of single versus dual task conditions emerged.

The presence of a dual task effect in the absence of a priority effect on grammar learning suggests that there may be some resource that is required for grammar learning and that is applied to RNG in an all-or-none way. One possibility is that the RNG task occupies the articulatory loop (Baddeley, 1986) and thus interferes with the acoustic or articulatory encoding of the artificial grammar strings. Further evidence for the role of such nonvisual codes in

artificial grammar learning is provided by the transfer results of Reber (1969) and of Mathews et al. (1989), discussed in Chapter One. Both Reber (1969) and Mathews et al. (1989) found that artificial grammar learning was minimally impaired by a change in the specific letter set used to construct the strings, so long as it was possible for subjects to determine the mapping between letter sets. Thus, it is not necessary for subjects to be exposed to the same visual information during acquisition and testing in order to access the knowledge base.

Experiment Eight provided evidence that artificial grammar learning is predominantly learned in a single mode and involves a single knowledge base. An early result by Reber (1976) indicated that giving subjects explicit rule-search instructions, rather than implicit memory instructions, deteriorated classification performance. Although this suggested that subjects could approach artificial grammar learning in distinct implicit or explicit modes, Reber et al. (1980) and several more recent studies have failed to replicate Reber's original result (Dulany et al., 1984; Hayes, 1987; Mathews et al., 1989; see Chapter One). Experiment Eight provided data consistent with the latter studies: Asking subjects to approach the task in an explicit rather than implicit manner had no effect on the classification and SLD tasks; further, it did not lead to any greater explicit knowledge, as indexed by free recall and by the correlations between the different knowledge measures. Note that Experiment Eight could detect a population difference in classification performance between implicit and explicit subjects of the size found by Reber (1976) with a power greater than .98. Thus, the failure to replicate Reber can be accepted as valid with some confidence, especially in light of the consistency of this null

result with all the later studies. To summarize the evidence with respect to learning modes, both the absence of an interaction between implicit versus explicit instructions and dual task conditions, discussed previously, and the absence of a main effect of implicit versus explicit instructions are consistent with subjects approaching artificial grammar learning with a single learning mode.

The evidence regarding the number of knowledge types acquired during artificial grammar learning parallels the evidence regarding learning modes. All the measures of artificial grammar learning - classification performance, SLD responding, and free recall - were affected in the same way by the dual task and priority manipulations. This result is consistent with a single knowledge base that is accessible by the classification and SLD tasks to an equal extent, but only inadequately by free recall.

General Discussion

The aim of Chapter Four was to explore implicit concept formation using the artificial grammar learning task (Reber, 1967, 1989) and employing the knowledge elicitation techniques used in Chapters Two and Three. Three experiments were reported that compared classification performance with a structured knowledge test - the SLD test - and with free recall, over a range of experimental manipulations. Interest focussed on the extent to which classification knowledge transferred to the SLD task, whether distinct learning modes could be applied to artificial grammar learning, and the extent to which the knowledge underlying classification performance could be regarded as implicit. These issues are discussed in turn.

Classification performance and SLD

The three experiments reported in Chapter Four jointly indicated a positive relation between classification performance and number of correct SLD responses; the Spearman's correlation across subjects over all experiments was .42[1], $p<.001$. Similarly, the Spearman's correlation over all experiments between d' on the SLD and classification tasks was .52[1], $p<.001$. Thus, there is evidence that the knowledge tests tapped the same knowledge base. But did they do so equally efficiently? This question is important in formulating the conceptual task that defines what subjects can do after exposure to artificial grammars; defining the conceptual task is important in assessing the nature of any potential implicit knowledge acquired in artificial grammar learning.

Evidence for the classification and SLD tasks accessing the

[1]This includes both within group and between group covariation.

same knowledge base with equal efficiency is provided by the similar d's for the two tasks over all the experiments: 0.61 and 0.60, respectively. A linear transformation was applied to the SLD responses to determine a predicted classification performance for each subject given that subject's level of SLD knowledge. A linear transformation was applied partly because it was one of the simplest transformations that could be used and partly because there is evidence that linear transformations accurately predict human judgements, even if people are not using a strictly linear strategy (Slovic & Lichtenstein, 1971). The linear transformation involved creating a value, SUM, based on adding the subject's confidence ratings for each successive letter considered "grammatical" and subtracting the subject's confidence ratings for each successive letter considered "nongrammatical". Exemplars above the mean SUM were then classified as grammatical and those below as "nongrammatical" (giving PPSUM, the predicted performance based on SUM). Experiments Six, Seven, and Eight provided two types of evidence for the psychological validity of this transformation. First, for each subject the Spearman's correlation over exemplars between SUM and tendency to say "grammatical" was calculated. These correlations were transformed according to Fisher's z. The mean value (converted back to a correlation) was .24. This is significantly different from zero, $t(111)=11.04$, $p<.0001$. Secondly, and most importantly, the Spearman's correlation over the nine group means for all three experiments between PPSUM and classification performance was .91, $p<.01$, and not significantly different from 1.0.

If the transformation is psychologically valid, what evidence does it provide for the transfer of classification

knowledge to the SLD test? In all three experiments, PPSUM closely matched actual classification performance across a range of manipulations that affected classification performance. That is, the SLD test could access the classification knowledge base with a sensitivity equal to classification performance. Thus, the conceptual task that specifically elicits the knowledge acquired during artificial grammar learning needs to encompass the SLD test. The conceptual task that subjects are able to do could be defined to be simply the recognition of well-formedness of exemplars or elements of exemplars. The elements can be in isolation, as shown by the high levels of SLD responding for small stem lengths (see Tables 15 and 23).

This conceptual task differs from that implicitly defined by Hayes (1987) as underlying implicit concept formation: Hayes believed that implicit concept knowledge did not transfer to his VPA measure which also involves presenting elements of exemplars in isolation. (In fact, Experiment One, Chapter Two, showed that the knowledge underlying classification performance did transfer to the VPA measure.) Also, this conceptual task suggests that the overall sense of "grammaticality" of an exemplar is analyzable by the subject when directly probed, and that the subject is capable of formulating general rules that capture the perceived grammaticality of an exemplar. Thus, while classification knowledge may be seen as implicit in some sense (discussed below), it is not

> "implicit in our sense that [subjects] are not consciously aware of the aspects of the stimuli which lead them to their decision. (Reber & Allen, 1978, p. 218)"

Learning modes and knowledge bases

One type of evidence for distinct implicit and explicit

learning modes concerns comparing different learning strategies on the same task (see e.g., Hayes, 1987, and in Broadbent, 1989; Reber, 1976). Another type of evidence relies on comparing subjects' performance on different tasks - tasks that can be plausibly classified as relying on implicit or explicit knowledge (see, e.g., Abrams & Reber, 1988; Berry & Broadbent, 1988; Mathews et al., 1989). Experiments Six to Eight provided data relevant to the first type of evidence, and will be discussed in this section. The comparison between Experiments Six to Eight and Experiments Three to Five provided data relevant to the second type of evidence, and will be discussed in the next section.

What evidence is there for different learning modes on the artificial grammar task? Reber (1976) argued that subjects could approach the task with either an implicit or explicit learning mode; when subjects were asked to search for rules, their classification performance deteriorated compared to the performance of subjects simply asked to memorize the strings. While this early result was encouraging, there has been a consistent failure to replicate, both in Reber's laboratory (Reber et al., 1980) and in a number of other laboratories (e.g., Dulany et al., 1984; Hayes, 1987; Mathews et al., 1989). Experiment Eight provided data consistent with the latter studies; explicit rather than implicit instructions had no influence on classification performance. Further, if the explicit rather than implicit instructions had resulted in more explicit knowledge there should have been more free recall knowledge for explicit rather than implicit subjects; but there was not. Also, if more of the knowledge had been explicit, there should have been greater correlations for explicit rather than implicit subjects between the different knowledge measures, as they would all be

tapping a single data base that could be consistently applied across the measures; however, the correlations were almost identical for implicit and explicit subjects.

Hayes (1987; and in Broadbent, 1989) seemed to provide a way of reconciling these findings with the existence of two learning modes: The effect of explicit instructions might show itself only under dual task conditions. However, Experiment Eight demonstrated that the RNG task used by Hayes interfered equally with implicit and explicit subjects. Experiment Eight also indicated that low power probably prevented Hayes from detecting a dual task effect on implicit subjects. In sum, the evidence from both explicit versus implicit instructions and the use of dual tasks is consistent with subjects approaching artificial grammar learning with a single mode of learning.

The data of Experiments Six to Eight were also consistent with a single knowledge type underlying classification performance. The strong relation between classification performance and the SLD test was discussed in the last section. Further, the Spearman's correlation over the nine groups of Experiments Six to Eight between classification performance and PPFR was .75, $p<.05$, and between PPSUM and PPFR was .91, $p<.01$. That is, changes in the group mean of one knowledge measure was mirrored by similar changes in the mean of either of the other measures. Classification performance and PPSUM had very similar group means; PPFR consistently underpredicted classification performance. These results are consistent with a single knowledge base, tapped with equal sensitivity by classification performance and the SLD test, but only inadequately by free recall. This conclusion is also supported by results from Mathews et al. (1989). They found that yoked subjects, using the

free recall instructions from experimental subjects exposed to the grammar, always classified fewer correct items than the experimental subjects, consistent with the insensitivity of free recall to the knowledge base; also that an increase in classification performance by the experimental subjects was matched by an increase by yoked subjects, consistent with a single knowledge base underlying classification performance and free recall.

Is the knowledge implicit?

If subjects learn the artificial grammar task in a single mode and acquire a single type of knowledge, is the knowledge implicit? The classification knowledge did not transfer to an immediate free recall test. This might seem to be a particularly good reason for regarding the knowledge as implicit because there are other cases, such as Experiments Four and Five, where classification knowledge did adequately transfer to free recall. However, the suprising specificity of transfer on one task as compared to another can only be regarded as good evidence for implicit knowledge if the tasks are equivalent in other relevant respects. One important difference between artificial grammar learning and the tasks used in Experiments Four and Five is the number of associations that need to be stored. This factor alone can more simply explain the failure of free recall in this chapter but not the last without invoking different implicit and explicit knowledge types.

Classification can be regarded as in part a recognition task. The associations presented in an exemplar can be recognized or not, and then some rule needs to be applied (e.g., a linear combination of the information, like that used to derive PPSUM) to make an overall classification decision. Recognition is generally

more sensitive than free recall (see e.g., Tulving, 1983), but the difference between free recall and recognition can be abolished by using stimuli from a suitably small set size (Davis, Sutherland, & Judd, 1961). In Experiments Four and Five, the subject may usefully form one or two associations per exemplar (between sentences and category response). In Experiments Six to Eight, the subject may usefully form up to $(n^2+n)/2$ associations for a n letter exemplar (each letter with any other in the exemplar). Under these conditions, and especially since some of the associations may be low confidence, it is not suprising that not all the associations are retrieved in free recall in Experiments Six to Eight.

If subjects perform poorly in free recall simply because of the amount of information to be retrieved, then they should perform adequately in a Think Aloud situation, where the information retrieved to classify the exemplar can also be given immediately as a justification (assuming low confidence responses are given). Responses given in free recall in Experiments Six to Eight were of the same type as those reported by Reber (e.g., Reber & Allen, 1978) as elicited in a Think Aloud situation (see Chapter One). That is, subjects report that a letter can or cannot appear in a certain position, or that combinations of letters can or cannot occur. Reber (e.g., Reber & Allen, 1978) reported that subjects sometimes failed to provide justifications for their decisions. The reason for this must be that the subjects had low confidence because when subjects were forced to respond (with the scoring task of Dulany et al., 1984), with the same type of response that they would have given anyway, the subjects' responses were capable of predicting the subjects' classification performance. Thus, suitably instructed subjects should be able to give sufficient information in a Think

Aloud protocol by responding on each trial, even if the response is low confidence; conversely, subjects who do not give sufficient information were not suitably instructed.

Another reason why it might be suggested that the knowledge in Experiments Six to Eight was of a different type to the knowledge in experiments Three to Five is that there was a difference in the variability in efficiency with which different tests tapped the underlying knowledge base. Considering data from all three artificial grammar experiments, the within-groups Spearman's correlation across subjects between PPSUM and classification performance was .13, between PPFR and classification performance was .23, and between PPSUM and PPFR was .24. Across Experiments Three to Five, the within-groups Spearman's correlations across subjects between different measures varied between .56 and .89. However, once again, this difference across experiments need not be explained in terms of a different type of knowledge; it can be explained in terms of the relative amount of knowledge. Given that the amount of information stored in artificial grammar learning is relatively large, and assuming that retrieval is selective, then it is relatively unlikely that the same information will be retrieved on two occasions, especially if the cues are different. However, one would expect the information to be just as accurate, on average, when different cues are used, so long as a sufficient number of observations are made. It appears that for the tests used in Experiments Six to Eight, the number of observations collected on a single subject was insufficient and so the averaging also needed to occur over subjects.

The defining criterion of implicit learning is its specificity of transfer; Experiments Six to Eight found little

evidence of this for artificial grammar learning, except for the insensitive free recall task. However, there are other aspects of artificial grammar learning that appear interesting; for example, it appears to be learned in a relatively passive way. Thus, asking subjects to search for rules did not enhance learning the rules of the grammar; also, shifting resources away from random number generation, presumably towards artificial grammar learning, under dual task conditions did not improve grammar learning. Not all concept formation tasks are passive in this way. As described in Chapter One, Mathews et al. (1989) found that learning a biconditional rule (but not a finite state grammar) was impaired by incidental rather than intentional learning conditions. Similarly, Abrams and Reber (1988) found that psychiatric patients as compared to normal controls were impaired in learning a biconditional rule, but not in learning a finite-state grammar.

Note also that the process by which subjects reached a classification decision could be modelled by an additive combination of information from multiple cues, a process Hammond et al. (1987) regarded as forming the basis of intuitive cognition. There is certainly an interesting learning process to be investigated and modelled here. What, for example, is the learning rule by which subjects learn the associations incidentally, and how are the associations combined? The linear combination used to derive PPSUM provided a good first approximation and also provided a constraint for future, more complete learning models: Such models will have to produce PPSUM values close to their actual classification values, on average. Such models should also be able to produce variability between individual classification and PPSUM values.

To summarize, Chapter Four did not provide evidence for a

distinctively implicit type of knowledge. The different results obtained in experiments Three and Four can be explained in terms of a single underlying system. It is still possible, however, to distinguish at a suitable functional level of explanation between classification tasks where the underlying knowledge can or cannot be retrieved by an immediate free recall test. The "implicit" tasks would involve learning a large amount of information, perhaps combined in an additive way, and the "explicit" tasks would involve learning a small amount of information. The theoretical problem would be to specify the different characteristics of the tasks in terms of a single system of learning principles.

Chapter Five

Modelling implicit learning: Artificial grammar learning

Introduction

Reber has long argued that subjects can learn finite-state grammars in an implicit way (e.g., Reber, 1967, 1989). Reber's claim that this learning can occur incidentally and is not helped by knowledge of the existence of rules was confirmed in Chapter Four. Moreover, Chapter Four confirmed Reber's claim that subjects show poor free recall of the acquired knowledge even though they can accurately classify grammatical and nongrammatical strings. This was interpreted in Chapter Four as indicating that subjects utilized a large number of associations, and not a few general rules, in guiding their classification performance. Reber's claims that distinct implicit and explicit learning modes can be applied to artificial grammar learning, and that distinct implicit and explicit knowledge types are formed, were not confirmed in Chapter Four. Nonetheless, the processes by which large numbers of associations are stored and combined incidentally, and are then utilized in making classification decisions, remain intriguing. These learning processes may be regarded as "implicit" at an everyday level of explanation - incidental learning followed by a failure to transfer to a free recall task. The aim of Chapter Five was to explore how these "implicit" processes of artificial grammar learning could best be modelled.

The problem is to model learning without any apparant hypothesis testing. Rumelhart and McClelland (1986) suggested that

lawful behaviour and judgements may be produced by a Connectionist (or PDP) network in which rules are not explicitly represented but emerge from the way that interacting units are connected. The knowledge represented in the pattern of connection strengths can only be accessed by the specific inputs coded by the network and elicited in the form of the specific output responses of the network. There is no reason for the network to "know" how its performance is best summarized, or to "know" which inputs would elicit the most informative outputs for interpreting the nature of the knowledge stored in the weights. That is, if subjects' knowledge could be modelled by such a network, there is no reason to suppose that they possess sufficient metaknowledge to describe how the network works. Thus, this style of modelling would be consistent with the knowledge being "implicit" in the everyday sense described above: Subjects acquire many associations but find it difficult to summarize their knowledge. It would be neutral with respect to the more theoretical sense of a learning system distinct from explicit learning (one may, or may not, be able to model hypothesis testing with such a model, where the units represent hypotheses, or attributes of hypotheses).

A similar approach to modelling "implicit" learning is the memory array or exemplar approach (e.g., Estes, 1986; Hintzman, 1986; Medin & Schaffer, 1978). In this case, rules are also not explicitly represented but emerge from the way in which test items are compared to stored exemplars. As reviewed in Chapter One, Brooks (1978) argued for an exemplar account of artificial grammar learning, and McAndrews and Moscovitch (1985) found that at least some of subjects' knowledge of artificial grammars could be accounted for by knowledge of specific exemplars.

This chapter investigates the usefulness of both the Connectionist and exemplar approaches in modelling artificial grammar learning. That is, learning is modelled as the application of a learning rule to a net that results in an appropriate set of connection weights or as the appropriate storage and deployment of exemplars.

This chapter proceeds by initially describing the models used, and then describing the experimental data used to evaluate the models. Next, some relationships between the models are indicated, and then the models are evaluated against the empirical data. Finally, characteristics of the most successful model are explored in more detail.

Description of models

The two types of model considered, Connectionist models and the memory array models of Estes (1986; also, Medin & Schaffer, 1978) and of Hintzman (1986), are now described in turn.

Connectionist models

Connectionism is having an increasing impact on psychology; indeed, Massaro (1988) called it a revolution, and Schneider (1987) a "paradigm shift". For relevant overviews of the impact of Connectionism see Cognition, volume 28, Hinton (1987), Journal of Memory and Language, volume 27, McClelland and Rumelhart (1986), Rumelhart and McClelland (1986), and Smolensky (1987). The Connectionist ideas specifically relevant to the models conducted for this chapter are first discussed, and then the models themselves are described.

Connectionism attempts to model human learning and performance through the use of a large number of simple

computational elements, or units. The state of a unit can be specified simply by the level of its activation. The activation of input units is clamped so as to represent some aspect of the environment. The response to the input is given by output units, which are connected to the input units, and optionally to mediating hidden units, with weights. The activation of an output unit is some function (the output function) of the weighted activation of the input and hidden units connected to it. The simplest output function, and the one used for the models in this chapter, is

$$o_i = \sum_j w_{ij} a_j$$

where o_i is the activation of the ith output unit, w_{ij} is the weight from the jth input unit to the ith output unit, and a_j is the activation of the jth input unit. In order to investigate the performance of the simplest models, this chapter does not use models with hidden units.

The architecture of a network of units is specified by the connectivity between the units, i.e. which weights are allowed to be nonzero. For example, in a pattern associator there are no connections between the input units, and output units receive connections only from input units. With suitable weights, the pattern associator will be able to associate a number of input and output patterns with each other. The (feed forward) auto associator may be considered to be a pattern associator where the input and output units code the same pattern and the weights between corresponding input and output units are zero. This is equivalent to a single set of input units connected to each other (but no connections to the same unit from itself) and the task is to produce a predicted activation for each unit based on the weighted sum of the activation of the other units. The auto associator is

interesting from the point of view of artificial grammar learning because the task of forming suitable weights in an auto associator is similar to the task of subjects learning the grammar: That is, to establish the predictability of each letter from the other letters in a grammatical string.

Part of the appeal of Connectionist networks is that there is often no need to set the weights by hand in order to produce appropriate behaviour; the network can learn to "program itself" by the use of local learning rules. Two rules commonly used in networks without hidden units are the Hebb rule and the delta rule (Rumelhart & McClelland, 1986). These rules are discussed in turn. Initially, the rule is stated. Then, previous research suggesting the usefulnes of the rule to modelling human concept formation is discussed. Finally, the application of the rule to artificial grammar learning is considered.

(a) Hebb rule.

The Hebb rule is so called because it was clearly espoused by Hebb (1949), although a previous statement of it appeared in James (1890) as his "law of neural habit". The Hebb rule is that the increment in weight between two units is dependent on the correlation between the activations of the two units. A common version of this rule (see e.g. J. A. Anderson, 1983), and the simplest version, is for the weight to be incremented on each learning trial by an amount equal to the product of the activations of the two units. That is,

$$\Delta w_{i,j} = a_i a_j$$

where a_i and a_j are the activations of the ith and jth units, respectively[2]. Thus, the weight between units i and j will be increased only if both units are on (have positive activations)

during the learning trial.

J. A. Anderson (1983) argued that a Hebbian pattern associator could provide a powerful model for concept formation. He showed that such a model could qualitatively simulate the performance of subjects in Posner and Keele's (1968, 1970) classic experiments. In these experiments, patterns of random dots generated on an oscilloscope screen were denoted as "prototypes". Subjects were exposed to distortions of these prototypes and classified them (with feedback) into categories defined by the prototype. Depending on conditions, prototypes may later be better classified than actually exposed exemplars, and remembered for longer. Similarly, Anderson found that a Hebbian pattern associator, associating exemplars (distortions) with "names" of the categories also learned to classify prototypes better than exposed exemplars under suitable conditions: The associator averages out the distortions in different exemplars. The relevance of Posner and Keele's (1968) paradigm to artificial grammar learning is suggested by their attempt to elicit introspections from a group of subjects; some of these subjects, similarly to some subjects in Experiments Six to Eight, did not report using any rules.

[2]With this version of the Hebb rule, $w_{i,j}$ can increase without limit. This is not a problem for the use of the Hebb rule in this thesis, but some people might find a learning rule more natural if it led to stable $w_{i,j}$ after some period of exposure to an ergodic sequence. It seems to me that the following rule could be used for this purpose:

$$w_{i,j,n+1} = w_{i,j,n} + LR(a_i a_j - w_{i,j,n})$$

where $w_{i,j,n}$ is the weight between the ith and jth units on trial n, and LR is a small learning rate. $w_{i,j}$ will be stable over a long sequence of trials when it equals the expected value of $a_i a_j$ over trials. Note that the pattern of $w_{i,j}$ produced by this rule after a sufficiently large number of trials will equal the pattern produced by the simpler rule given in the text after a single iteration through all exemplars. Thus, the simpler rule is actually used in this thesis to model grammar learning.

In the case of modelling artificial grammar learning, the interest is not in associating exemplars with different categories, but in determining a "central tendency" of all the exemplars. As noted above, this is a task for an auto associator rather than a pattern associator. The weights matrix, **W**, of a Hebbian auto associator can be determined very simply. Let $\mathbf{a}_k$ be the column vector of activations for the kth pattern presented to a Hebbian auto associator, and $\mathbf{a}^T$ its transpose. Then after presentation of all m patterns,

$$\mathbf{W} = \sum_{k=1}^{m} \mathbf{a}_k \mathbf{a}^T.$$

Note that the i,jth element of **W** will be a measure of the extent to which units i and j are simultaneously active over all the patterns. That is, **W** constitutes an approximate sample covariance matrix for the units (J. A. Anderson, 1983). This observation enables the insights of the matrix algebra of correlation matrices - factor analysis - to be brought to bear on understanding the nature of the knowledge acquired by the Hebbian auto associator. The eigenvectors of a correlation matrix give the principal components of principal components analysis (PCA, i.e. factor analysis with constant communalities down the diagonal of the correlation matrix). Thus, the eigenvectors of **W** are analogous to the principal components of the matrix of correlations between the features coded by the units of **W**. That is, the principal eigenvectors of **W** will contain appreciable loadings for features that are mutually highly correlated. Extracting this underlying structure in the patterns may be regarded as a form of concept formation. Indeed, Child (1970), in his book on factor analysis, compares extracting factors

to a child forming concepts.

There are two differences with between **W** and the correlation matrix used by PCA that should be noted. First, PCA conventionally uses ones down the diagonal of a correlation matrix; **W** contains zeros down the diagonal, as no unit connects to itself. Changing the diagonal entries of a correlation matrix by a constant amount, c, does not change the eigenvectors in any way; it simply reduces each eigenvalue by c[1]. Thus, in this respect, the eigenvectors of **W** do not differ from principal components.

The second difference between **W** and the correlation matrix used by PCA is that the entries in **W** are strictly covariances only if the mean activation of each unit over learning trials is zero. The entries in **W** are correlations only with the further requirement that the standard deviations of the activations of each unit are one. These requirements are not easily met in a natural coding scheme. For example, a readily interpretable coding scheme is for a unit to have an activation of 1 if the feature it codes (e.g., a particular letter in a particular position) is present in an exemplar, and an activation of 0 (or -1) otherwise. The entries in **W** will then reflect not only the covariance between the activations of two units, but also their frequency of occurrence. Thus, a unit will be strongly represented in the dominant eigenvector not only if

[1]Consider a square matrix **A** with eigenvectors $\mathbf{v}_i$ and corresponding eigenvalues λ_i. That is,

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

Let **B** be a matrix obtained from **A** by subtracting s constant, c, from the leading diagonal of **A**. That is,

$$\mathbf{B} = \mathbf{A} - c\mathbf{I}.$$

Thus,

$$\begin{aligned} \mathbf{B}\mathbf{v}_i &= (\mathbf{A} - c\mathbf{I})\mathbf{v}_i \\ &= \mathbf{A}\mathbf{v}_i - c\mathbf{v}_i \\ &= (\lambda_i - c)\mathbf{v}_i \end{aligned}$$

Thus, $\mathbf{v}_i$ is an eigenvector of **B** with eigenvalue $(\lambda_i - c)$. QED.

its activation correlates highly with the activations of other units (as in PCA), but also if it has a high base rate of occurrence. This difference to PCA is desirable if the auto associator is to be sensitive to base rate effects; i.e. to both first order dependencies (as in PCA) and also zero order dependencies.

By analogy with PCA, if there are only a few eigenvectors of **W** with appreciable eigenvalues, these principal eigenvectors may be regarded as having extracted the "central tendencies" of the exposed exemplars. In this sense, the Hebbian auto associator may be regarded as having learned a "concept". But how might this knowledge of the concept actually be expressed?

Let the vector of predicted activations across all units for the kth input pattern be given by

$$\mathbf{p}_k = \mathbf{W}\mathbf{a}_k.$$

If the exemplars learned by a Hebbian auto associator are coded as orthogonal vectors, then the auto associator will be able to reproduce each input vector entirely (to a scalar multiple) without interference from the others. That is, the $\mathbf{a}_k$ will form the eigenvectors of **W**. In the artificial grammar learning task used in Experiments Six to Eight, the exemplars possessed a strong family resemblance structure defined by the finite state grammar. Any scheme for coding the exemplars that captured this family resemblance structure must represent the exemplars in a nonorthogonal way. Thus, with such a coding scheme, there is likely to be a dominant eigenvector that almost entirely captures the variance in all the exposed exemplars. Test strings that are highly "prototypical" - that is, close to the dominant eigenvector - will be little changed (to a scalar multiple) by multiplication by **W**. That is, the correlation between input and predicted activations as

given by

$$a_k \cdot p_k / |a_k||p_k|$$

where $|a_k|$ is the magnitude of a_k, will be close to one. If however, a test string is not close to the dominant eigenvector, the predicted vector will nonetheless be pulled towards it[1], and so the correlation between input and predicted vectors will be somewhat less than one. The variation in the correlation between input and predicted vectors can be used as a means for classifying test strings as "grammatical" or not.

To summarize, a Hebbian auto associator provides a measure of the central tendencies of a set of exemplars in terms of principal components analysis. Test strings can be classified according to how well they match the "principal components" of the studied exemplars, where the "principal components" have been modified by base rate effects.

(b) The delta rule.

Whereas the Hebb rule was developed as a plausible way in which neurons might learn (Hebb, 1949; James, 1890), the delta rule was developed as an optimal solution to a computational problem (see Hinton, 1987; Stone, 1986; and Widrow & Hoff, 1960). The delta rule is the procedure that will produce a set of weights that minimizes the following "error measure" of the difference between the desired output vectors and the actual output vectors produced by the network:

$$E = 1/2 \sum_{i,k} (o_{i,k} - d_{i,k})^2$$

where $o_{i,k}$ is the actual activation of output unit i in exemplar k and $d_{i,k}$ is the desired activation.

How should the weights be changed to produce the least mean

[1]Strictly, towards the eigenvector, with the largest eigenvalue, to which it is nonorthogonal.

square solution? Consider a plot of the error obtained against $w_{i,j}$ for a given i and j (and with the weights for all other i,j held constant). If $w_{i,j}$ is changed by a small amount in the direction opposite to the gradient, then the error will decrease (as long as the change is small enough); if the amount of change is proportional to the gradient, then with no error, no change will be produced. This procedure is known as gradient descent. That is, we want

$$\Delta w_{i,j} = - LR\delta E/\delta w_{i,j}$$

where LR is a suitably small constant. It can be shown (e.g., Hinton, 1987) that this implies the delta rule

$$\Delta w_{i,j} = \sum_k LR(d_{i,k} - o_{i,k})o_{i,k}$$

if the output of each unit is simply the weighted sum of its input. Here LR becomes the learning rate for the delta rule. The rule is often implemented (see Hinton, 1987), and is implemented in this thesis, by changing each $w_{i,j}$ after each pattern (instead of after presentation of all patterns). Thus, if the delta rule is followed, the network is guaranteed to move in the direction that minimizes the squared error; if there is a set of weights for which the error is zero, it can be approached with any desired degree of accuracy.

Understanding the properties of the delta rule may be helped by noting its correspondence to procedures in two other domains that psychologists are already familiar with. First, the delta rule is an iterative method of producing the standard regression coefficients for predicting each unit from the other units (see Stone, 1986), as may be expected from the fact that it produces the least mean square solution. An important property of the regression coefficient for a variable is that it does not simply reflect the correlation between the variable and the dependent variable, but rather the correlation between the variable and the dependent

variable with the variance due to all the other independent variables taken out. Second, the delta rule is essentially identical to the Rescorla-Wagner rule (Rescorla & Wagner, 1972) that has been very successful in summarizing a large body of animal learning data. In particular, it has been successful in accounting for the phenomena of blocking and overshadowing, in which different CSs compete to develop associative strength with a US. Thus, in contrast to the Hebb rule, the $w_{i,j}$ produced by the delta rule for the jth input unit will partly reflect how well the ith output unit is already predicted by other input units, just as for regression coefficients and conditioned associative strengths.

Recently it has been argued that the delta rule is relevant in understanding human concept formation (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; McClelland & Rumelhart, 1985, 1986; Medin & Edelson, 1988; Shanks, 1990). Estes et al., Gluck and Bower, Medin and Edelson, and Shanks contrasted a delta rule network with the exemplar model of Estes (1986) and Medin and Schaffer (1978) (described on page 204) in its ability to account for how categorization performance is affected by category base rates. These results may be relevant to modelling artificial grammar learning: In this situation, some letters in a grammatical string of letters occur in a given position more frequently than others. McClelland and Rumelhart more generally indicated how a delta rule auto associator can behave as if it has acquired concepts. The above studies are now discussed in more detail.

The studies by Estes et al. (1989), Gluck and Bower (1988), and Shanks (1990) used a task in which subjects classified patterns of four symptoms into one of two disease categories, one of which was more common than the other. Subjects were assessed on their

ability to classify complete symptom patterns (with feedback) and also on their ability to predict each disease given information on only one of the symptoms. The delta rule network used to model this task by all three studies consisted of four input units coding each symptom connected to a single output unit coding the disease category. Each unit had an activation of 1 if the disease or symptom was known to be present, and 0 otherwise. All three studies (except Shanks, Experiment Three) found that subjects' asymptotic classification data of complete symptom patterns were close to matching the normative Bayesian values. With the stimuli used in all three studies (except Shanks, Experiment Three), both the network model and the exemplar model (with no forgetting) predict average asymptotic matching, but they differ in their trial by trial predictions. At asymptote, the error produced by the delta rule network on any given trial will be small, and so the weights will not fluctuate much across trials. At asymptote, the predictions of the exemplar model will be changed each trial by the acquisition of a new exemplar, regardless of the error in its prediction. Thus, due to local sequence effects, the predictions of the exemplar model may deviate more widely than the network model from Bayesian matching on any given trial. Estes et al. found that subjects trial by trial performance could be better accounted for by the network rather than exemplar model, mainly because of the deviations from Bayesian matching predicted by the exemplar model. The stability of the delta rule's predictions at asymptote differentiates it from the Hebb rule as well as from the exemplar model.

Shanks (1990; Experiment Three) tested the network model further by devising a stimulus set for which the network model did not predict Bayesian matching, but rather base-rate neglect. In

this situation, subjects also showed base-rate neglect in their classification of complete symptom patterns.

Estes et al. (1989), Gluck and Bower (1988), and Shanks (1990) found that when subjects were asked to indicate the probability of a disease given information on only one symptom, they showed a degree of base rate neglect. This neglect was not well predicted by either the exemplar or network models. Estes et al. (1989) and Gluck and Bower (1988) did derive some degree of base rate neglect from the network model because the network model could not distinguish the absence of a feature from the absence of knowledge about a feature: Both situations were encoded as 0. Thus, the network model does not behave in a Bayesian way when there is absence of knowledge about a feature. However, actual subjects did distinguish the absence of knowledge about a feature from the absence of a feature. This indicates that the network model was used inappropriately in this situation. Instead of using 0 for both the absence of a feature and the absence of knowledge, a better coding scheme for this task might be -1 for the absence of a feature and 0 for lack of knowledge. The importance of how absent features are encoded is considered below.

Gluck and Bower (1988) tested another feature of the delta rule that distinguishes it from the exemplar model and also from the Hebb rule. Namely, if a symptom is redundant with another symptom in predicting a disease, then the weight for that symptom will be attenuated (compare blocking and overshadowing in the animal learning literature, and the reduction in regression weights for correlated variables in multiple regression). The data of Gluck & Bower (1988) supported this prediction of the delta rule: Subjects predicted a lower probability for a disease given a redundant

symptom rather than a nonredundant symptom.

Medin and Edelson (1988) presented results consistent with the delta rule as a model for concept formation. Their results also illustrate the importance of considering how absent features are encoded in determining the predictions of a delta rule model. Subjects classified symptom patterns into one of six diseases. A typical stimulus structure for two diseases would be symptom pattern "a,b" indicating disease "one" and "a,c" indicating disease "two"; disease "one" occurred three times as frequently as disease "two". If subjects were using base rate information appropriately, given symptom "a" they should respond "one" more often then "two". This is indeed what subjects did, and it is also the behaviour predicted by Gluck & Bowers' (1988) network model. If subjects were using base rate information appropriately, they should also classify the pattern "b,c" as "one" more often than as "two". In fact, subjects classified reliably in the other direction. Gluck & Bower's (1988) network, with a separate node for each disease and the absence of a feature coded as 0, would not show this inverse base rate effect. Markman (1989) points out that if the absence of a feature or disease is coded as -1 (and not as 0), a delta rule network can learn the inverse base rate effect. In this case, the absence of feature "a" in the "b,c" test case produces a negative input to both diseases because it is coded -1. But because the weight from "a" is greater to the common rather than rare disease (because of its greater frequency of presentation), the common disease receives a greater negative input, and the inverse base rate effect is obtained.

The relevance of these studies to modelling artificial grammar learning depends on whether the tasks employed in these

studies tap the same concept formation processes as artificial grammar learning. In the absence of contrary evidence, the simplest assumption would be that the same concept formation processes are involved. The relevance of these studies might also be questioned because the performance of a network is as much determined by its architecture as by the learning rule used; with artificial grammar learning, the appropriate architecture would be an auto associator rather than the simple network used by Gluck and Bower (1988). Nonetheless, the relative success in the simple case examined by Gluck and Bower indicates that the delta rule should not be dismissed in examining artificial grammar learning. Further, McClelland and Rumelhart (1985, 1986) have argued for a delta rule auto associator in understanding human concept formation. They found that the auto associator could extract a central tendency or prototype from a set of patterns that were random distortions of the prototype (cf. J. A. Anderson, 1983), and that it could do this for several different prototypes simultaneously. Further, representations of specific exemplars could coexist in the same set of connections with knowledge of the prototype. The ability of the model to store nonorthogonal prototypes and patterns was dependent on the use of the delta rather than Hebb rule. These qualitative results are encouraging in considering modelling artificial grammar learning with a delta rule autoassociator.

Understanding the type of concept acquired by a delta rule auto associator is essential if our understanding of what the associator is learning is to go beyond a simple enumeration of its weights. Unfortunately, McClelland and Rumelhart (1985, 1986) make only the most general points about the type of concept acquired by a delta rule auto associator. A few further elementary comments are

made here. As for the Hebb rule, let the vector of predicted activations for the kth input pattern be given by

$$p_k = Wa_k$$

and let the a_k be classified as "grammatical" according to their correlation with their p_k. In contrast to a Hebbian auto associator, the weights matrix, **W**, produced by a delta rule auto associator is not a covariance matrix but a matrix of regression weights. Thus, in contrast to a Hebbian auto associator, factor analysis does not provide an appropriate analogy for understanding the nature of the knowledge acquired by a delta rule auto associator. However, predicted activations will be close to actual activations if the actual activations are close to the dominant eigenvector of **W**, as for a Hebbian auto associator. Thus, characterizing the dominant eigenvector of **W** would enable a characterization of what is learned by the auto associator. If the exemplars learned are coded as linearly independent, then, when learning asymptotes, any test exemplar that is a linear combination of the learning exemplars will be an eigenvector of **W** with eigenvalue equal to one. All such test exemplars would be classified as "grammatical". Any pattern that is not a linear combination of the learning exemplars would not be an eigenvector of **W**. Thus, to the extent that a test exemplar deviated from a linear combination of learning exemplars, it would be classified as "nongrammatical". If the learning exemplars are linearly dependent, it is not a priori clear how to characterize the dominant eigenvector in the general case. More will be said on this topic when the coding scheme actually used to model artificial grammar learning has been described.

Details of the Connectionist models.

Now the models specifically used in this chapter to model artificial grammar learning are considered. The models were all variants of a (feed forward) auto associator. That is, the model attempted to predict each feature of the exemplar applied based on some set of the remaining features of that exemplar. The models differed along four dimensions: The learning rule used; the coding of letter features; the coding of absent features; and the use of successive versus simultaneous prediction. These dimensions are discussed in turn.

1. The learning rule used.

The two rules used were the Hebb rule and the delta rule. During learning, the Hebb rule is parameter free. The pattern of weights produced does not depend on a learning rate, the number of iterations through the training exemplars, nor the sequence of exemplar presentation. Two learning parameters need to be considered for the delta rule: LR, the learning rate, and NI, the number of iterations through the exemplars. Before asymptotic performance, the sequence of exemplar presentation may also be important.

As long as LR for the delta rule is below a maximum value (see Stone, 1986, for what this is), it does not influence the final pattern of results, only how long it takes to get there. Thus, a "parameter free" version of the delta rule can be produced by determining the asymptotic pattern of weights. In practice, the asymptotic weights were determined by setting LR at .01 or .02 and running the model for 500 - 1000 iterations (evidence will be presented later that the weights were indeed asymptotic under these conditions). Note that there is a peak after fewer than six

iterations in the ability of the model to classify, but the pattern of classification does change after this point. Thus, for each delta rule model, its predictions were tested, first, with asymptotic weights; and, second, with weights produced by six iterations (subjects were exposed to the exemplars six times), with the exemplars presented in the same order as for subjects, and with the approximately optimal learning rate for that model. "Optimal" means the learning rate that appeared to maximize classification performance, as determined by a rough "hand" exploration of LR space. The first type of delta rule model will be called "asymptotic", and the second type "pre-asymptotic".

2. The coding of letter features.

The material presented to the models was the same material presented to subjects in Experiments Six to Eight; see Figure 2, Chapter Four, for a list of the acquisition and test exemplars. The exemplars to be learned were up to six letters in length, and each letter position could be filled (or not filled) with any of five different letters according to the rules of the grammar; see Figure 1, Chapter Four, for these rules. In single letter models, 30 units were used, one unit for each letter in each position. In digram models, in addition to single letter coding, digrams were also coded. As with single letters, the same digram in different positions was coded by a different unit. Thirty-seven units were used to code the 37 allowable digrams, with one unit corresponding to one digram. Five additional units coded non-allowable digrams, one unit for each of the five possible digram positions. Thus, in total 72 units were used for digram models.

Digram coding allows the model to learn interactive relations between the letters. Interactive relations were not

common in the grammar used, but they did exist; for example, an R preceded by a V can only be followed by an X; but an R preceded by an X can only be followed by an R or M.

3. The coding of absent features.

In all models, if a feature was present, the unit coding it was given an activation of 1. If a feature was absent, the unit coding it could have an activation of 0, for one type of model, or -1. for the other type of model. In the first case, the model is sensitive to the frequency of cooccurrence of features; that is, to the frequency of $(f_i + f_j)$, where f_i indicates the presence of the ith feature, and f_j of the jth feature. In the second case, the model can be additionally sensitive to contingency between features; that is, to the frequency of $(f_i + \tilde{f}_j)$, or vice versa, as well as of $(f_i + f_j)$. The two types of model will therefore be called cooccurrence and contingency models, respectively.

In a Hebb cooccurrence model, w_{ij} is a direct tally of the frequency of cooccurrence of features i and j (coded for by units i and j, respectively). In a delta rule cooccurrence model $w_{i,j}$ is a measure of the extent to which the occurrence of feature j uniquely predicts the occurrence of feature i. In a Hebb contingency model, w_{ij} will be decremented if feature i is present but not feature j, or vice versa. In fact, in this model, w_{ij} is a direct measure of the extent to which features i and j behave similarly; that is, of the sum of the frequencies of $(f_i + f_j)$ and of $(\tilde{f}_i + \tilde{f}_j)$ minus the sum of the frequencies of $(f_i + \tilde{f}_j)$ and of $(\tilde{f}_i + f_j)$. In a delta rule contingency model, again, of course, it is only the unique prediction of contingency that is important for $w_{i,j}$.

Note that coding the absence of a feature as -1 rather than 0 implies the active coding of the absence of a feature by the

subject. This is plausible when subjects are exposed to stimuli with a only small set of well-learned features. It is possible that in the grammar used in experiments Six to Eight, the absence of a feature was as noticable as its presence.

4. Successive versus simultaneous prediction.

In successive prediction, each unit only connected to units in previous positions. This might correspond to the case where the subject reads each stimulus from left to right. For successive models, a single permanently active 'initial unit' was used to predict features in the first position. Thus, these models had 31 units for single letter coding, and 73 units for digram coding. In simultaneous prediction, each unit was connected to all other units. This would correspond to the case where the subject used both previous and succeeding letters to constrain the identity of the letter in any given position.

All four dimensions were fully crossed to produce 16 different types of model. Apart from the differences discussed, all models followed the same procedure. In the learning phase, the model was exposed to each of the 20 grammatical acquisition exemplars used in Experiments Six to Eight and Dulany et al. (1984). For each exemplar, weights were changed according to the learning rule involved. One iteration through the exemplars was used for the Hebb rule models, six iterations for the pre-asymptotic delta rule models, and 500 - 1000 iterations for the asymptotic delta rule models. In the test phase, the model classified each of the 25 grammatical and 25 nongrammatical test strings used in Experiments Six to Eight and in Dulany et al. (1984). To do this, the weights matrix of the model, **W**, was used to predict the activation of each unit based on the other units (all other units, or only previous

ones, depending on the model) to produce a vector of predicted activations for the kth test string, p_k,

$$p_k = Wa_k.$$

As described on page 190, if a_k is close to a "central tendency" extracted by the network, then p_k will closely match a_k. Thus, the correlation, C, was calculated between a_k and p_k, where $C = a_k.p_k/|a_k||p_k|$.

In order to convert C into a response probability, the procedure adopted by Estes et al. (1989), Gluck & Bower (1988), McClelland and Elman (1986), and McClelland and Rumelhart (1986) was employed; that is, first, the "strength" of the "grammatical" response was taken to be $e^{kC - T}$ and the "strength" of the "nongrammatical" response was taken to be $e^{-(kC - T)}$, where k is a scaling parameter and T is a threshold. Second, the decision rule of Luce (1959) was applied to the "grammatical" and "nongrammatical" strengths to yield probability of responding "grammatical" to an exemplar

$$p("g") = 1/(1 + e^{-2(kC - T)}).$$

This procedure was adopted here because of its conventional use, the simplicity of the Luce rule, and because it introduces the parameter k which gives a degree of freedom in adjusting predicted to actual overall response probabilities. T is adjusted to give equal numbers of "grammatical" and "nongrammatical" responses.

From the p("g") for each exemplar, the proportion of exemplars expected to be (1) classified correctly overall (Pc), (2) classified correctly twice in a row (CC), (3) classified correctly once and in error once (CE), and (4) classified in error twice in a row (EE) over two classification blocks could be calculated. If k could be adjusted to give the same pattern of values for Pc, CC, CE,

and EE as were obtained in Experiments Six to Eight, this would provide an existence proof that the models could match overall characteristics of the experimental data given appropriate parameter tweaking. If for a given model, there was no value of k for which the experimental pattern of values could be obtained, then the model would be clearly inadequate.

From the C for each exemplar, a rank order of difficulty for grammatical and nongrammatical exemplars was constructed for each model. For the Hebb rule and asymptotic delta rule models, these rank orderings were parameter free predictions of the models. For pre-asymptotic delta rule models, the parameters were not determined by their influence on the rank orderings. As long as the rank orderings are different for different models, the correspondence between experimentally obtained and predicted rank orderings can be used to competitively test the different models.

Memory Array Models

Three types of memory array model were considered: The exemplar model of Estes (1986; also, Medin and Schaffer, 1978), the feature array model of Estes (1986), and the multiple trace model of Hintzman (1986). These are discussed in turn.

(a) The exemplar model of Estes (1986) and Medin and Schaffer (1978).

According to the exemplar model, exemplar information is stored as an array of feature values. In the simplest model, all acquisition exemplars are stored perfectly. The first step in categorizing a test exemplar is to determine its similarity to each of the acquisition exemplars. This computation is done by entering a parameter, s_i ($0 \leq s_i < 1$), for each feature i where the test and acquisition exemplars have different values, entering a "1" for each

feature with common values, and taking the product. Thus, if the test and acquisition exemplars differ on n features, then the computed similarity is

$$\prod_{i=1}^{n} s_i$$

or s^n, if $s_i = s$, for all n. s_i can be regarded as reflecting the salience of the contrast between different values of the ith feature: The lower s_i is, the greater the salience. The probability of categorizing a test exemplar as "grammatical" is a function of the sum, A, of its similarities to all acquisition exemplars. The probability of responding "grammatical" is

$$p("g") = 1/(1 + e^{-kA + T})$$

as for the Connectionist models.

Medin and his colleagues have provided considerable evidence that the multiplicative relationship used in combining similarities for (experimenter-defined) features is important in accounting for classification performance across a range of tasks. This multiplicative relationship allows the model to be sensitive to correlations between features, and not just their independent effects. Note that this aspect of the exemplar model distinguishes it from prototype theories and also networks using the Hebb or delta rules, which employ an additive combination of information. That is, the prototype and Hebb and delta network models can only solve linearly separable classification tasks: The category to which an exemplar belongs must be predictable from a linear combination of the feature values used to encode the exemplar. Medin and Schwanenflugal (1981) showed that subjects learned nonlinearly separable classification tasks just as easily as linearly separable

ones. Further, Medin, Altom, Edelson, and Freko (1982) found that even when a classification task could be solved in a linearly separable way, subjects still prefered test exemplars that preserved correlations. Kemler-Nelson (1984; see also Kemler-Nelson, 1988; Ward and Scott, 1987) showed that incidental but not intentional learners found a nonlinearly separable task easier than a linearly separable one. This is interesting because of the incidental conditions under which subjects typically learn artificial grammars. While these data are consistent with an exemplar model of concept formation, they do not rule out Hebb and delta rule models if these models include an initial nonlinear process that presents combinations of experimenter-defined features to the Hebb and delta rule networks (e.g., consider the digram coding employed in the Connectionist models above).

Four exemplar models were considered - ex1, ex2, ex3, and ex4. For all of them, the features used to code each exemplar were the 30 letter position features used for the single letter Connectionist models. The four models differed according to how s_i changed with letter position. For ex1, $s_i = s = .1$ for all i (the exact value of s made little difference to the model over the range .001 - .5). This model assumed that all letter positions were equally salient to the subjects. For ex2, s_i increased linearly with letter position, from .1 for letter position 1 and .6 for letter position 6. This model assumed that the initial letters were most salient and the final letters least salient. For ex3, s_i varied quadratically with letter positions 1 to 6, with a minimum of .1 for letter positions 1 and 6, and a maximum of .6 for positions 3 and 4. This model assumed that positions 1 and 6 were most salient to subjects. And for ex4, s_i varied quadratically with the

beginning and end of the test exemplar, regardless of the absolute letter position. This model assumed that the beginning and end letters (at whatever absolute letter position) were most salient.

(b)The feature array model of Estes (1986).

The feature probability array model of Estes (1986) uses the same memory array as the exemplar model. Categorization relies on the "perceived frequencies", f_i, of each feature value i contained in the test exemplar over all acquisition exemplars. f_i is incremented by one for each acquisition exemplar containing feature value i, and by s, $0 \leqq s < 1$,for each exemplar not containing feature value i. The probability of classifying a test exemplar is a function of

$$L = \prod_{i=1}^{n} f_i$$

where i is over all the feature values characterizing the test exemplar.

$$p("g") = 1/(1 + e^{-kL + T})$$

as before.

The features used to code each exemplar were the 30 letter position features used for the single letter Connectionist models and the exemplar models.

Estes (1986) has shown that when each cue of a pattern independently predicts category membership with a given probability, the feature and exemplar array models fare equally well in accounting for subject performance. However, as for the Connectionist models, the feature array model cannot predict learning when it is only combinations of the features encoded by the model that predict category membership; in this situation, the exemplar array model fares better (Estes, 1986).

(c) The multiple trace model of Hintzman (1986).

The MINERVA 2 model of Hintzman (1986) was an attempt to show how abstract concepts could be acquired and represented in a system that stored only episodic traces. Briefly, when a probe is presented to primary memory it activates all traces in secondary memory according to how similar they are to the probe. This results in an echo with intensity and content returning to primary memory. Echo intensity depends on the total amount of secondary memory activation triggered by the probe, and forms the basis of judgements of familiarity. Echo content depends on which particular features in secondary memory are strongly activated.

Traces are stored as feature lists where 1 codes the presence of a feature and -1 its absence. If P_j represents the value of the jth feature of the probe, and T_{ij} represents the value of the jth feature of trace i, then the similarity of the probe to trace i is given by the correlation between the two vectors, that is

$$S_i = (1/N) \sum_{j=1}^{N} P_j T_{ij}$$

where N is the number of relevant features. So far similarities have been combined additively across different features. The degree of activation of trace i is

$$A_i = S_i^3.$$

The cubic function increases the signal to noise ratio in the echo (see Hintzman, 1986). Raising S_i to a power greater than one also introduces some multiplicative terms between similarities from different features, as in Medin and Schaffer's (1978) model.

Intensity is found by summing activation over all m traces

$$I = \sum_{i=1}^{m} A_i.$$

The activation of each feature in the echo (i.e., echo content) is

$$C_j = \sum_{i=1}^{m} A_i T_{ij}.$$

Hintzman (1986) showed that MINERVA 2 could simulate a number of results from the concept formation literature: Better classification of prototypes than old exemplars with a delayed test (Posner & Keele, 1970), effects on classification of category size and of the extent of the distortion used to generate exemplars from prototypes (Homa & Vosburgh, 1976). and the effect of within category similarity amongst exemplars (Elio & Anderson, 1981).

For this chapter, MINERVA 2 was used to classify exemplars based on either echo intensity or echo content. In both cases, it was assumed that all m = 20 acquisition exemplars had been stored. Each of the 50 test exemplars were used as probes. When echo intensity was used to classify

$$p(g) = 1/(1 + e^{(-kI + T)})$$

as with the PDP models. When echo content was used, the correlation C between the probe and the echo content was calculated, and classification was given by .

$$p(g) = 1/(1 + e^{(-kC + T)}).$$

The basis of classification (intensity vs content) was crossed with type of feature coding (single letters alone vs single letters and digrams) to produce four versions of the model. For these models, k in the last two equations corresponds to the number of traces stored of each exemplar.

For all memory array models - the exemplar array, feature array, and MINERVA 2 models - k was adjusted so as to produce Pc, CC, CE, and EE values as close as possible to experimental values. The rank ordering of exemplar difficulty depended on the parameter s_i for the exemplar and feature array models, but was a parameter free prediction of the MINERVA 2 models.

The experimental data

The Pc, CC, CE, and EE values, and the rank orderings of exemplar difficulty, obtained from the models were evaluated against experimental data obtained from two sources: Experiments Six to Eight of this thesis and Dulany et al. (1984). The thesis provided data from nine separate subject groups, with a total of 112 subjects, and Dulany et al. (1984) provided data from four separate subject groups, with a total of 50 subjects.

The average Pc, CC, CE, and EE values for the thesis and Dulany et al. (1984) are shown in Table 30. Although the average values happen to be very similar across the two studies, the results from Chapter Four indicate that the precise values vary depending on experimental condition. The range of individual subject scores is also quite large; for example, in Experiments Six to Eight, Pc varied between chance and 80%. Thus, an adequate model should be able to produce Pc values near 80%.

Another constraint on the models concerns the relationship between CE and EE. Reber (e.g., 1989) has emphasized the theoretical importance of this relationship; specifically, he has argued that if the subjects' CE and EE are very similar, then their knowledge can be regarded as representative of the grammar. Experiments Six to Eight and Dulany et al. (1984) consistently found

a slightly but significantly greater EE than CE value. Reber also often obtains values similar to the means displayed in Table 30 (see Dulany et al., 1984, p. 546 for discussion on this point). The important aspect of these results that will be used to constrain the models is that EE should be slightly, but only slightly, greater than CE; the difference EE - CE was taken to be .05. Thus, a model that could obtain high enough Pc values only by inflating EE considerably above CE, would not be adequate.

Table 30

	Pc	CC	CE	EE
Thesis	.63	.48	.15	.22
Dulany et al.	.63	.51	.15	.20

For the rank orderings, it is important that they represent a reliable aspect of subject performance. The obtained rank orderings may, for example, simply represent random scatter. Indeed, Mathews et al. (1989) emphasized the divergence between knowledge representations of different subjects, as indexed by the contents of free recall reports. However, the extent of the overlap across subjects in the knowledge representations underlying classification performance remains an open question (see Chapter One). To assess the reliability of the experimental rank orderings, separate rank orderings were determined for each of the nine thesis groups. Rank orderings were determined by summing the number of correct responses to each exemplar. The 36 Spearman's correlations between each group and all the other groups were transformed by Fisher's z and averaged. The mean Fisher's z converted back to a correlation was .57 for grammatical exemplars (with a standard deviation, SD_{n-1}, in z scores of .21), and .51 for nongrammatical

exemplars (SD_{n-1}=.18). Thus, there is considerable overlap between the groups in which exemplars were found difficult. The agreement among subjects run in all the thesis groups can also be indexed by Cronbach's α: For grammatical exemplars, α=.56, and for nongrammatical exemplars, α=.65.

The data from the thesis subjects were combined to provide a single rank ordering (call it THESIS). The THESIS rank ordering did not include subjects who were run in the POSNEG group of Experiment Six. The POSNEG subjects were exposed to nongrammatical exemplars as well as grammatical exemplars during learning, and so were exposed to a different set of exemplars to the other subjects, and also to the models. The Spearman's correlation between THESIS and the rank ordering derived from Dulany et al.'s (1984) data[1] was .68 for grammatical exemplars, and .59 for nongrammatical exemplars.

In summary, there was considerable consistency in the rank ordering of exemplar difficulty between subjects. An adequate model of artificial grammar learning should be able to account for this consistency.

The data from the thesis and Dulany et al. (1984) were combined to produce a single rank ordering (call it TOTAL) to test the models with. It may be objected that by averaging over all subjects, a rank ordering is obtained that is not representative of any single subject. While recognizing this possibility, the simple assumption will be made here, based on the high levels of between subjects consistency found above, that the obtained average rank ordering represents a central tendency, around which each subject deviated by a random error.

[1]Many thanks to Don Dulany for making this data available.

Relationships between models.

This section explores the influence on the rank order of exemplar difficulty of the ways in which the models varied. The purpose of this section is to indicate which models made different predictions. The Connectionist models are considered first. Correlations relevant to each of the dimensions along which they varied are presented for the Hebb and asymptotic delta rule models, and then the relationship between the asymptotic and pre-asymptotic delta rule models are discussed. Relationships involving the memory array models are considered second. For all correlations, $r(crit)_{.05}$=.32, one-tailed, or .38, two-tailed.

Connectionist models

Table 31 shows the correlations between models differing only in the type of learning rule used.

Table 31. Hebb vs asymptotic delta.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Gramm	.92	.21	.89	.22	.26	.05	.47	.21
Nongr	.78	.68	.89	.81	.48	.48	.54	.54

For both learning rules there was more agreement on the rank order of exemplar difficulty for nongrammatical rather than grammatical exemplars. For grammatical exemplars, there was only a small positive relationship between the Hebb and delta rules.

Table 32 shows correlations between models differing only in whether single-letter or digram coding was used.

Table 32. Single letter vs digram coding.

	Hebb				Asymptotic Delta			
	Coocc		Cont		Coocc		Cont	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Gramm	.97	.98	.98	.98	.93	.90	.82	.71
Nongr	.86	.97	.97	.99	.87	.84	.77	.85

In general, whether single-letter or digram coding was used made little difference to the performance of the model.

Table 33 shows correlations between models differing only in whether they were of the cooccurrence or contingency type.

Table 33. Cooccurrence vs Contingency.

	Hebb				Asymptotic Delta			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Gramm	-.22	.13	-.44	-.01	.57	.95	.35	.79
Nongr	.06	.37	-.15	.26	.55	.84	.38	.84

The effect of the cooccurrence-contingency distinction was large for both learning rules. For the Hebb rule, there was sometimes a negative relationship between the two types of model. For the delta rule, there was always a positive relationship, though sometimes small (e.g., the digram successive model).

Table 34 shows correlations between models differing only in whether successive or simultaneous prediction was used.

Table 34. Successive vs simultaneous.

	Hebb				Asymptotic Delta			
	Coocc		Cont		Coocc		Cont	
	Single	Dig	Single	Dig	Single	Dig	Single	Dig
Gramm	.96	.96	.91	.92	.45	.60	.56	.16
Nongr	.93	.96	.83	.85	.69	.77	.69	.77

The effect of successive versus simultaneous prediction depended on the learning rule used. For the Hebbian models, whether successive or simultaneous coding was used made little difference to the performance of the model. For delta rule models, it did make a difference, particularly for grammatical exemplars.

In summary, the type of learning rule used, the sensitivity of the model to cooccurrence or contingency, and the use of successive versus simultaneous prediction produced considerable variation in the predicted rank order of exemplar difficulty, particularly for grammatical exemplars. The use of single-letter or

digram coding had little effect.

Consider now the relationship between the asymptotic and pre-asymptotic delta rule models. What evidence is there that the asymptotic delta rule models actually achieved asymptotic weights? One way of answering this is to compare the behaviour of the models to their ideal behaviour based on assuming asymptotic weights. If the model was trained on linearly independent exemplars, then it should be able to complete each of them perfectly, i.e. **p** should be identical to **a**. A standard procedure given in elementary linear algebra textbooks (e.g., Venit & Bishop, 1985) was used to determine the linear dependence of the training exemplars. A matrix was formed in which the coded exemplars formed columns. Thus, for single letter models the matrix was 30 by 20, and for digram models it was 72 by 20. The row-reduced echelon form of the matrix was determined by Jordan-Gauss elimination (using a Basic program suggested by Venit & Bishop, 1985). If the row-reduced echelon form of the matrix contains only elementary columns, then the exemplars are linearly independent, otherwise they are linearly dependent and the nature of the dependence is indicated by the row-reduced matrix (see Venit & Bishop, 1985).

For the digram models (regardless of whether they were of the cooccurrence or contingency type), the acquisition exemplars were linearly independent. Thus, **p** should be identically **a** for all acquisition exemplars for the asymptotic simultaneous models, if they are truly asymptotic. Indeed, for the contingency model, this was true to six decimal places; and for the cooccurrence model, this was true to three decimal places. Thus, the simultaneous digram models can be regarded as actually asymptotic. What about the successive digram models? In this case, the linearly independence

of the acquisition exemplars is not relevant, because each feature is predicted only by features in previous positions. Thus, these models will be able to reproduce the acquisition exemplars perfectly only if the feature at each position is a linearly separable function of features in the previous position. Because this is certainly not true of features in the first position (the first letter can be one of two letters), these models will never be able to complete the training exemplars perfectly. In fact, for the contingency successive model, the average correlation between **p** and **a** over all acquisition exemplars was .88 (range .83 - .95). However, because this correlation had remained constant to at least two decimal places over at least the last 100 iterations of learning, the successive digram models were also regarded as actually asymptotic.

For the single letter models (regardless of whether they were of the cooccurrence or contingency type), the acquisition exemplars were linearly dependent. Specifically, considering the ordering of the acquisition exemplars given in Figure 2, Chapter Four, exemplar 14 (i.e., VXRRRR - call it e14) can be obtained by subtracting exemplar 7 (MVRXRR, e7) from 6 (MVRX, e6) and adding to exemplar 12 (VXRR, e12); similarly, e17 = e2 - e1 + e8 - e10 + e15, and e18 = e4 - e1 + e8 - e10 + e15. Thus, the single letter models will never be able to complete the acquisition exemplars perfectly. However, the correlations between **p** and **a** remained constant to at least two decimal places over at least the last 100 iterations of learning, and so the single letter models were also regarded as actually asymptotic.

For the pre-asymptotic models, the correlations between **p** and **a** were less than for their asymptotic counterparts. For

example, the correlations for the pre-asymptotic contingency digram simultaneous model ranged between .62 and .87, whereas for the asymptotic model they were virtually unity. Nonetheless, the rank orderings of exemplar difficulty were comparable between asymptotic and pre-asymptotic models; Table 35 displays the correlations in rank order of exemplar difficulty between asymptotic and pre-asymptotic models.

Table 35. Asymptotic versus pre-asymptotic delta rule models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Gramm	.91	.75	.90	.76	.80	.86	.90	.63
Nongr	.80	.87	.93	.89	.79	.88	.94	.91

Memory array models.

Table 36 shows the correlations between the four MINERVA 2 models, and also the Hebb contingency single-letter simultaneous model, HC, as representative of the Hebb contingency models. I stands for categorization by intensity, C for by content; S stands for single-letter coding, D for digram coding.

Table 36. MINERVA 2 models.

Grammatical exemplars.

	DI	DC	SI	HC
DI				.96
DC	.88			.91
SI	.97	.88		.99
SC	.79	.97	.82	.85

Nongrammatical exemplars.

	DI	DC	SI	HC
DI				.98
DC	.95			.96
SI	.98	.95		.99
SC	.90	.97	.92	.93

The MINERVA 2 models all made essentially the same prediction of the rank orderings of exemplar difficulty. and this prediction is almost identical to that made by the Hebb contingency

models (the correlations with one such model are given as an example).

Table 37 shows the correlations between the four exemplar models, the frequency model, and also HC.

Table 37. Memory array models of Estes (1986).

	Grammatical exemplars.				
	ex1	ex2	ex3	ex4	freq
ex2	.92				
ex3	.68	.41			
ex4	.85	.63	.85		
freq	.41	.12	.86	.72	
HC	.69	.46	.90	.87	.89
	Nongrammatical exemplars.				
	ex1	ex2	ex3	ex4	freq
ex2	.80				
ex3	.72	.80			
ex4	.83	.84	.97		
freq	.59	.57	.84	.75	
HC	.70	.71	.87	.83	.96

The frequency model made substantially different predictions than the exemplar models, but, like the MINERVA2 models, its predictions were very similar to the Hebb contingency models. The exemplar models made somewhat different predictions dependent on the nature of s_i, particularly for grammatical exemplars, and their predictions could not be entirely subsumed by any other model.

Evaluation of the models

Models were evaluated both in terms of the Pc, CC, CE, and EE values they could produce, and also in terms of the rank ordering of exemplar difficulty that they predicted. The k parameter used to scale response probabilities in all the models was adjusted so as to maximize Pc with the constraint that EE-EC ≈ .05. In general, increasing k would increase both Pc and EE. If the maximum Pc so obtained for a model exceeded empirically obtained values, then, with suitable parameter tweaking, the models values could be made to

match experimental values (by reducing k and/or adding noise to deteriorate performance). On the other hand, if the maximum Pc was below empirically obtained values, then no parameter tweaking could rescue the model.

Pre-asymptotic delta rule models were tested against the THESIS rank ordering; all other models were tested against the TOTAL rank ordering. The reason for testing the different models against different data is that the pre-asymptotic models were sensitive to the order of presentation of the exemplars, and the order used for the pre-asymptotic models was the same as that used by the thesis subjects, but not the same as that used by Dulany et al. (1984). Thus, only the THESIS data was appropriate for the pre-asymptotic models. On the other hand, the other models were not sensitive to presentation order, and the TOTAL rather than THESIS data gives a better estimate of the subjects' rank order of exemplar difficulty independent of presentation order.

The Hebbian models are considered first, then the delta rule models, and finally the memory array models.

For all correlations, $r(crit)_{.05}=.32$, one-tailed.

Hebbian models

Table 38 shows the Pc, CC, CE, and EE values for the different Hebbian models.

Table 38 Hebbian models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Pc	.70	.73	.67	.83	.56	.64	.55	.64
CC	.57	.61	.54	.76	.37	.51	.34	.47
CE	.12	.11	.13	.06	.19	.14	.21	.16
EE	.18	.16	.19	.11	.25	.21	.25	.21

The Pc values produced by the contingency Hebbian models are

too low, and so these models are not adequate models of artificial grammar learning. The cooccurrence models pass this initial test.

Table 39 shows the correlations between the rank ordering of exemplar difficulty predicted by the Hebbian models and TOTAL.

Table 39. Hebbian models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Grammatical	-.24	-.26	-.19	-.20	-.08	-.29	-.06	-.26
Nongramm.	.48	.52	.44	.54	-.14	.12	-.20	.08

The cooccurrence models could significantly predict the rank order of nongrammatical exemplar difficulty. However, all the Hebbian models failed to predict the empirical rank order of grammatical exemplar difficulty. Thus, none of the Hebbian models are adequate models of artificial grammar learning.

Delta rule models

Table 40 shows the Pc, CC, CE, and EE values produced by the asymptotic delta rule models, and Table 41 shows the values for the pre-asymptotic delta rule models.

Table 40. Asymptotic delta rule models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Pc	.73	.75	.72	.78	.73	.70	.67	.78
CC	.62	.66	.60	.69	.63	.58	.53	.69
CE	.11	.10	.12	.09	.11	.13	.14	.09
EE	.16	.15	.16	.14	.16	.18	.19	.13

Table 41. Pre-asymptotic delta rule models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Pc	.72	.87	.72	.87	.72	.79	.71	.81
CC	.61	.83	.61	.83	.61	.72	.59	.73
CE	.12	.04	.11	.04	.11	.08	.12	.07
EE	.16	.09	.17	.09	.16	.13	.17	.12

All the delta rule models could produce reasonable values for Pc. Interestingly, in many cases, as the ability of the delta rule models to complete the acquisition exemplars improved (from pre-asymptotic to asymptotic), their ability to generalize to new exemplars deteriorated. Brooks (as reported in McAndrews & Moscovitch, 1985) suggested an exemplar model interpretation of this pattern of performance; the results here show that the pattern is not in itself indicative of an exemplar model.

Tables 42 and 43 show the correlations between the predicted and actual rank order of exemplar difficulty for the asymptotic and pre-asymptotic models, respectively. The actual rank ordering was TOTAL for the asymptotic models and THESIS for the pre-asymptotic models.

Table 42. Asymptotic delta rule models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Grammatical	.02	.57	-.02	.61	.01	.56	-.17	.69
Nongramm.	.34	.40	.32	.53	.16	.52	.12	.38

Table 43. Pre-asymptotic delta rule models.

	Cooccurrence				Contingency			
	Single		Digram		Single		Digram	
	Succ	Sim	Succ	Sim	Succ	Sim	Succ	Sim
Grammatical	-.12	.40	-.09	.43	-.19	.56	-.17	.40
Nongramm.	.50	.52	.54	.53	.06	.52	.04	.36

The successive-simultaneous distinction was of most importance in distinguishing the different delta rule models: The successive models were unable to predict the rank order of grammatical exemplars; the simultaneous models could significantly predict the rank order of both grammatical and nongrammatical exemplars. The single letter-digram and cooccurrence-contingency

distinctions did not influence the ability of the auto associator to predict the rank order of exemplar difficulty.

Memory array models.

Table 44 shows the Pc, CC, CE, and EE values produced by the memory array models.

Table 44. Memory array models.

	ex1	ex2	ex3	ex4	fre	DI	DC	SI	SC
Pc	.71	.78	.56	.63	.56	.63	.73	.62	.72
CC	.60	.69	.33	.48	.36	.47	.62	.45	.61
CE	.13	.09	.22	.16	.19	.16	.11	.17	.11
EE	.17	.15	.22	.22	.25	.21	.16	.21	.16

The ex1 and ex2 exemplar models, and the MINERVA 2 models that classified by content rather than intensity, showed adequate Pc values. The other models were not adequate in this respect.

Hintzman (1986) suggested that the performance of MINERVA 2 could be improved in some situations if the echo returning from a probe is itself fedback to secondary memory to produce a new echo; this procedure can be repeated for a number of iterations. When this procedure was followed, the Pc values for the MINERVA 2 models actually deteriorated. After one iteration, the Pc for the intensity models fell to chance. After four iterations, the performance of the content models also fell to chance. (A similar procedure can be followed for the Connectionist models: **W** can be applied a number of times to the resulting **p** until a stable state is reached. This essentially amounts to comparing each **a** to its nearest eigenvector. As for MINERVA 2, this procedure decreases Pc values for all the models except the Hebb cooccurrence models, which are only marginally improved.)

Table 45 shows the correlations between the predicted and actual rank order of exemplar difficulty for the memory array

models.

Table 45. Memory array models.

	ex1	ex2	ex3	ex4	fre	DI	DC	SI	SC
Grammatical	-.08	-.14	-.04	-.05	-.16	-.22	-.12	-.30	-.08
Nongramm.	.41	.43	.18	.23	.08	.03	.17	.12	.27

ex1 and ex2 could significantly predict the rank order of nongrammatical exemplar difficulty, but no memory array model could predict the rank order of grammatical exemplar difficulty.

Properties of simultaneous delta rule models

The only class of model that could produce adequate Pc values and predict the rank order of both grammatical and nongrammatical exemplars was the simultaneous delta rule model. The fact that there were other models inadequate in each of these respects indicates that the achievement is not trivial. It is worth noting that no other model could predict the rank ordering of grammatical exemplar difficulty. The properties of simultaneous delta rule models are now explored further. Initially, their performance on the SLD test used in Chapter Four is considered, and then the issue of what higher level description could be given to the performance of these models is discussed.

The SLD test.

The SLD test, introduced in Chapter Four, presents subjects with an initial grammatical sequence of letters (the stem), and probes for which letters could come next. Experiments Six to Eight showed that the knowledge derived from this test could be used to derive a predicted classification performance (PPSUM) nonsignificantly different from actual classification performance. There was considerable between subject variability: For some

subjects, PPSUM was considerably above classification performance, and for other subjects the reverse was true. Can this pattern of results be modelled by a simultaneous delta rule auto associator?

The SLD test was applied to the auto associator by coding the SLD stem as a vector **a** with 1 for the presence of each feature and 0 or -1 (depending on whether the model was cooccurrence or contingency type) for the absence of a feature in the stem and with 0 for all features succeeding the stem. The weights matrix was applied

$$\mathbf{p} = \mathbf{Wa}$$

and the predicted activation levels in **p** for the next letter position was taken as the model's response to the stem. If the predicted activation was positive for a letter, then the model's response was taken to be that the letter could occur after the stem; conversely, if the predicted activation was negative or zero, the response was taken to be that the letter could not occur. The magnitude of the predicted activation, even for finite activations, could vary for different responses by a factor of about 10,000. Because the magnitude of subjects' responses was constrained to vary by a factor of five or less, the magnitude of all the model's responses was set to one. That is, the model's response to a possible next letter could be either +1 (legitimate next letter) or -1 (illegitimate next letter). From the matrix of responses to all the stems, a PPSUM could be derived according to the procedure described in Chapter Four.

The SLD test probes the simultaneous models in a successive manner. A problem arises in asking the models about initial letters - the simultaneous models were only trained to predict initial letters from succeeding ones. The pre-asymptotic cooccurrence and

contingency single-letter models were retrained with an extra "context" unit that was permanently on (an activation of +1). The weights from this unit to the first position units were used to give the model's responses about first letters (the weights from this unit also contributed to all other responses). Adding the extra context unit did not substantially alter the other predictions of the model. The Pc values were the same to two decimal places, and all exemplars retained the same ranking to within three rank values.

Table 46 shows the results for the contingency single-letter model. The learning rate, LR, is the only parameter free to vary and produce different classification performances and PPSUMs for different subjects. The scaling parameter, k, was fixed by the requirement that EE - CE ≈ .05 during classification (k was not used for PPSUM).

Table 46.
SLD performance of the contingency single-letter model.

LR	PPSUM	Pc (Classification performance)
10^{-7}	.55	.63
10^{-4}	.55	.63
.001	.53	.70
.01	.74	.83
.025	.71	.79
.06	.71	.65
.065	.66	.59
.07	.59	.52

The model behaves in an interesting way as LR is increased from a very small value. Initially, both PPSUM and Pc increase with increasing LR, but PPSUM consistently underpredicts Pc. At about LR=.01, a maximum is reached for both PPSUM and Pc, and both start to decline. However, Pc declines at a faster rate than PPSUM and so eventually PPSUM overpredicts Pc. This pattern of performance could mimic the empirical results, depending on the distribution of LR over subjects. That is, with a suitable mixture of LRs, average

PPSUM could equal Pc, with only a small positive correlation between them.

Table 47 shows the results for the cooccurrence single-letter model.

Table 47.
SLD performance of the cooccurrence single-letter model.

LR	PPSUM	Pc (Classification performance)
10^{-7}	.68	.71
.001	.68	.72
.05	.74	.87
.3	.68	.69
.32	.53	.64
.34	.55	.54

The cooccurrence single letter model behaves like the contingency model in that both PPSUM and Pc increase with increasing LR, reach a maximum, and then decline. However, in this case, PPSUM consistently lags behind Pc until performance is next to chance. Thus, the cooccurrence model is not consistent with the finding of equal average PPSUM and Pc values in subjects.

The digram models were also retrained with a single context unit. The addition of the context unit left the predictions for Pc unchanged to two decimal places, and did not substantially change the rank orders for exemplar difficulty. For the contingency model, exemplars retained their ranking to within two rank values (except for two exemplars that changed by five rank values), and for cooccurrence models, exemplars retained their ranking to within three rank values.

Table 48 shows the results for the contingency digram model.

Table 48.
SLD performance of the contingency digram model.

LR	PPSUM	Pc (Classification performance)
10^{-4}	.53	.64
.01	.66	.81
.027	.58	.66

For the contingency digram model, as for the cooccurrence

single-letter model, PPSUM consistently underpredicts Pc. Thus, the contingency digram model is not consistent with the finding of equal average PPSUM and Pc values in subjects.

Table 49 shows the results for the cooccurrence digram model.

Table 49.
SLD performance of the cooccurrence digram model.

LR_{-4}	PPSUM	Pc (Classification performance)
10^{-4}	.68	.78
.025	.68	.87
.19	.74	.62
.195	.61	.57

For very low learning rates, PPSUM consistently underpredicts Pc. Pc achieves its optimal learning rate before PPSUM, and after this point PPSUM consistently overpredicts Pc. Thus, for a suitable mixture of learning rates, the digram cooccurrence model could produce equal average Pc and PPSUM values with a small positive correlation between them.

To summarize, the performance of the simultaneous delta rule models on the SLD test indicated that the single-letter contingency and digram cooccurrence models were more consistent with subject data than the other two models. However, it should be noted that consistency with the data can only be achieved by an arbitrary selection of learning rates. Unfortunately, there is no clear mapping of LR onto a measurable subject characteristic, and so, while the models are consistent with the experimental results, they do not illuminate or explain them.

Characterizing the knowledge of simultaneous delta rule models.

This section attempts to derive some symbolic rules that describe the behaviour of the simultaneous delta rule models. This is important for three reasons. First, if the models can be described in a reasonably high level way, a better understanding is

obtained of the behaviour of the models, and there is a greater chance that further predictions can be clearly derived from the models. Second, the more general the class of models that empirical results speak to, the more informative the results are (Broadbent, 1980). Thus, characterizing the knowledge acquired by simultaneous delta rule models in as general a way as possible would allow the results obtained in this chapter to speak to the whole class of models that satisfy the characterization. And thirdly, the relationship of the characterization of the model's knowledge to the model might correspond to the relationship between grammatical rules and the subject: The subject or the model obeys the rules, but does not represent them symbolically. This would be consistent with the failure of the subject to describe such rules in free recall. The question of what - if any - characterization or higher level description can be given to the knowledge acquired by the simultaneous delta rule models will be addressed here with respect to their predictions of rank order of exemplar difficulty.

All the simultaneous delta rule models made very similar predictions for rank order of exemplar difficulty for the test exemplars used, so characterizing the knowledge of one model will approximately do so for the others. Consider an asymptotic model trained on linearly independent exemplars. As discussed on page 198, all the training exemplars will form eigenvectors of the **W** for the model and so the training exemplars will all be considered perfectly grammatical by the model. In fact, any linear combination of the training exemplars will form an eigenvector of **W**, and so will also be considered as perfectly grammatical. What sort of exemplar is a linear combination of training exemplars?

For illustration, consider the acquisition exemplars used in

this chapter and in Chapter Four, as shown in Figure 2, Chapter Four. Consider a single-letter model trained on the first 13 exemplars, which are linearly independent. If exemplar three (MTV) is subtracted from exemplar four (MTVRX), the stem ...RX remains (where a "." indicates an empty position) and can be added to any three letter exemplar. If it is added to exemplar 11 (VXM), for example, the linear combination VXMRX is obtained (which is nongrammatical by the finite state grammar).

In general, if any two exemplars are the same in k positions and different in (n - k) positions, where n is the maximum length of the exemplars, and a third exemplar is similar to any one of the two exemplars in each of the (n - k) positions, then an exemplar formed from the third one by substituting the feature of the dissimilar for the similar original two exemplars in all (n - k) positions will also be accepted as "grammatical" by the model. Call this the principle of combination.

As described on pages 216-217, the acquisition exemplars were linearly dependent for single-letter models, but were linearly independent for digram models. The effect of using digram coding with the above principle of combination is that no linear combination of acquisition exemplars will have illegitimate digrams, and so the correspondence between the set of exemplars judged "grammatical" that actually are grammatical will, in general, be greater for the digram than for the single-letter models. This difference between digram and single-letter models was not well exploited by the test exemplars used, as almost all of the nongrammatical exemplars incorporated illegitimate digrams. Considering the exemplars shown in Figure 2, Chapter Four, the above principle of combination produces the following "rules" that are

followed by the digram models:

..TTVT = ..TVT (1)

...VRX = ...VT (2)

MTV.. = VXV.. (3)

....V = ...VT (5)

.TV = .TTVT (6)

.XM = .XRR (7)

....X = ...XV (8)

etc.

For example, rule (1) is produced by taking the difference between exemplars 15 (VXTTVT) and 17 (VXTVT) . Any exemplar fitting the mould on the left or right side, where "." means any letter, can be transformed to the mould on the other side (keeping the "." letters constant), and will still be regarded as "grammatical" by the digram models. All these rules operating on grammatical exemplars will produce only grammatical exemplars.

In general, a digram model will accept as "grammatical" nongrammatical exemplars if a finite-state grammar is used with the following necessary and sufficient properties: (1) the same letter in the same position can occur through different routes; (2) the letter can be followed by at least one letter in common through the different routes; and (3) the letter can be followed by at least one different letter through the different routes. With a suitable set of grammatical training exemplars, the digram model will accept as "grammatical" an exemplar in which the common letter has been swapped for the different letter. The finite state grammar used here does not have these properties. Thus, all the rules abstracted by the model will be representative of the grammar. However, a model trained on only a subset of exemplars will not in general have

a complete representation of the grammar. Interestingly, this is the state of affairs claimed by Reber (e.g., 1989) for subjects: They have representative but incomplete knowledge of the grammar. However, if the delta rule model is a valid model of subjects' knowledge, then subjects will not abstract representative knowledge from a set of grammatical exemplars if an appropriate finite-state grammar is used.

In order to determine the linear dependence of the test exemplars on the acquisition exemplars, all 70 exemplars were coded as column vectors in a 72 by 70 matrix. The row reduced echelon form of this matrix revealed that 13 of the 25 grammatical test exemplars were linear combinations of the acquisition exemplars, but, of course, none of the nongrammatical test exemplars were linear combinations. These 13 grammatical test exemplars were, of course, the 13 best classified exemplars by the digram models. Eleven of these 13 were amongst the 13 best classified grammtical exemplars by the single-letter models, which approximate the digram models with this data set. Were these 13 exemplars anything special as far as subjects were concerned? In fact, 11 of these 13 were amongst the 13 best classified grammatical exemplars in the TOTAL data. This is what would be predicted by any model that consisted of all the rules that could be produced by combining the acquisition exemplars according to the principle of combination described above.

Does the simultaneous delta rule work well because, as it approaches asymptote, it approximates these rules, or do the rules work because they approximate a real delta rule auto associator? The asymptotic delta rule models do a better job at predicting rank order of exemplar difficulty than pre-asymptotic models, and so it appears that the closer the delta rule models approximate the rules,

the better the job they do. However, the delta rule models do more than just embody the rules - they can also classify to varying degrees grammatical exemplars that do not fit the rules. Consider the 12 grammatical test exemplars that were not linear combinations of the acquisition exemplars, and therefore did not follow from the rules. The rules in themselves do not speak to how these exemplars should be classified, but the delta rule models still do. Does the rank ordering of the delta rule models for these exemplars match that of subjects? The correlations between TOTAL and the predicted rank order predicted by the cooccurrence and contingency digram models were .35 and .78, respectively (.50 is needed for significance at the 5% level), indicating some degree of match. Also, note the significant correlations between TOTAL and the predictions of the digram models for the nongrammatical test exemplars. Thus, the digram models provide a measure of the extent to which the rules are broken that matches the measure of subjects.

To summarize, the digram simultaneous delta rule model trained on linearly independent acquisition exemplars can be regarded as embodying (1) a set of incomplete but (for this grammar) representative rules; and also (2) a measure of deviation from the rules. If the measure of deviation can itself be characterized, then the empirical results for rank order of exemplar difficulty would support all models of this class. Future research could usefully characterize the measure of deviation and also the knowledge of models trained on linearly dependent acquisition exemplars.

Conclusion

This chapter tested a range of Connectionist and exemplar models for their ability to account for artificial grammar learning. The criteria used were the ability to produce adequate levels of Pc and of (EE-EC), and to predict the rank order of both grammatical and nongrammatical exemplars. Only one class of model - the simultaneous delta rule model - could satisfy all the criteria. In fact, the simultaneous delta rule model was the only model that could predict the rank order of grammatical exemplar difficulty. It was shown that some versions of this model were consistent with the empirical results obtained with the SLD test in Chapter Four, and that the classification knowledge of the digram versions could be regarded as embodying representative but incomplete rules of the finite state grammar.

Although successful by the criteria of this chapter, it would not be difficult to falsify simple models like the simultaneous delta rule auto associator. For example, without the addition of further processes, it would not be able to account for the transfer of knowledge from one letter set to another (Mathews et al., 1989; Reber, 1969). Also, certain aspects of the encoding characteristics appear arbitrary: The same letter in different positions is as different to the model as different letters in different positions. In fact, it seems likely that a subject might not clearly differentiate a T in the fourth position or a T in the fifth position. And why should a subject start off with a unitary representation of each digram, as assumed by the digram models?

A future direction for modelling artificial grammar learning may be to introduce a procedure for extracting features by creating

useful higher level units out of lower level ones. This could be achieved by, for example, competitive learning algorithms with hidden units (Rumelhart & Zipser, 1986) or the algorithm used by Wolf to segment prose into meaningful units (Wolf, 1975, 1977, 1980). The use of hidden units might allow the creation of unitary representations of abstract structures in the grammar not tied to particular letters; a representation of the fact that the same letters in different positions are the same letters; and also the creation of unitary representations of digrams in a natural way. Perhaps the type of model used in this chapter is best regarded as a second stage operating on the emerging units created by a first feature extraction stage.

Another issue deserves comment. The models used in this chapter were particularly geared to artificial grammar learning. What aspects of the models are relevant to understanding implicit learning generally? This question would be best addressed by considering what sort of model could model learning in other paradigms, and what sort of model does so successfully. Accordingly, Chapter Six addresses the issue of modelling the dynamic control tasks, investigated by the Broadbent group.

Chapter Six

Modelling implicit learning: The control tasks

Introduction

Chapter Five indicated that the delta rule auto associator could model artificial grammar learning with some success. However, if the delta rule auto associator is to be a useful model of implicit learning in general, then it should be relevant to understanding subject performance on a number of tasks, not just artificial grammar learning. Chapter Six addresses the generality of the model by considering its application to learning the dynamic control tasks. Initially, the sort of knowledge that could be acquired by subjects learning the control tasks is discussed. Then an existing Connectionist algorithm - reinforcement learning - that creates such knowledge is considered as a model for learning the control tasks. Finally, the use of an auto associator to embody principles similar to those employed by reinforcement learning is considered.

In the control tasks (e.g., Berry & Broadbent, 1984), the subject controls the level of one variable (e.g, work force in a sugar production factory) in order to reach target values on another variable (e.g., the amount of sugar production). What sort of knowledge could be acquired by a subject in learning the control tasks? Broadbent, Fitzgerald, and Broadbent (1986) suggested that implicit knowledge may be represented as a "look-up table" in which successful actions are stored together with the associated situations. The best action for a particular situation can then be determined by matching the current situation to the most similar

stored one and emitting the associated action. Indeed, Marescaux et al. (1989) provided evidence that subjects' knowledge of appropriate action on the control tasks is linked to particular situations. Specifically, Marescaux et al. found that subjects were more likely to respond correctly to a situation that they had encountered before and which had been followed by the target than to other situations.

Barto (e.g., Barto, Sutton, & Anderson, 1985; Barto, Sutton, & Brouwer, 1981) has suggested a Connectionist learning rule that creates a look-up table; he calls the procedure reinforcement learning. Reinforcement learning is now described and its application to the control tasks considered. This sets the stage for considering how principles similar to reinforcement learning could be applied by an auto associator.

Reinforcement learning

Reinforcement learning is a procedure for acquiring control over a system with unknown dynamics when the only feedback is the extent to which an outcome is favourable (see Barto, Sutton, and Anderson, 1985). For every action that can be performed there is an "adaptive element". The output activation, o_j, of the jth element is a weighted sum of the input, c_i, from units coding environmental context, plus random noise, N, with mean zero normal distribution (standard deviation = SD):

$$o_j = \sum_i w_{j,i} c_i + N.$$

For mutually exclusive actions, the output activation of each adaptive element is compared, and the action associated with the element with the highest activation is chosen. Its activation

is set to 1, and that of the other elements to 0. A reinforcement, Z, is fedback from the environment to the element, the value of Z depending on the consequences of the action; Z may be positive (rewarding) or negative (punishing), discrete or continuous.

A separate element attempts to predict the Z for a given context according to the delta rule, with learning rate L:

$$PZ = \sum_i wz_i \, c_i$$

$$\Delta wz_i = L(Z - PZ)c_i$$

The change in weight from the environmental context in which the action was taken to the element that represents the action chosen is simply proportional to the difference between the Z and the predicted Z for that context; the constant of proportionality is the learning rate LR:

$$\Delta w_{i,j} = LR(Z - PZ)o_j c_i.$$

In order to model learning the control tasks, a reinforcement function for Z needs to be specified, and what is encoded by the context units, c_i, needs to be specified. There are then three free parameters: SD (the standard deviation of the normally distributed noise), LR (the learning rate for the weights between context and action), and L (the learning rate for the weights between context and reinforcement). There are reasons for expecting the optimal value of each to be nonzero, but in practice this may not be so for SD and L.

A nonzero LR provides the essential component of the model: It allows nonzero weights between situations and actions. If LR is zero, then the model will perform at chance.

The presence of noise (corresponding to a nonzero SD) would seem to be necessary to allow a variety of actions to be attempted

in a given situation; but perhaps the absence of noise would allow a more efficient and systematic exploration of action space.

A nonzero L allows the model to calculate an expected level of reinforcement for a given situation. Only if a reinforcement is better than expected are the weights to an action strengthened. This is to allow learning in a situation where the reinforcement is negative but is the best that could be obtained, or where the reinforcement is strongly positive but it is possible to do even better. There is evidence that animals are sensitive to expected reinforcement: For example, the existence of learning when the animals are trained with negative reinforcement or the avoidance of punishment. Further, animals show "successive negative contrast" effects during instrumental conditioning: If a positive reinforcer is substituted with a less favorable one, performance deteriorates below that of animals trained with just the less favorable reinforcer (see Mackintosh, 1974, p. 213). But perhaps this extra mechanism is not needed for modelling the learning of the control tasks. If L is zero, then a rewarded action will always be strengthened and a punished one weakened.

Three versions of the control task were simulated: The person interaction task with Clegg (Berry & Broadbent, 1984), and Persons S and U (also called the Salient and Nonsalient tasks: Berry & Broadbent, 1988; Hayes and Broadbent, 1988). In the person interaction tasks, the subjects control their level of friendliness to a computer personality on a 12 point scale, ranging from very rude to loving. Their aim is to move the computer personality's friendliness to a target value, and maintain it there. The target that has typically been used is 9 (very friendly). Let S be the subject's friendliness, PS the subject's friendliness on the

previous trial, C the computer personality's response to S, and PC the computer personality's response on the previous trial. Then for the Clegg computer personality,

$$C = 2S - PC + R$$

where R is randomly -1, 0, or +1.

For Person S,

$$C = S + 2 + R.$$

And for Person U,

$$C = PS + 2 + R.$$

Z was set to be one minus the absolute difference between the target behaviour and the actual behaviour. Thus, Z varies continuously, and the further from target that the computer personality's behaviour is, the more negative Z is. Z is positive only when the computer personality's behaviour is on target. The behaviours of the previous trial were coded as the "environmental context" for the next trial; additionally, there was one context unit permanently on to code for the invariant features of the context, such as the target.

Table 2 shows the experimental scores, taken from Berry and Broadbent (1984) for Clegg (the first score is the number of times on target in the first set of 30 trials, the second score is for the second set of 30 trials), and Broadbent and Hayes (1988) for Persons S and U (the score is the number of times on target in the last 10 of 30 trials for Person S and in the last 10 of 50 trials for Person U).

Every set of parameter values attempted for each version of the control task was simulated 10 times by starting each simulation with a different random number seed (the random numbers were used in generating the noise and in determining the random component, R, for

the computer personality's response). The mean for each set of 10 simulations will be reported with the standard deviation in brackets (sums of squares divided by nine).

Table 2.

Task	Clegg	Person S	Person U
	9, 16	7	7

Consider first Persons S and U. Estimated chance performance was determined by setting all parameters to zero except for noise. With SD=.005, the score for Person S was 3.0 (1.9), and for Person U, 2.6 (1.4). With LR=.01, SD=0, and L=0, the score for Person S was 9.5 (1.3), and for Person U 9.3 (2.2). That is, the model does not need the noise component (SD=0) nor the reinforcement prediction component (L=0) to learn almost perfectly. In this case, the model starts with the first action, which is negatively reinforced, and so moves on to the second, and so on, until the correct action is chosen. (For Persons S and U, the correct action is independent of the situation.) Note that for Person U, the action responsible for the reinforcement is not the one that receives it; the action on trial (n+1) receives the reinforcement produced by the action on trial n. The reason the model still learns Person U is partly a result of the reinforcement function used: If the computer personality's response, C, is one off target (e.g., a friendliness of eight instead of nine, where nine is target), the model's action is neither punished nor rewarded and so the action may be repeated on the next trial. Thus, because the model moves through the actions sytematically and in order, the correct action will be repeated and appropriately rewarded. It should be noted

that subjects do not behave in this systematic way, as will be discussed on page 243.

The systematicity in the model's behaviour can be abolished by letting SD be nonzero. With LR=.01 and SD=.005, the score for S was 9.7 (0.7) and that for U was 3.1 (2.1). Increasing L above zero can make the model's responses "sticky": It quickly learns to expect a low level of reinforcement, and so the model may stick to the first action it does that's just a bit better than that. For example, with SD=0, LR=.01, and with L=1, an extreme value, the score for S was 3.3 (2.9), and that for U was 1.8 (2.1). The low values were obtained because the model sticks with a single response. Adding noise, by increasing SD, can eliminate the stickiness, as can reducing L. By balancing various values of SD, LR, and L, almost any score can be obtained for Persons S and U. U requires an optimal amount of stickiness so that it repeats itself now and again: This allows the reinforcement to apply to the actual action responsible for it. For example, with SD=.005, LR=.01, L=.01, the score for U was 8.2 (2.0); that for S was 8.6 (1.9).

Hand exploration of parameter space suggests that the model could learn Person S over large regions of parameter space, Person U over smaller regions, and Clegg over yet smaller regions. Respectable scores for Clegg could only be obtained by finely tuning the parameters. With SD=.1, LR=.05, and L=.05, the scores for two successive blocks were 7.8 (5.6) and 14.0 (4.2). The reinforcement prediction component is important: If L=0 (SD=.1, LR=.05) then the scores are 4.8 (1.5) and 3.7 (2.3), a significant reduction in the second score, $\underline{t}(18)=7.08$, $p<.05$. The noise component appears less important: If SD=0 (L=.05, LR=.05) then the scores were 9.1 (2.2) and 11.1 (4.1), the $\underline{t}$ for the reduction in the second score being

1.68, ns. With these apparantly "optimal" parameter values for Clegg (SD=.1, LR=.05, L=.05), the scores for S and U were 8.3 (1.6) and 3.9 (3.8) respectively. There may be regions of parameter space for which both Clegg and Person U could be learned, but they were not readily apparant.

The reason that Person U, at least, was learned over smaller regions of parameter space than Person S is that when the model was learning Person U the wrong action was generally being reinforced. Introducing an exponentially decaying "trace" so that previous actions were reinforced to some extent (Barto, Sutton, & Anderson, 1985) did not solve the problem as the reinforcement of irrelevant actions still swamped the reinforcement of the relevant action.

Hayes (1987) argued on the basis of inter-response latencies that three consecutive responses by the subject should be treated as a single functional response; this suggestion was incorporated into the simulation by reinforcing groups of three responses according to the Z obtained after the third response; also, each response in a group was generated from the previous one by randomly adding -1, 0, or +1. As expected, this did increase the robustness of Person U to parameter changes. For example, considering the parameters that were optimal for Clegg with no response grouping (SD=.1, LR=.05, L=.05), the score for U was 6.1 (1.1). With the response grouping, however, the model did not learn Clegg: With these parameters, the scores for Clegg were 5.3 (3.3) and 8.9 (3.4). No set of parameter values could be found that substantially increased the scores for Clegg. Thus, grouping the responses does not appear to be of general use to the model.

The detailed pattern of responding of actual subjects was analyzed to provide clues for assessing and modifying the models.

Unfortunately, the trial by trial responses for subjects from Hayes and Broadbent (1988) were not available (for Persons S and U); but they were from Berry and Broadbent (1984) (for Clegg). The data from 24 subjects were used. On the first trial, before any learning could have occurred, responses were not distributed randomly over the 12 response categories, as they were for the models. The distribution of these first responses is shown in Table 3. Subjects mainly used values of 7, 8, or 9 on their first trial. One explanation for this subject bias is that subjects might presume that moving the computer's behaviour to the upper middle of the computer's scale should require them to respond in the upper middle of their scale.

The initial bias was incorporated into the model by hand setting the initial weights to the action units from the context unit that was permanently on. The resulting distribution of initial responses is shown in Table 3. Table 3 shows that people both started with a response bias and maintained a restricted range of responses: The distribution of responses over the first 30 trials was very similar to the distribution of initial response (though significantly different, Hotellings $T^2(11,13)=98.59$, $p<.005$). The model behaves in a similar, but not identical way, maintaining a restricted range of responses. It is informative to know how the model would perform with this response distribution in the absence of any learning. With LR=0 and L=0, the scores for Clegg were 4.8 (2.9) and 5.3 (2.8). For Person U, the score was 5.9 (1.8). This is very close to the level of performance obtained by Hayes and Broadbent (1988), but the response bias is not able to explain the level of performance reported by them. In Hayes and Broadbent (1988), subjects also interacted with a version of Person U in which

two was subtracted from, rather than added to, the subjects previous behaviour to produce the computer personality's behaviour. With this version, the chance level of performance was 1.1 (1.0). As subjects' score for this version of the task was 5-7 over three experiments, learning must have occurred for the subjects.

With the previously optimal parameters for Clegg (SD=.1, LR=.05, L=.05) the performance of Clegg was not benefitted by bias (the scores for two successive blocks were 7.1 (3.4) and 11.0 (7.1) with a response bias in the model; compare the values of 7.8 (5.6) and 14.0 (4.2) given earlier for no response bias). With these parameters, Person U did not evidence any learning. When the computer personality responded two greater than the model's previous behaviour, the score was 6.6 (3.0); when the computer personality responded two less than the model's previous behaviour, the score was 3.1 (3.1). Neither of these scores for Person U were significantly above chance (using the chance values calculated in the previous paragraph). In summary, introducing an initial response bias to match that of subjects does not improve the performance of the model.

Table 3. Response distribution for interaction with Clegg.

Response:	1	2	3	4	5	6	7	8	9	10	11	12
People:												
Initial	.00	.00	.00	.00	.04	.13	.25	.29	.25	.00	.00	.04
Over 30 trials	.00	.00	.00	.01	.02	.09	.15	.31	.27	.10	.03	.00
Model:												
Initial	.01	.01	.00	.01	.09	.18	.22	.24	.21	.00	.00	.02
Over 30 trials	.01	.01	.00	.00	.03	.05	.08	.22	.40	.06	.04	.11

An important experimental finding concerning the sort of knowledge acquired by subjects interacting with Clegg was provided by Marescaux et al. (1989). It is thus interesting to see if the

reinforcement learning algorithm produces similar data. Marescaux et al. demonstrated how dependent giving the correct response was on the existence of a previous situation in which the correct response was given. They trained subjects on the Clegg task for two blocks of 30 trials. Subjects were then presented with situations (Clegg's three previous responses) and were asked to decide which action was necessary to bring Clegg to target. Some questions were general for all subjects and others were specifically selected for each subject. The selected questions referred to situations that the subject had encountered in the last 30 trials of training and for which the subject's response had been followed by the target. The key result was that subjects answered a greater proportion of selected than unselected questions correctly: .71 (standard deviation .33) compared to .52 (standard deviation .23). The model should be able to produce this type of result because it was explicitly designed to produce a look-up table. The responses subjects gave to the selected questions matched the responses they gave to the same situation in training 57% (standard deviation 35%) of the time (Marescaux et al called this measure the concordance index). It is less clear that the model would produce only 57% concordance: A perfect look-up table would produce 100% concordance.

The "situations", i.e. the context, coded by the model are Clegg's previous response and the model's previous response. With SD=.1, LR=.05, and L=.05, the model was trained for 60 trials and then tested with all possible situations, defined by the possible permutations of Clegg's previous response and the model's previous response. To test the model, the relevant context units were activated, random noise (SD=.1, same as that used during learning) was added to the output activations of the elements, and the element

with the most activation was selected as the response. The situations were classified as old situations (in the last 30 trials) in which the model had earlier produced a response that was followed by the target (OldT), old situations in which the model had earlier produced a response that was not followed by the target (OldNT), and new situations (New). The proportions of questions answered correctly (i.e., would produce the target or the adjacent response) were 1.00 (0.00), .02 (.03), and .22 (.09), respectively. Thus, like subjects, the model responded more correctly to OldT situations than other situations. However, it overpredicted subjects' performance in OldT situations (subjects' performance was .71; the difference was significant, $t(11)=3.04$, $p<.01$). The concordance index for OldT was 100% (0%), and for OldNT 13% (4%). Thus, like subjects, the model had a substantial concordance index for OldT, but it also overpredicted subjects' actual concordance rate for OldT (57%); the difference was significant, $t(11)=4.26$, $p<.01$. That is, the model appears to be a more effective look-up table than subjects.

One possible explanation for why subjects had less than 100% concordance is that the "situation" presented to subjects by Marescaux et al. (1989) might have only partially overlapped the situation actually encoded by subjects. To determine the effect of partial situations on the model, the model was only presented with Clegg's previous behaviour as the situation for questions, but was trained on both Clegg's and the model's previous behaviour as context. This hardly affected the results, and the concordance rate was still 100%.

In summary, for the person interaction tasks, the model can learn all three computer personalities, though with different sets

of parameter values for Person U and Clegg. Neither grouping each successive set of three responses nor introducing an initial response bias enabled the model to learn Clegg with any more robustness. When learning Clegg, the model, like subjects, acquires a situation-specific knowledge of the correct response, though the model's responses are more tightly linked to situations than are subjects'.

The Barto reinforcement learning model has been usefully applied here to learning the person interaction tasks. However, two characteristics of the model in relation to the empirical data will be mentioned because of the implications they may have for a possibly more satisfying model. First, one of the key empirical results in the control task literature is that, for some computer personalities, subjects can learn to produce a correct response in a given situation even though they cannot predict the result of an incorrect response. That is, subjects lack predictive knowledge. This has been replicated in several studies for Clegg (e.g., Berry & Broadbent, 1984; Berry, 1990). The lack of predictive knowledge is also apparently true for Person U (Berry & Broadbent, 1988; Hayes & Broadbent, 1988), though these results may require some empirical clarification (this point is discussed on page 259). This lack of predictive knowledge is built into the Barto model, as the weights only go from context to action and not vice versa. Thus, the Barto model cannot explain the lack of predictive knowledge. A more satisfying model would allow the possibility of weights forming from the model's behaviour units to the resulting computer personality's behaviour units. This greater inter-connectedness would be most naturally modelled with an auto associator. The lack of predictive knowledge in such a model would then be an interesting finding.

A second characteristic of the Barto reinforcement learning model is that it behaves more like a look-up table than subjects do; that is, the concordance index for the model is 100% rather than about 50%. Marescaux et al. (1989) found that subjects behaved only somewhat like a look-up table. A plausible explanation is that subjects engage in a number of strategies, only one of which is modelled by reinforcement learning. However, the more the available data can be explained in terms of a single mechanism, the more satisfying the explanation is. An auto associator, which does not so deliberately embody a look-up table, may be able to provide a better match to the results of Marescaux et al. (1989).

The auto associator

The discussion of the Barto model lead to the idea that the auto associator might provide a suitable model of learning the control tasks. The auto associator is also interesting because Chapter Five found that a delta rule auto associator could model artificial grammar learning with some success. The auto associator would acquire more general importance as a model of implicit learning if it could also learn the control tasks.

How could such an architecture learn the control tasks? The first step is to decide what are the relevant features to be encoded. As for the Barto model, these are: The context, the behaviour to be produced, and the reinforcement. As for the Barto model, what constitutes the context units needs to be specified and a reinforcement function needs to be specified. The reinforcement function would determine the pattern of activation over the reinforcement units as a function of the outcome of the model's action. Also as for the Barto model, it may be necessary to add

noise to the behaviour units so that the correct action has a chance to be elicited.

The auto associator could then work in the following way. The aim is to choose the action in a particular context that produces the most positive reinforcement. The model could do this by seeing which action unit is most strongly activated when the reinforcement units are coding the most positive reinforcement, and the context units are coding the current context, and some random noise is added to the activation of the action units. The model could learn the consequences of this action by changing weights when the action units code the action chosen, and the reinforcement units code the reinforcement following the chosen action.

In terms of the features encoded and the use of noise the auto associator can be set up in the same way as a Barto model. There are also three intrinsic differences between the models. One is the way that reinforcement acts: In the Barto model, reinforcement acts by combining multiplicatively with context activation in determining weight change; in the auto associator, reinforcement acts by combining additively with context activation in determining behaviour activation. Thus, in the auto associator, context-action links are not reinforced as such. This difference between the models may be relevant to modelling the variability with which subjects respond to a previously reinforced situation, as well as providing an opportunity for some consistency (Marescaux et al., 1989).

The second difference between the models is that only the auto associator allows the possibility of the model acquiring predictive knowledge. Thus, the absence of predictive knowledge in the model would be an interesting result for the auto associator,

but is not for the Barto model.

The third difference is the number of parameters that need to be determined. For the auto associator, there are only two, rather than three: The standard deviation of the noise, SD, and the learning rate, LR. The Barto model also uses the parameter L, for the prediction of reinforcement.

Two types of auto associator were constructed depending on how the context was coded. Both used the computer personality's and the model's behaviour on the previous trial as context, but coded them in different ways. Both attempted to learn Person's S and U and Clegg.

For the first model, the context was not coded by a separate set of units, but by the persisting activation in the units from the previous trial. Twelve units coded each of computer personality's 12 possible behaviours (the CP units), 12 units coded each of the model's possible behaviours (the behaviour units), and four units coded reinforcement. An input of 1 coded the presence of a feature, and 0 its absence, or -1 for negative reinforcement. The model cycled through the following procedure. An activation vector $\mathbf{a}_1$ was produced by setting all reinforcement units to 1, and leaving all behaviour units and CP units as they were on the previous trial. This activation was then allowed to pass to the rest of the net through **W**:

$$\mathbf{a}_2 = \mathbf{W}\mathbf{a}_1.$$

Random noise was added to the activation of each of the model's behaviour units (mean 0, standard deviation, SD). The behaviour unit with the most activation was selected to represent the model's behaviour; the activation of that unit was set to 1, that of the other behaviour units to 0. The computer personality's behaviour

was calculated; the activation of that CP unit was reset to 1, that of the other CP units to 0. If the computer personality's behaviour was on target, or only one off target, the activation of all reinforcement units was set to 1. In all other cases, the activation of all reinforcement units was set to -1. Let a_3 represent the current state of activation across the net. The activation was then allowed to pass through **W**:

$$a_4 = \mathbf{W}a_3.$$

The weights between all units were changed according to the delta rule, where a_3 provided the target activation values, and a_4 the actual values. The weight from each unit to itself was always zero. Then the cycle began again for the next trial. Thus, the behaviour units and the CP units remained active as context for the next trial.

Note that the behaviour units and the CP units were used in two ways. First, their persisting activations from the previous trial allowed them to act as context in vector a_1 when choosing the appropriate behaviour for the current trial. Second, their activations in a_3 allowed the behaviour units to code the current action and the CP units to act as the consequences of the current action. When the weights were changed, the behaviour and CP units were fulfilling their second role. Comment is now made on the meaning of this, first, for the CP units and, second, for the behaviour units.

The psychological assumption here for the CP units is that although, when the weights were changed, the CP units coded events that occured after those coded by the behaviour units, some "backward conditioning" occurred from the CP units to the behaviour units. Thus, when the CP units were acting as context in a_1, their

residual activation influenced the choice of behaviour. Backward conditioning is known to occur in animals (see Mackintosh, 1983, p. 209), and was included in this model mainly because of the assumption that all units were connected in both directions. The effect of the backward conditioning was to introduce a tendency for the model to repeat the same behaviour, unless there was sufficient negative reinforcement.

The behaviour units were only nominally used in two roles (to be consistent with the use of the CP units in two roles): Maintaining activation in the behaviour units in $\mathbf{a}_1$ produced the same result as setting the activation of all behaviour units to zero in $\mathbf{a}_1$. This is because the weight from any behaviour unit to any other was always zero (and this in turn was a consequence of coding the absence of a feature as zero), and so the pattern of activation across the behaviour units in $\mathbf{a}_2$ was unaffected by the pattern of activation across the behaviour units in $\mathbf{a}_1$.

Ten different random number seeds were used for 10 runs of the model for each set of parameter values with each of Person's S and U and Clegg. As for the Barto model, Person S could be learned over large regions of parameter space, Person U and Clegg over smaller regions. With noise alone, the performance was 2.8 (1.1) for Person S, 2.8 (1.1) for Person U, and 3.6 (1.3), 4.6 (1.8) for the two successive blocks with Clegg. With no noise (SD=0) and LR=.0001, Person's S and U perform adequately, 7.4 (2.0) and 8.8 (2.0) respectively, and Clegg performs substantially above chance, 9.8 (3.1) and 11.8 (3.8). However, as for the Barto model, with no noise the model behaves in a systematic fashion not characteristic of subjects. Optimal parameter's for Clegg were found by hand exploration to be SD=10^{-5}, LR=10^{-4}; the scores were 9.0 (4.7) and

15.0 (4.1). With these parameter values, the scores for Person's S and U were 6.4 (3.5) and 9.2 (1.5), respectively.

Because the model did not code context with distinct units, the model did not possess weights from context units to behaviour units. However, the backward conditioning from CP units to the behaviour units could well result in a tendency for the model to give the same behaviour to the same situation (as defined by the computer personality's previous behaviour). It is thus an open question as to whether the model behaves like a look-up table, in the same way that the subjects of Marescaux et al. (1989) did. To answer this question, the model was tested in the same way as the Barto model. After training for 60 trials with $SD=10^{-5}$ and $LR=10^{-4}$, the model was presented with situations consisting of Clegg's previous behaviour (note that the Barto model could be tested with situations defined by both Clegg's and the model's previous behaviour). As for the Barto model, activation was passed through **W** to the behaviour units, noise was added to the behaviour units ($SD=10^{-5}$), and the behaviour unit with the most activation was selected as the response. This response was scored as correct if it would result in a computer personality's behaviour that was on target, or just one off target. The proportion of questions answered correctly was .72 (.23) for old situations that were followed by the target, or an adjacent response, in training (OldT). The corresponding proportion found by Marescaux et al. for subjects was .71 (.33). The proportion of questions answered correctly was .32 (.13) for old situations not followed by the target (OldNT), and .21 (.33) for new situations (New). Marescaux et al. reported .52 (.23) questions answered correctly for an unselected and unknown mixture of OldT, OldNT, and New questions. The proportion of

questions for which the subject gave the same response during this testing as during training (i.e., the concordance index) was 79% (25%) for OldT questions, and 61% (27%) for OldNT questions. Marescaux et al. reported a concordance index of 57% (35%) for OldT questions (not significantly different from the model's, $\underline{t}(20)=1.71$, $\underline{p}>.1$), but they did not report an index for OldNT questions. In summary, the model did behave like a look-up table, and in a similar way to the subjects of Marescaux et al..

What predictive knowledge did the model acquire? When the weights were changed, the model only represented the model's current behaviour and the computer personality's resulting behaviour, and neither of the behaviours on the previous trial. The model could have no predictive knowledge for Clegg: Clegg's behaviour on trial n depends on its behaviour on trial (n-1), but the coding of Clegg's behaviour on trial (n-1) in the model could not influence the coding of Clegg's behaviour on trial n in the model. Nonetheless, after learning Clegg with optimal parameters ($SD=10^{-5}$, $LR=10^{-4}$), the model was tested for predictive knowledge in the following way. The activation of one of the CP units was set to one (to code the computer personality's behaviour on the previous trial), and the activation of one of the model's behaviour units was set to one. Activation was allowed to pass back to the CP units through **W**. The CP unit with the most activation was selected as the computer personality's behaviour. The model was probed with each of the 12 (CP units) X 12 (behaviour units) = 144 possible situations. The proportion of predictive knowledge questions correct was .22 (.02). The equivalent proportion obtained for subjects was .32, and chance performance (i.e. with all responses equiprobable, which is not true of the model) can be calculated as .16.

Similarly, when learning Person U, the model could contain no representation of the connection between computer personality's behaviour and the model's behaviour on the previous trial. The model could not simultaneously code its behaviour on the previous and current trials; if it were given both pieces of information to code in a predictive knowledge question, the coding of the behaviour on the previous trial would be overwritten by the coding of the behaviour on the current trial. All predictive knowledge questions given to subjects about Person U (e.g., Hayes & Broadbent, 1988), have presented possible subject's behaviours on the both the previous and current trials (and then asked subjects to predict the computer personality's behaviour). Thus, the model could have no predictive knowledge for Person U.

Subjects who have learned Person S can predict the computer personality's behaviour based on their own behaviour (e.g., Berry & Broadbent, 1988). The model could also in principle learn this connection because the relevant features are simultaneously coded by the model when the weights are changed. After learning Person S, the model was tested for predictive knowledge in the same way as for Clegg. When trained with SD=10^{-5}, LR=10^{-4}, the mean proportion of predictive knowledge questions correct was .37 (.21); the mean proportion expected by chance alone is .22. With SD=0 and LR=10^{-4}, the mean was .52 (.03). Thus, for some parameter values the model could show appreciable predictive knowledge for Person S, though somewhat less than that shown by subjects (proportions of .80 to .85 were found by Berry & Broadbent, 1988, and Hayes & Broadbent, 1988).

To summarize, when the model coded context not by distinct units, but by persisting activation in the units, it could learn all three person interaction tasks, and it could acquire appreciable

predictive knowledge only for Person S.

The absence of predictive knowledge for Person U and Clegg in the model was not arbitrary in the sense that one-way connections from context to action were built into the model (as for the Barto model). On the contrary, the connections between everything that was coded did go in both directions. However, the absence of predictive knowledge in the model was a consequence of the way context was coded; i.e., as persisting activation. The question arises as to what predictive knowledge would be acquired by the model if the model's and computer personality's behaviours on the previous trial were distinctly coded with separate units, along with the current behaviours. Thus, a second auto associator was set up that simultaneously encoded all these features.

Twelve units represented the model's behaviour on the previous trial and 12 units represented the computer personality's behaviour on the previous trial. At the start of a cycle, the activations of these units were set to their appropriate values, the activation of 12 reinforcement units was set to 1, and the activation of all other units to zero. Activation passed through **W** to the model's behaviour units. Random noise was added to this activation, and the behaviour unit with the most activation was chosen as the response. The activation of this behaviour unit was set to 1, and the activation of the other behaviour units to zero. The computer personality's behaviour was calculated; the activation of the relevant CP unit was set to 1, that of the other CP units to zero. If the computer personality was on target, or one off target, the 12 reinforcement units were set to 1; otherwise four were set to -1 and the rest to zero (this reinforcement function appeared to be the best for learning Clegg; the precise reinforcement function made

less difference for Person's S and U). At this stage all the features were simultaneously represented: The model's behaviour on the current and previous trials, the computer personality's behaviour on the current and previous trials, and the reinforcement. The weights were then changed according to the delta rule. The model has the opportunity to learn not only the appropriateness of the behaviour and the situation it occurs in, but also predictive knowledge for all three computer personalities. The activations were then cleared and the cycle repeated for the next trial.

With SD zero or LR zero the model behaved in the same way as the previous auto associator. Also, more parameter tweaking was needed to obtain a respectable score for Clegg than for Person's S or U. With $SD=10^{-5}$ and $LR=10^{-4}$, the scores for Clegg were 9.7 (4.1) and 13.9 (5.3). The corresponding scores for Persons S and U were 7.9 (2.9) and 5.7 (3.2), respectively.

The model was assessed for how it functions as a look-up table, in the manner of Marescaux et al. (1989). as had been done for the Barto model and the previous auto associator model. After training for 60 trials with $SD = 10^{-5}$, $LR = 10^{-4}$, the model was presented with all possible situations defined by the combinations of the model's and Clegg's previous behaviour. The proportion of questions answered correctly was .78 (.31) for old situations that were followed by the target, or an adjacent response, in training (OldT), .16 (.14) for old situations not followed by the target (OldNT), and .23 (.05) for new situations (New). The concordance index was 90% (32%) for OldT, and 100% (0) for OldNT. The concordance index of the model was higher than that of subjects (57%, (35%)), $t(20)=2.31$, $p<.05$. The lower concordance rate found by Marescaux et al. (1989) may have been due to presenting subjects

with only partial situations; e.g., Clegg's previous behaviour but not the subject's. The model was trained as before, but tested with situations defined only by Clegg's previous behaviour. The proportion of questions answered correctly was .76 (.32) for OldT, .12 (.09) for OldNT, and .05 (.11) for New. The concordance index was 90% (32%) for OldT and 78% (31%) for OldNT.

The concordance index produced by this model (.90) for OldT was significantly different to the concordance index produced by subjects (.57). The concordance index produced by the first auto associator model, which coded context as persisting activation, (.79) was not significantly different to the concordance index produced by subjects. Thus, there is a suggestion that the first model fits the subject data better than the second model does. It should also be said that the high concordance index for the second auto associator does not reflect a sensitivity to different contexts. On the contrary, during training this model appeared to stick with a single response over a range of situations, changing it only after a few trials of consistent negative reinforcement. When given the tests of Marescaux et al., this second model gave a single response to all the presented situations (OldT, OldNT, and New). The first auto associator model, on the other hand, gave a range of different responses to the different situations. Unfortunately, Marescaux et al. (1989) do not report the range of responses given by a typical subject to different situations.

Predictive knowledge was assessed by presenting the model with situations consisting of: The model's behaviour on the previous trial, the resulting computer personality's behaviour on the previous trial, and the model's behaviour on the current trial. The model attempted to predict the resulting computer personality's

behaviour on the current trial by passing activation through **W**, as before. When the model was assessed with all possible situations after training with $SD=10^{-4}$ and $LR=10^{-4}$, the proportion of predictive knowledge questions correct was .37 (.14) for Clegg, close to the equivalent proportion obtained for subjects, .32 (Berry & Broadbent, 1984). Chance performance can be calculated as .16. The relatively greater level of control performance rather than predictive knowledge shown by the model is probably due to the greater number of associations that need to be learned in the latter rather than former case. In order to control Clegg, the model needs to learn mainly the correct actions for commonly occurring situations.

The proportions of predictive knowledge questions correct for Persons S and U were .31 (.13) and .47 (.13), respectively. Unfortunately, equivalent proportions for subjects are not available. Chance can be calculated as .22. Hayes and Broadbent (1988) assessed subjects' predictive knowledge of Person's S and U, but only after subjects had been exposed to a change in the rule governing the computer personality's behaviour. The proportions of predictive knowledge questions answered correctly by subjects for Persons S and U were then .80 and .10 (Hayes & Broadbent). Berry and Broadbent (1988) assessed predictive knowledge of subjects after interacting with Persons S and U for 60 trials; the proportions were .85 and .28, respectively. However, in this study, subjects maintained Person U on target for only 6 trials out of the last 20, a score that is close to chance. A group that had been informed of the relevance of their behaviour on the previous trial achieved a performance more in line with the model's: About 10 trials on target in the last 20. The predictive knowledge in this case was

.77, although perhaps this score reflects a different mode of learning (Berry & Broadbent, 1988). An experiment in which subjects are not exposed to conflicting rules and in which they achieve a level of performance comparable to that found by Hayes and Broadbent (1988) has not yet been done with Person U.

One fact appears clear from the results with subjects that is not mirrored by the results with the model: Predictive knowledge is greater for Person S than Person U (significantly so, $\underline{t}(18)=2.76$, $p<.01$). The results for the model go the other way round. The reason is that the model learns the correct behaviour for Person S in less trials than for Person U. Further, the more different actions the model tries out, and the more trials it does this over, the greater the opportunity to acquire the associations underlying predictive knowledge. Thus, more predictive knowledge is acquired by the model for Person U than S. Therefore, the difference in predictive knowledge acquired by subjects for Persons S and U would have to be accounted for by a separate learning mechanism operating in the case of Person S, as postulated by Berry and Broadbent (1984).

In summary, an auto associator that coded the behaviours of both the current and previous trials could learn all three person interaction tasks. When learning Clegg, the model acquired little predictive knowledge. When learning Person S, the model acquired less predictive knowledge than subjects.

Discussion

Chapter Six considered the usefulness of reinforcement learning (Barto, Sutton, & Brouwer, 1981) and the auto associator in modelling how subjects learn the control tasks.

Reinforcement learning was quite successful: It could learn Persons S and U (Berry & Broadbent, 1988; Hayes & Broadbent, 1988) and Clegg (Berry & Broadbent, 1988). Different parameter values were needed for Person U and Clegg, but this may not be a problem if parameter values could be adjusted on-line by the subject according to some heuristic (Broadbent, 1971, 1977). The weaknesses of the reinforcement learning algorithm as a model of control task learning lie in its arbitrary rather than emergent lack of predictive knowledge, and its lack of variability in responding to old situations that were followed by the target (Marescaux et al., 1989).

Two auto associator models were constructed, both of which could learn all three person interaction tasks. The first model coded behaviours on the previous trial not by distinct units but by persisting activation. This model could in principle only acquire predictive knowledge about Person S. The second model coded behaviours on both the current and previous trials with distinct units, and could in principle acquire predictive knowledge about all three computer personalities. In fact, it only acquired substantial predictive knowledge for Person U. The lack of predictive knowledge for Clegg is consistent with data obtained with subjects (Berry & Broadbent, 1984). The relative lack of predictive knowledge for Person S rather than Person U is not consistent with empirical data (e.g., Berry & Broadbent, 1988), and its explanation would require the introduction of an additional learning process for Person S (Berry & Broadbent, 1988). Further experimental work is needed to determine the predictive knowledge of subjects for Person U under the conditions modelled. A further result favored the first rather than the second auto associator: The first but not the second auto

associator gave the same response to old situations that were followed by the target a similar proportion of the time as subjects did (Marescaux et al., 1989).

The success of the auto associator in learning the control tasks as well as in artificial grammar learning suggests its general importance as a model of implicit learning. This has implications, that will be discussed in Chapter Seven, for understanding the relationship between artificial grammar learning and the control tasks. The results of this chapter and Chapter Five show that the mechanisms underlying learning the control tasks and artificial grammar learning could, in principle, be very similar; that is, like a delta rule auto associator. Three additional processes were used in applying the auto associator to the control tasks rather than artificial grammar learning: The addition of noise activation to some units, the selection of one unit rather than another to produce as a behaviour, and the coding of reinforcement. These processes may make the control tasks seem a more active task to subjects than artificial grammar learning. But in both cases the actual learning mechanism (in the model) involved simple association formation according to the delta rule.

Chapter Seven

Summary and conclusions

Introduction

This chapter summarizes the major points of the thesis and draws out the implications of those points for an understanding of implicit learning. First, the conceptual framework and background of the thesis are considered. Second, the empirical research conducted in this thesis are summarized. Third, the results of the computational modelling are described. Finally, the overall implications for implicit learning are discussed. Throughout the chapter some ambiguities in the present research are acknowledged, and specific recommendations are made for future research.

Conceptual framework and background

According to the framework adopted by this thesis, and outlined in Chapter One, implicit knowledge can be regarded as a distinct database that shows itself by two characteristics. The first characteristic is that implicit rather than explicit knowledge shows specificity of transfer. The exact nature of the specificity of the tasks is a matter for empirical investigation. The second characteristic is that implicit as compared to explicit knowledge is of a fundamentally different type. That is, there should be evidence suggesting distinct principles of storage and retrieval for tasks showing specificity of transfer.

Evidence for both characteristics would suggest a distinct knowledge base underlying an experimental task only if it were possible to rule out simple alternative interpretations. For example, if evidence of specificity of transfer from a task could be interpreted in terms of the degree to which low confidence knowledge

is elicited, there would be little evidence for a distinct knowledge base underlying the task. If such reinterpretations are not ruled out, the investigation would indicate that an experimental task is learned "implicitly" only at an everyday level of explanation.

Chapter One reviewed empirical studies in the concept formation literature. The review showed that specificity of transfer had been demonstrated mainly by contrasting classification performance, which might force the use of low confidence knowledge, with free recall, which would not. Thus, these previous studies do not provide strong evidence for a distinct knowledge base underlying classification performance. One noticable exception is Hayes (1987), who showed a failure of classification performance to transfer to recognition of isolated elements of the exemplars. This restriction of classification knowledge to the recognition of whole exemplars also appeared to be a possible restriction on the conceptual task underlying artificial grammar learning. Unfortunately, Hayes (1987) had failed to test for the transfer of all plausible knowledge, and so his results were ambiguous. In summary, Chapter One established that in order to demonstrate implicit learning with concept formation tasks, a more rigorous testing of the specificity of transfer of classification knowledge was needed.

Summary of empirical research

Chapter Two presented the results of two experiments that extended previous methodology to provide more complete transfer tests of the knowledge underlying classification. Experiment One employed the Residents Task of Hayes (1987), and extended his transfer task so that it was more complete. Subjects classified

"computer people" into one of three towns. Each computer person was described by four phrases, including one unique phrase. After reaching a performance criterion, the transfer task of Hayes (1987) was given in which subjects were asked to recognize the unique phrases and to indicate the town associated with each. There was also an additional transfer test in which subjects were asked to recognize unique combinations of shared phrases and to indicate the town associated with each. The main result was that knowledge of unique phrases was sufficient to account for classification performance. A reconsideration of the data from Hayes (1987), taking into account the probabilities of performing on the classification and transfer tasks by chance alone, indicated that his results were also consistent with this conclusion.

A problem with the Residents Task, even with the extended transfer task, is that there are different probabilities for performing well on the classification and the transfer tests by chance alone. Experiment Two was a first attempt to test for implicit concept formation by appropriately adjusting these probabilities. Additionally, subjects were exposed to the information incidentally to discourage an explicit mode of learning. The stimuli consisted of twelve computer people, each described by four phrases and living in one of three towns. One phrase was unique to a town (rather than to a person) and the other phrases were counterbalanced across the towns. In the transfer test, subjects underlined the important sentence(s) for each town and indicated the rule connecting the sentence(s) to the town. Thus, like the Hayes (1987) transfer test, this task tests for the application of classification knowledge to elements of the exemplars in isolation. The results showed that subjects performed on the

classification and transfer tests at a chance level.

Chapter Three reported three experiments that continued to probe for the possibility of implicit learning in simple concept formation tasks by addressing possible problems with Experiment Two. The task used in Experiment Two may not have induced implicit learning because of the way the stimuli were displayed, because of the underlying rule to be learned, or because the cover task biased subjects to process the stimuli at an inappropriate level. Thus, the tasks used in Chapter Three addressed the above three points: The stimuli were presented as continuous variables; a rule was introduced that involved a couple of unique sentences (rather than just one); and different cover tasks were used, depending on the experiment. Experiment Three employed a partial report task. Like Experiment Two, this procedure results in chance classification and transfer test performance. In Experiments Four and Five, subjects memorized which income went with which person. Experiment Four was exploratory and had unequal probabilities for performing well on the classification and transfer tasks by chance alone. When the underlying rule involved a single unique phrase per category, performance on the classification and transfer tasks was good and equivalent on both. When the rule involved two phrases determining category membership, performance was better on the classification than the transfer task. Experiment Five introduced additional classification blocks to equalize baseline probabilities on the classification and transfer tasks; when this was done, performance on the classification and transfer tasks was equivalent. Thus, the apparent lack of transfer in Experiment Four was artifactual.

To summarize, Chapters Two and Three presented the results from five experiments that indicated an equivalence between

classification performance and a structured transfer test. These results may indicate the absence of an effect to be found or the failure to establish the necessary conditions to find implicit learning. Rather than continuing to probe this issue with new paradigms, Chapter Four returned to a paradigm in the concept formation literature where it was claimed that the necessary conditions for implicit learning had been established. Specifically, the artificial grammar learning paradigm of Reber (e.g., 1976) was employed for Experiments Six, Seven, and Eight reported in Chapter Four. Reber (e.g., 1988) claimed that this task is learned implicitly, but it had not been demonstrated that classification did not transfer to a structured knowledge test. Thus, in Chapter Four a structured transfer test similar to those used in Chapters Two and Three was applied to the artificial grammar learning task.

In the artificial grammar learning task, subjects were first asked to memorize strings of letters generated by a finite state grammar, and then were asked to classify grammatical and nongrammatical strings. In Chapter Four, two transfer tasks were used: Free recall, a distinctively explicit knowledge test, in which the subject was asked to describe as completely as possible how she actually classified; and the test of Sequential Letter Dependencies (the SLD test), the structured transfer test in which the subject was presented with allowable initial letter sequences, of length 0 letters upwards, and asked to indicate which letters could occur next. The SLD test is similar to the Hayes (1987) transfer test and to the transfer test used in Experiments Two to Five in that it assesses the subject's ability to apply her knowledge to elements of exemplars presented in isolation. In

Experiment Six, classification knowledge transfered to the SLD test, but not to free recall. This is not suprising: Free recall is not expected to be a sensitive test. More importantly, free recall failed to correlate with either classification or SLD, which did correlate with each other to a small but significant extent. This result provides tentative evidence for distinct knowledge bases underlying free recall on the one hand, and classification and SLD performance on the other.

Experiment Seven attempted to strengthen the case for separate knowledge bases by employing a dual task methodology used by Hayes (1987; and as reported in Broadbent, 1989). Hayes found that concurrent random number generation (RNG) did not interfere with learning the artificial grammar (as measured by classification) when incidental learning instructions were used; but concurrent RNG did interfere if subjects were asked to search for rules during learning. Thus, concurrent RNG may interfere with an explicit but not implicit mode of learning. Thus, one interesting possibility is that under incidental instructions, RNG may interfere with free recall, indexing explicit knowledge, but not classification and SLD, potentially indexing implicit knowledge. In fact, Experiment Seven showed that with concurrent RNG <u>all</u> knowledge tests suffered. This result raised the possibility of different task priorities (as a function of different task demands) accounting for the interference found by Hayes for explicitly but not implicitly instructed subjects. In Experiment Eight, task priorities were systematically manipulated for both implicitly and explicitly instructed subjects; and classification, Free Recall, and SLD measures of artificial grammar learning were taken. The results showed dual task conditions interfered with all knowledge measures for both

implicitly and explicitly instructed subjects. The priority manipulation per se had no influence on classification performance. Further, the results indicated that Hayes (1987) may have failed to find an effect of dual task conditions on implicitly instructed subjects simply because his study lacked power.

In summary, Chapter Four showed that classification knowledge in the artificial grammar learning task could transfer to recognition judgements of part exemplars, and that there was no evidence that it represented a different type of knowledge to that elicited by free recall. Nonetheless, in that classification knowledge failed to transfer to free recall, it could be regarded as "implicit" at an everyday level of explanation.

Computational modelling

In order to investigate the sorts of mechanisms that could learn the tasks used in Experiments Six to Eight, the next two chapters turned to computational modelling. Chapter Five investigated the performance of different auto associators and exemplar models in learning the artificial grammar used in Experiments Six to Eight. An auto associator is a mechanism that attempts to predict each part of a presented pattern according to the remainder of the pattern. An auto associator can classify grammatical and nongrammatical strings according to how well each part of a string is completed. A range of auto associators were considered differing along four dimensions: The learning rule used (Hebb or Delta); letter vs digram coding; sensitivity to cooccurrence or contingency; and simultaneous vs successive prediction. Two exemplar models were considered: The MINERVA 2 model of Hintzman (1986) and the array model of Estes (1986). Both

exemplar models involved determining an overall similarity between each test string and the stored exemplars. Each model was used to produce a rank ordering of string difficulty, and this was compared to experimental rank orderings (derived from Experiments Six to Eight and from Dulany et al., 1984). The results showed that the simultaneous delta rule auto associator models fare best in accounting for the empirical data. These models passed a number of tests failed by the other models. Further, the knowledge used by these models for classification performance did transfer to the SLD test. It was shown that these models could be regarded as abstracting a set of representative but incomplete rules of the grammar. The necessary and sufficient conditions to be satisfied by a finite state grammar for the models to abstract representative rules were specified. Future research could investigate subject performance on grammars for which the models abstract nonrepresentative rules.

In order to determine the generality of auto association in providing a possible mechanism for implicit learning, Chapter Six investigated the performance of different Connectionist models, including auto association, in learning the dynamic control tasks. Two Connectionist architectures were considered: Reinforcement learning (e.g., Barto, Sutton, & Brouwer, 1981), and auto association. It was shown that reinforcement learning could learn the dynamic control tasks. Such a system, like subjects, has poor predictive knowledge, but that is because it is hard-wired that way. A more interesting architecture is the auto associator, where all units are connected to all others. Thus, it is logically possible for the auto associator to learn not only what action to produce in a given context (performance), but also what context results from a

given action (predictive knowledge). In fact, it was shown that when the dynamic control tasks are simulated by an auto associator learning by the delta rule, the model could learn to perform with little predictive knowledge when Clegg was used but not when Person U was used. Subjects' lack of predicitive knowledge for Clegg is well documented (Berry, 1990; Berry & Broadbent, 1984; Sanderson, 1989; Squire & Frambach, 1990), but for Person U it has been shown only when subjects show poor control performance (Berry & Broadbent, 1988) or have been exposed to contradictory relationships (Hayes & Broadbent, 1988). Future research needs to establish the extent of subjects' predictive knowledge for Person U under a wider range of experimental conditions.

Implications

A major aim of this thesis was to develop concept formation tasks to test the specificity of the knowledge underlying classification performance. A review of the literature indicated that a possible source of specificity in classification knowledge was its failure to transfer, under certain conditions, to recognition of elements of exemplars in isolation. Eight experiments showed that there was transfer to isolated elements in a concept formation paradigm claimed to rely on knowledge applicable only to whole exemplars (Hayes, 1987), in a paradigm claimed to rely on implicit knowledge (Reber, 1967), and in several new paradigms.

There may well be other sources of specificity that would entitle classification knowledge to be regarded as genuinely implicit under certain circumstances. As the evidence stands, however, the strongest that can be said is that some classification tasks, such as artificial grammar learning, can be regarded as

"implicit" at an everyday level of explanation. The evidence for specificity in the control tasks is more convincing (see Chapter One). Thus, whether there may be another source of specificity in concept formation may be best addressed by considering the relationship between concept formation and the control tasks, and the source of specificity in the control tasks.

Chapters Five and Six demonstrated that a delta rule auto associator could learn both the control tasks and to categorize strings defined by a finite state grammar in a similar manner to subjects. Further, both the failure of subjects after learning Clegg to transfer to a predictive knowledge test and the success of subjects after learning the finite state grammar to transfer to the SLD test could be modelled by the auto associator. If the control tasks are assumed to be learned implicitly, this result shows that the artificial grammar could have been learned by the same implicit processes. Thus, it may not be suprising that classification knowledge did transfer to isolated elements of exemplars, even if classification knowledge was genuinely implicit. So where would the specificity lie for classification knowledge?

One source of specificity in the knowledge underlying the control tasks is its failure to transfer when there are changes in the perceptual embodiment of a task. Berry and Broadbent (1988) found no transfer of performance between two control tasks embodying the same rule but different cover stories, and negative transfer if subjects were informed of the mapping between the tasks. The failure of transfer between different cover stories has recently been replicated by Squire and Frambach (1990) for both normal and amnesic subjects. Berry (in preparation) found positive transfer of performance across two blocks of the Clegg task if the subjects

typed their responses on both blocks and not if the subjects spoke their responses on the first block and typed them on the second.

Two other studies have found similar effects of perceptual embodiment on procedural tasks (Stadler, 1989; Willingham et al., 1989). Stadler replicated the finding of Lewicki, Czyzewska, and Hoffman (1987) that reaction time decreased in a visual search task when the general location of a target could be predicted by a set of rules that subjects could not later report. Stadler showed further that the knowledge did not transfer when the precise location of the target was changed, although the general location was still predictable. Willingham et al. demonstrated that when subjects had learned a repeating spatial sequence on a serial reaction time task, there was no transfer of performance if the perceptual characteristics (colours) of the display were changed.

The effects of changes in perceptual embodiment in concept formation tasks needs systematic investigation. Reber (1969, 1989) has argued that the knowledge underlying artificial grammar learning is not specific in this way because the knowledge transfers to different letter sets embodying the same grammar. This finding has been replicated by Mathews et al. (1989). It shows that identical visual information during acquisition and testing is not necessary to access the classification knowledge base. However, it is likely that strings are at least partly processed in terms of an articulatory code. It is possible that subjects articulatorily recode strings in the new letter set in terms of the old letter set and that this constant articulatory coding sustains transfer. If both visual and articulatory codes could be rendered inapplicable, more specific transfer effects might be found. Indeed, certain results from Experiment Eight are consistent with this

interpretation. In Experiment Eight it was found that concurrent RNG interfered with artificial grammar learning, but this interference was independent of the priority assigned to the tasks. This result suggests that there was some common resource taken up by RNG in an all-or-none fashion. A plausible resource is the articulatory loop. Future research could usefully determine if concurrent articulatory suppression interferes with artificial grammar learning and abolishes the transfer to different letter sets.

McGeorge and Burton (in press) also argued that implicit classification knowledge is not perceptually based. In their experiments, subjects were initially exposed to 30 four-digit numbers while performing an incidental cover task (e.g., counting the number of horizontal lines). Each number contained at least one "3". Subjects later classified as "old" more numbers that contained one "3" rather than no "3". McGeorge and Burton argued that the knowledge was semantic rather than perceptual because the performance transfered to numbers written as words rather than digits. Again, it seems likely that there would be a constant articulatory coding of the stimuli, and this may have sustained transfer of the classification knowledge. Future research could determine if articulatory suppression abolishes transfer on this task.

The possible dependence of concept formation and control tasks on perceptual characteristics might provide a point of convergence between implicit learning and implicit memory. Tests of implicit memory are defined by not requiring deliberate recollection of a prior event, whereas tests of explicit memory are defined by requiring recollection (Lewandowski, Kirsner, & Dunn, 1990;

Schacter, 1987). Examples of implicit memory tasks are perceptual identification, lexical decision, or fragment completion. Explicit memory tasks are recall and recognition. Typically, implicit and explicit memory tasks are stochastically independent (e.g., Hayman & Tulving, 1989b). An important finding has been that study-test modality shifts can reduce or eliminate learning on implicit but not explicit memory tasks (see e.g. Bassili, Smith, & MacLeod, 1989). Thus, transfer of the knowledge underlying implicit memory is dependent on the perceptual characteristics of the task.

Chapter Four indicated a possible analogy between artificial grammar learning and implicit memory which reflected the variability with which same and different cues access a common knowledge base. Specifically, Experiments Six to Eight found that there was a substantial correlation between probability correct on successive classifications of each exemplar, but a negligible correlation between classification of an exemplar and the same knowledge elicited by appropriate SLD stems. The first and second correlations can be regarded, respectively, as (inversely) referring to the variability with which identical and different cues access the knowledge base. This difference in variability for identical and different cues is closely paralleled by results in the primed fragment completion paradigm obtained by Hayman and Tulving (1989a). They found that successive tests of the same fragment were highly stochastically dependent, but successive tests of different fragments of the same word were virtually stochastically independent. They suggested that their results could be explained if fragment completion relied on a "traceless" procedural memory system that learned by

"refining, modifying, or optimizing the activity or responding dictated by a perceptually present stimulus and the state

of the system (p. 953)".

Just such a system - the delta rule auto associator - was found in Chapter Five to be a good model of artificial grammar learning.

Two further results point to a correspondence between implicit learning and implicit memory. First, note that implicit learning involves the formation of new associations. Schacter (e.g., Graf & Schacter, 1989; Schacter & Graf, 1989; Schacter & McGlynn, 1989) has argued that implicit memory paradigms can also reveal the formation of new associations. If subjects study unrelated word pairs, they are more likely to evince learning on a later stem completion task if the stem is presented in the context of the same word it was paired with rather than a different one. This associative effect is reduced by a study-test modality shift (Schacter & Graf, 1989).

Second, the correspondence between implicit learning and memory is further highlighted by the results of Marescaux et al. (1989) discussed in Chapters One and Six, and modelled in Chapter Six. Marescaux et al., employing the Clegg task, found that subjects tended to respond in the same way to situations previously encountered if the previous response was rewarded. Thus, the Clegg task can be seen as a form of implicit memory task. Whether the same processes are involved in learning Clegg as in learning new associations in traditional implicit memory tasks could be investigated by exploring both the Clegg paradigm and the traditional implicit memory paradigm.

In terms of the Clegg paradigm, research needs to investigate whether the tendency to respond consistently to rewarded

situations is independent of the subjects' ability to recall the response to the situation and whether it is reduced by modality shifts. The dependence of the subjects' tendency to respond consistently on the subjects' ability to recall the response in that situation would be important in assessing whether Clegg is learned unconsciously. If consistent responding was dependent on recall, and subjects could justify their responding in terms of the episodic memory, the claim for unconscious learning (Hayes, 1987; Stanley et al., 1989) would need reconsidering. In terms of the effect of modality shifts, an important complementary investigation would be to determine whether transfer of performance on the control tasks under modality shifts occurs when a task relying on explicit knowledge is used. This result would be needed to assess whether transfer (or the lack of it) is a property of the putative implicit processes or a property of the control tasks in general.

In terms of the implicit memory paradigm, research needs to investigate whether implicit memory for new associations can occur selectively for rewarded stimuli, and the dependence of such selective learning on free recall ability and modality shifts. The implicit memory for new associations paradigm could also be used to explore implicit concept formation. If during study, a class of words was associated with a particular class of context, would subjects complete word stems according to the appropriate class when the appropriate context was present, and would this tendency be abolished by modality shifts?

If classification knowledge was found, under certain circumstances, to be modality-specific, as revealed by artificial grammar learning, implicit memory, or other paradigms, further research to establish distinctive principles of storage or retrieval

would be necessary to establish the claim for implicit concept formation fully. This thesis attempted to provide evidence for distinctive principles by the use of a secondary task. Experiment Eight showed that, with the artificial grammar learning paradigm, implicitly and explicitly instructed subjects were affected equally, and so were different measures of possibly different knowledge types. Future research might profitably explore the use of different types of secondary task. A different tack is suggested by the results of Graf and Schacter (1987). They found that implicit but not explicit memory for new associations was unaffected by pro- and retroactive interference manipulations. Future research might usefully explore the use of interference manipulations in concept formation paradigms.

Concluding note

The research reported in this thesis has demonstrated that the knowledge underlying classification performance is not specific in a way previously thought plausible, nor is it spared from the effects of secondary tasks. This thesis has demonstrated the value of computational modelling of implicit learning paradigms, and provided tentative computational models of artificial grammar learning and learning the control tasks. Finally, the thesis has pointed to a number of research directions that could reveal possible specificity in classfication knowledge and illuminate the nature of implicit knowledge in general.

REFERENCES

Abrams, M., & Reber, A. S. (1988). Implicit learning: Robustness in theface of psychiatric disorders. Journal of Psycholinguistic Research, 17, 425-439.

Allen, R., & Reber, A. S. (1980). Very long term memory for tacit knowledge. Cognition, 8, 175-185.

Allport, A. (1989). What concept of consciousness? In A. J. Marcel & E. Bisiach (Eds), Consciousness in contemporary science (pp 159-182). Oxford: Clarendon Press.

Anastasi, A. (1976). Psychological testing, 4th edition. New York: MacMillan.

Anderson, J. A. (1983). Cognitive and psychological computation with neural models. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 799-815.

Anderson, J. R. (1983). The architecture of cognition. Cambridge, Ma.: Harvard University Press.

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

Attneave, F. (1959). Applications of information theory to psychology. New York: Holt.

Baddeley, A. D. (1966). The capacity for generating information by randomisation. Quarterly Journal of Experimental Psychology, 18, 119-129.

Baddeley, A. D. (1986). Working memory. Oxford: Clarendon Press.

Barto, A. G., Anderson, C. W., & Sutton, R. S. (1982). Synthesis of nonlinear control surfaces by a layered associative search network. Biological Cybernetics, 43, 175-185.

Barto, A. G., & Sutton, R. S. (1981). Landmark learning: An illustration of associative search. Biological Cybernetics, 42, 1-8.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 834-846.

Barto, A. G., Sutton, R. S., & Brouwer, P. S. (1981). Associative search network: A reinforcement learning associative memory. Biological Cybernetics, 40, 201-211.

Bassili, J. N., Smith, M. C., & MacLeod, C. M. (1989). Auditory and visual word-stem completion: Separating data driven and conceptually driven processes. Quarterly Journal of Experimental Psychology, 41, 439-453.

Bekerian, D. A., & Bowers, J. M. (1983). Eyewitness testimony:

Were we mislead? Journal of Experimental Psychology: Learning, Memory, and Cognition, 9, 139-145.

Berry, D. C. (in preparation). Learning while we watch: The role of action in controlling complex systems.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. Quarterly Journal of Experimental Psychology, 36, 209-231.

Berry, D. C., & Broadbent, D. E. (1987). The combination of explicit and implicit learning processes in task control. Psychological Research, 49, 7-15.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. British Journal of Psychology, 79, 251-272.

Bowers, K. S. (1984). On being unconsciously inluenced and informed. In K. S. Bowers & D. Meichenbaum (Eds), The unconscious reconsidered (pp 227-271). New York: Wiley.

Brewer, W. F. (1974). There is no convincing evidence for operant or classical conditioning in adult humans. In W. B. Weimer & D. S. Palermo (Eds), Cognition and the symbolic processes (pp 1-42). Hillsdale, N.J.: Erlbaum.

Broadbent, D. E. (1971). Decision and stress. New York: Academic Press.

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. Quarterly Journal of Experimental Psychology, 29, 181-201.

Broadbent, D. E. (1980). The minimization of models. In A. J. Chapman & D. M. Jones (Eds), Models of man (pp 113-128). Leicester: The British Psychological Society.

Broadbent, D. E. (1989). Lasting represntations and temporary processes. In Roediger III, H. L. and Craik, F. I. M. (Eds), Varieties of memory and consciousness: Essays in honor of Endel Tulving (pp 211-227). Hillsdale, N.J.: Erlbaum.

Broadbent, D. E., & Aston, B. (1978). Human control of a simulated economic system. Ergonomics, 21, 1035-1043.

Broadbent, D. E., & Broadbent, M. H. P. (1977). Effects of recognition on subsequent recall: Comments on "Determinants of recognition and recall: Accessibility and generation" by Rabinowitz, Mandler, and Patterson. Journal of Experimental Psychology: General, 106, 330-335.

Broadbent, D. E., Fitzgerald, P., & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. British Journal of Psychology, 77, 33-50.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds), Cognition and categorization (pp 169-211). Hillsdale, N.J.: Erlbaum.

Brunswick, E. (1956). Perception and the representative design of psychological experiments. Berkeley, CA: University of California Press.

Brunswick, E., & Herma, H. (1951). Probability learning of perceptual cues in the establishment of a weight illusion. Journal of Experimental Psychology, 41, 281-290.

Carlson, R. A., & Dulany, D. E. (1985). Conscious attention and abstraction in concept learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 45-58.

Child, D. (1970). The essentials of factor analysis. London: Holt, Rinehart, & Winston.

Cohen, N. J. (1984). Preserved learning capacity in amnesia: Evidence for multiple memory systems. In L. R. Squire & N. Butters (Eds), Neuropsychology of memory (pp 83-103). New York: Guilford Press.

Cormier, S. M., & Hagman, J. D. (Eds) (1987). Transfer of learning: Contemporary research and application. Orlando, FL: Academic Press.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104, 268-294.

Davis, R., Sutherland, N., & Judd, B. (1961). Information content in recognition and recall. Journal of Experimental Psychology, 61, 422-429.

Dulany, D. E. (1962). The place of hypotheses and intentions: An analysis of verbal control in verbal conditioning. In C. W. Eriksen (Ed.), Behaviour and awareness (pp 102-129). Durham, NC: Duke University Press.

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgement: How conscious and how abstract? Journal of Experimental Psychology: General, 113, 541-555.

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1985). On consciousness in syntactic learning and judgement: A reply to Reber, Allen, and Regan. Journal of Experimental Psychology: General, 114, 25-32.

Dulany, D. E., & O'Connell, D. C. (1963). Does partial reinforcement dissociate verbal rules and the behaviour they might be presumed to control? Journal of Verbal Learning and Verbal Behaviour, 2, 361-372.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. Psychological Review, 95, 91-101.

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

397-417.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. Cognitive Psychology, 6, 159-171.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, Ma.: MIT Press.

Eriksen, C. W. (1962). Figments, fantasies, and follies: A search for the subconscious mind. In C. W. Eriksen (Ed.), Behavior and awareness (pp 3-26). Durham, NC: Duke University Press.

Estes, W. K. (1986). Memory storage and retreival processes in category learning. Journal of Experimental Psychology: General, 115, 155-174.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retreival models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 556-571.

Fitzgerald, P., Tatersall, A., & Broadbent, D. E. (1988). Seperating central mechanisms by POCs: Evidence for an imput-output buffer. Quarterly Journal of Experimental Psychology, 40, 109-134.

Flint, C. R. (1979). The role of consciousness in memory. Unpublished D. Phil thesis, University of Oxford.

Fodor, J. A. (1983). The modularity of mind. Cambridge, Ma.: MIT Press.

Fodor, J. A. (1985). Precis of the modularity of mind. Behavioural and Brain Sciences, 8, 1-42.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. Journal of Memory and Language, 27, 166-195.

Graf, P., & Schacter, D. L. (1987). Selective effects of interference on implicit and explicit memory for new associations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 45-53.

Graf, P., & Schacter, D. L. (1989). Unitization and grouping mediate dissociations in memory for new associations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 930-940.

Graham, C., & Evans, F. J. (1977). Hypnotizability and the deployment of waking attention. Journal of Abnormal Psychology, 86, 631-638.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgement. IEEE Transactions on Systems, Man, and Cybernetics, SMC-17, 753-770.

Hanfmann, E. (1941). A study of personal patterns in an intellectual performance. Character and Personality, 9, 315-325.

Hartman. M., Knopman, D. S., & Nissen, M. J. (1989). Implicit learning of new verbal associations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1070-1082.

Hayes, N. A. (1987). Systems of explicit and implicit learning. Unpublished D. Phil thesis, University of Oxford.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. Cognition, 28, 249-276.

Hayman, C., & Tulving, E. (1989a). Contingent dissociation between recognition and fragment completion: The method of triangulation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 228-240.

Hayman, C., & Tulving, E. (1989b). Is priming in fragment completion based on a "traceless" memory system? Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 941-956.

Hebb, D. O. (1949). The organization of behaviour. New York: Wiley.

Heidbreder, E. (1934). A study of the evolution of concepts. Psychological Bulletin, 31, 673. (Abstract)

Heidbreder, E. (1936). Language and concepts. Psychological Bulletin, 33, 724. (Abstract)

Hinton, G. E. (1987). Connectionist learning procedures. Technical Report CMV-CS-87-115, Computer Science Department, Carnegie-Mellon University.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple trace memory model. Psychological Review, 93, 411-428.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. Annual Review of Psychology, 41, 109-139.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioural and Brain Sciences, 9, 1-66.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. Journal of Experimental Psychology: Human, Learning, and Memory, 2, 322-330.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts: An experimental study. Psychological Monographs, 28, (whole issue, no. 123).

James, W. (1950/1890). The principles of psychology, vol. I. New York: Dover Publications.

Kellogg, R. T. (1982). Hypothesis recognition failure in conjunctive and disjunctive concept-identification tasks. Bulletin of the Psychonomic Society, 19, 327-330.

Kellogg, R. T., Robbins, D. W., & Bourne, L. E., Jr (1978). Memory for intratrial events in feature identification. Journal of Experimental Psychology: Human, Learning, and Memory, 4, 256-265.

Kellogg, R. T., Robbins, D. W., & Bourne, L. E., Jr (1983). Failure to recognize previous hypotheses during concept learning. American Journal of Psychology, 96, 179-199.

Kemler-Nelson, D. G. (1984). The effect of intention on what concepts are acquired. Journal of Verbal Learning and Verbal Behaviour, 23, 734-759.

Kemler-Nelson, D. G. (1988). When category learning is holistic; A reply to Ward and Scott. Memory & Cognition, 16, 79-84.

Kolers, P. A., & Roediger III, H. L. (1984). Procedures of mind. Journal of Verbal Learning and Verbal Behaviour, 23, 425-449.

Lewandowski, S., Dunn, J. C., Kirsner, K. (Eds) (1989). Implicit memory: Theoretical issues. Hillsdale, N.J.: Erlbaum.

Lewicki, P. (1986). Nonconscious social information processing. New York: Academic Press.

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 523-530.

Lewicki, P., Hill, T., & Sasaki, I. (1989). Self-perpetuating development of encoding biases. Journal of Experimental Psychology: General, 118, 323-337.

Marescaux, P-J, Luc, F., & Karnas, G. (1989). Modes d'apprentissage selectif et nonselectif et connaissances acquises au control d'un processes: Evaluation d'un modele simule. Cahiers de Psychologie Cognitive, 9, 239.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). Journal of Experimental Psychology: General, 118, 417-421.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.

McAndrewsm, M. P., & Moscovitch, M. (1985). Rule-based and exemplar-based classification in artificial grammar learning. Memory and Cogntion, 13, 469-475.

McClelland, J. L., & Elman, J. (1986). Interactive processes in speech perception: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds), Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and

biological models (pp 58-121). Cambridge, Ma.: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-188.

McClelland, J. L., & Rumelhart, D. E. (Eds) (1986a). Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models. Cambridge, Ma.: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1986b). A distributed model of human learning and memory. In J. L. McClelland & D. E. Rumelhart (Eds), Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models (170-215). Cambridge, Ma.: MIT Press.

McGeorge, P., & Burton, A. (1989). The effects of concurrent verbalization on performance in a dynamic systems task. British Journal of Psychology, 80, 455-465.

McGeorge, P., & Burton, A. (in press). Semantic processing in an incidental learning task. Quarterly Journal of Experimental Psychology,

Mackintosh, N. J. (1974). The psychology of animal learning. London: Academic Press.

Mackintosh, N. J. (1983). Conditioning and associative learning. Oxford: Clarendon Press.

McNemar, Q. (1969). Psychological statistics, 4th edition. New York: Wiley.

McNicol, D. (1972). A primer of signal detection theory. London: George Allen & Unwin.

Massaro, D. W. (1988). Some criticisms of Connectionist models of human performance. Journal of Memory and Language, 27, 213-234.

Mathews, R. C., Buss, R. R., Chinn, R., & Stanley, W. B. (1988). The role of explicit and implicit learning processes in concept discovery. Quarterly Journal of Experimental Psychology, 40, 135-165.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J-R, & Druhan, B. (1989). The role of implicit and explicit processes in learning from examples: A synergistic effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1083-1100.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 37-50.

Medin, D. L., Dewey, G. I., & Murphey, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is

not automatic. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9, 607-625.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. Journal of Experimental Psychology: General, 117, 68-85.

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

Medin, D. L., & Schwanenflugal, P. J. (1981). Linear seperability in classification learning. Journal of Experimental Psychology: Human, Learning, and Memory, 7, 355-368.

Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), Information theory in psychology (pp 95-100). Glencoe, Illinois: Free Press.

Navon, D. (1984). Resources - a theoretical soup stone? Psychological Review, 91, 216-234.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. Psychological Review, 86, 214-255.

Navon, D., & Gopher, D. (1980). Task difficulty, resources, and dual task performance. In R. S. Nickerson (Ed.), Attention and performance VIII (pp297-315). Hillsdale, N.J.: Erlbaum.

Nelson, T. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. Journal of Experimental Psychology: Human, Learning, and Memory, 4, 453-468.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. Cognitive psychology, 7, 44-64.

Norman, D. A., & Bobrow, D. G. (1976). On the analysis of performance operating characteristics. Psychological Review, 83, 508-510.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 282-304.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? Journal of Experimental Psychology: General.

Phelan, J. G. (1965). A replication of a study on the effects of attmepts to verbalize on the process of concept attainment. Journal of Psychology, 59, 283-293.

Polanyi, M. (1969). Knowing and being. Chicago, IL: University of Chicago Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Posner, M. I., & Keele, S. W. (1968). Retention of abstract ideas. Journal of Experimental Psychology, 83, 304-308.

Putnam, H. (1981). Reason, truth, and history. Cambridge: Cambridge University Press.

Reber, A. S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behaviour, 6, 855-863.

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. Journal of Experimental Psychology, 81, 115-119.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. Journal of Experimental Psychology: Human, Learning, and Memory, 2, 88-94.

Reber, A. S. (1989). Implicit learning and tacit knowledge. Journal of Experimental Psychology: General, 118, 219-235.

Reber, A. S., & Allen, R. (1978). Analogic and abstraction strategies in synthetic grammar learning: A functionalist interpretation. Cognition, 6, 189-221.

Reber, A. S., Allen, R., & Regan, S. (1985). Syntactic learning and judgements: Still unconscious and still abstract. Journal of Experimental Psychology: General, 117, 17-24.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. Journal of Experimental Psychology: Human, Learning, and Memory, 6, 492-502.

Reber, A. S., & Lewis, S. (1977). Implicit learning: An analysis of the form and structure of a body of tacit knowledge. Cognition, 5, 333-361.

Rescorla, R. A., & Wagner, A. D. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds), Classical conditioning II: Current research and theory (pp 64-99). New York: Appleton-Century-Crofts.

Restle, F. (1962). The selection of strategies in cue learning. Psychological Review, 69, 329-343.

Roediger, H. L., III & Blaxton, T. A. (1987). Effects of varying modality, surface features, and retention interval on priming in word fragment completion. Memory & Cognition, 15, 379-388.

Rommetveit, R. (1960). Selectivity, intuition, and halo effects in social perception. Oslo: Oslo University Press.

Rommetveit, R. (1960). Stages in concept formation and levels of cognitive functioning. Scandanavian Journal of Psychology, 1, 115-124.

Rommetveit, R., & Kvale, S. (1965). Stages in concept formation. IV. Scandanavian Journal of Psychology, 6, 75-79.

Rumelhart, D. E., & Hinton, G. E., & Williams, R. J. (1986). Learning internal repersentations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds), Parallel distributed processing: Explorations in the microstructure of cognition. Vol. I: Foundations (pp 318-362). Cambridge, Ma.: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds) (1986a). Parallel distributed processing: Explorations in the microstructure of cognition. Vol. I: Foundations. Cambridge, Ma.: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986b). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds), Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models (pp 216-271). Cambridge, Ma.: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart & J. L. McClelland (Eds), Parallel distributed processing: Explorations in the microstructure of cognition. Vol. I: Foundations (pp 151-193). Cambridge, Ma.: MIT Press.

Ryle, G. (1949). The concept of mind. San Francisco: Hutchinson.

Sanderson, P. M. (1989). Verbalizable knowledge and skilled task performance: Association, dissociation, and mental models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 729-747.

Schacter, D. L. (1987). Implicit memory: History and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 501-518.

Schacter, D. L., & Graf, P. (1986). Effects of elaborative processing on implicit and explicit memory for new associations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 432-444.

Schacter, D. L., & Graf, P. (1989). Modality specificity of implicit memory for new associations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 3-12.

Schacter, D. L., & McGlynn, S. (1989). Implicit memory: Effects of elaboration depend on unitization. American Journal of Psychology, 102, 151-181.

Schneider, W. (1987). Connectionism: Is it a paradigm shift for psychology? Behavioural Research Methods, Instruments, and Computing, 19, 73-83.

Schneider, W., Dumais, S. T., & Schiffrin, R. M. (1984). automatic and control processing and attention. In R. Parasuraman, & R. Davies (Eds), Varieties of attention (pp 1-28). New York: Academic Press.

Schroth, M. L. (1987). Memory for intratrial events in learning disjunctive concepts: Implications for hypothesis-testing models. Journal of General Psychology, 114, 373-381.

Schwartz, S. H. (1966). Trial-by-trial analysis of processes in simple and disjunctive concept attainment tasks. Journal of Experimental Psychology, 72, 456-465.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. Quarterly Journal of Experimental Psychology, 42, 209-237.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retreival failure? Science, 168, 1601-1603.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing (pt 2): Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. Organizational Behavior and Human Performance, 6, 649-744.

Smoke, K. L. (1932). An objective study of concept formation. Psychological Monographs, 42, whole no. 191.

Smolensky, P. (1988). On the proper treatemnt of connectionism. Behavioral and Brain Sciences, 11, 1-74.

Squire, L. R., & Frambach, M. (1990). Cognitive skill learning in amnesia. Psychobiology, 18, 109-117.

Stadler, M. (1989). On learning complex procedural knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1061-1069.

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction, and practice in a simulated process control task. Quarterly Journal of Experimental Psychology, 41, 553-577.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds), Parallel distributed processing: Explorations in the microstructure of cognition. Vol. I: Foundations (pp 444-459). Cambridge, Ma.: MIT Press.

Treisman, M., & Faulkner, A. (1985). On the choice between choice theory and signal detection theory. Quarterly Journal of Experimental Psychology, 37, 387-405.

Treisman, M., & Faulkner, A. (1987). Generation of random sequences by human subjects: Cognitive operations or psychophysical process? Journal of Experimental Psychology: General, 116, 337-355.

Truijens, C. L., Trumbo, D. A., & Waganaar, W. A. (1976). Amphetamine and barbiturate effects on two tasks performed singly and in combination. Acta Psychologica, 40, 233-244.

Tulving, E. (1983). Elements of episodic memory. Oxford: Clarendon Press.

Tulving, E. (1985). How many memory systems are there? American Psychologist, 40, 385-398.

Venit, S., & Bishop, W. (1985). Elementary linear algebra, 2nd edition. Boston: Prindle, Weber, & Schmidt.

Verplanck, W. S. (1962). Unaware of where's awareness: Some verbal operants - notates, monents, and notants. In C. W. Eriksen (Ed.), Behavior and awareness - a symposium of research and interpretation (pp 130-158). Durham, N.C.: Duke University Press.

Waganaar, W. A. (1970). Subjective randomness and the capacity to generate information. Acta Psychologica, 33, 233-242.

Walk, R. D. (1952). Effects of discrimination reversal on human discrimination learning. Journal of Experimental Psychology, 44, 410-419.

Ward, T. B. (1988). When is category learning holistic? A reply to Kemler-Nelson. Memory & Cognition, 16, 85-89.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family resemblance concepts. Memory & Cognition, 15, 42-54.

Widrow, C., & Hoff, M. E. (1960). Adaptive switching circuits. Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, 4, 96-104.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1047-1060.

Wilson, A. (1975). The inference of covert hypotheses by verbal reports in concept learning research. Quarterly Journal of Experimental Psychology, 27, 313-322.

Wolff, J. G. (1975). An algorithm for the segmentation of an artificial language analogue. British Journal of Psychology, 66, 79-90.

Wolff, J. G. (1977). The discovery of segments in natural language. British Journal of Psychology, 68, 97-106.

Wolff, J. G. (1980). Language acquisition and the discovery of phrase structure. Language and Speech, 23, 255-269.

Wolitzky, D., & Spence, D. (1968). Individual consistencies in the random generation of choices. Perceptual and Motor Skills, 26, 1211-1214.

Appendix for Chapter Three

Analysis of Variance Summary Tables

Experiment Three: Dependent variable is number of correct classifications. "Task" refers to double sentence versus single sentence task; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on structured transfer task.

Source	df	SS	MS	F
Task	1	0.29	0.29	0.22
Error	22	29.05	1.32	
Perf	1	2.30	2.30	1.73
Perf X Task	1	0.46	0.46	0.35
Error	22	29.26	1.33	

Experiment Three: Dependent variable is number of correct classifications. "Task" refers to double sentence versus single sentence task; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on Free Recall.

Source	df	SS	MS	F
Task	1	0.01	0.01	0.00
Error	22	30.15	1.37	
Perf	1	0.08	0.08	0.11
Perf X Task	1	0.05	0.05	0.07
Error	22	15.61	0.71	

Experiment Three: Dependent variable is reaction time. "Task" refers to double sentence versus single sentence task; "Correct" refers to correct versus incorrect responses.

Source	df	SS	MS	F
Task	1	12.59	12.59	0.70
Error	22	393.01	17.86	
Correct	1	0.21	0.21	0.13
Task X Correct	1	0.27	0.27	0.17
Error	22	35.30	1.60	

Experiment Four: Dependent variable is number of correct classifications. "Task" refers to double sentence versus single sentence task; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on structured transfer task.

Source	df	SS	MS	F	
Task	1	63.94	63.94	9.36	p=.006
Error	22	150.30	6.83		
Perf	1	20.02	20.02	10.24	p=.004
Perf X Task	1	18.01	18.01	9.21	p=.006
Error	22	43.03	1.96		

Experiment Four: Dependent variable is number of correct classifications. "Task" refers to double sentence versus single sentence task; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on Free Recall.

Source	df	SS	MS	F
Task	1	63.94	63.94	10.69 p=.004
Error	22	134.56	5.98	
Perf	1	18.25	18.25	13.46 p=.001
Perf X Task	1	18.01	18.01	13.28 p=.001
Error	22	29.84	1.36	

Experiment Four: Dependent variable is reaction time. "Task" refers to double sentence versus single sentence task; "Correct" refers to correct versus incorrect responses.

Source	df	SS	MS	F
Task	1	2.96	2.96	0.15
Error	22	445.43	20.25	
Correct	1	0.30	0.30	0.07
Task X Correct	1	0.12	0.12	0.03
Error	22	88.00	4.00	

Experiment Five: Dependent variable is number of correct classifications. "Criterion" refers to above versus below criterion in the learning phase; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on structured transfer task.

Source	df	SS	MS	F
Criterion	1	3.48	3.48	0.84
Error	16	66.30	4.14	
Perf	1	0.36	0.36	0.61
Perf X Crit	1	0.09	0.09	0.15
Error	16	9.42	0.59	

Experiment Five: Dependent variable is number of correct classifications. "Criterion" refers to above versus below criterion in the learning phase; "Perf" refers to predicted versus actual classification performance. Predicted Performance based on Free Recall.

Source	df	SS	MS	F
Criterion	1	4.84	4.84	0.70
Error	16	110.00	6.88	
Perf	1	0.44	0.44	0.38
Perf X Crit	1	0.00	0.00	0.00
Error	16	18.84	1.18	

Appendix for Chapter Four

Figure 3.

Test of Sequential Letter Dependencies (SLD test)

Stem:	Next letter: M	V	X	R	T
START					
M					
MV					
MT					
MTT					
MTTT					
MTTTT					
MTV					
MTTV					
MTTTV					
MTVR					
MTTVR					
MTVRX					
MVR					
MVRX					
MVRXV					
MVRXR					

M V X R T

V

VX

VXV

VXVR

VXVRX

VXT

VXTT

VXTTT

VXTV

VXTVR

VXTTV

VXR

VXRR

VXRRR

Analysis of Variance Summary Tables

Experiment Six: Dependent variable is proportions. "Group" refers to POS versus POSNEG; "Type" refers to EE versus AV.

Source	df	SS	MS	F
Group	1	21.53	21.53	4.53
Error	38	188.22	4.95	
Type	1	196.81	196.81	18.74 p=.000
Type X Group	1	1.38	1.38	0.13
Error	22	392.12	10.50	

Table 1.
Classification performance and position of violation for nongrammatical items

Position of violation	Group POS	POSNEG
PA	.64 (.28)*	.78 (.17)*
P1	.73 (.21)	.58 (.16)
P2	.58 (.22)*	.55 (.18)*
P3	.71 (.24)	.60 (.26)
P4	.55 (.23)	.56 (.16)
P5	.60 (.31)*	.46 (.26)
P6	.75 (.38)	.53 (.38)

Note Standard deviations appear in parentheses.
*Significantly different from chance (.5), $p<.05$.

Experiment Six: Dependent variable is proportion correct. "Group" refers to POS versus POSNEG; "Type" refers to type of nongrammatical string (position of violation in first position versus second position . . . versus sixth position); "Ti" refers to the trend in Type of degree i.

Source	df	SS	MS	F
Group	1	0.72	0.72	4.61 p=.038
Error	38	5.91	0.16	
T1	1	0.04	0.04	0.64
Error	38	2.60	0.07	
T2	1	0.13	0.13	2.05
Error	38	2.49	0.07	
T3	1	.07	.07	1.58
Error	38	1.67	0.04	
T4	1	0.24	0.24	5.76 p=.020
Error	38	1.60	0.04	
T5	1	0.11	0.11	4.40 p=.043
Error	38	0.97	0.03	
Type	5	0.60	0.12	2.45 p=.035[1]
Type X Group	5	0.36	0.07	1.46
Error	190	9.21	0.05	

[1] p=.048 with Huyn-Feldt correction.

Table 2.
Proportion correct and position probed on SLD test

	Group	
Position	POS	POSNEG
Q1	.69 (.20)	.66 (.23)
Q2	.67 (.17)	.69 (.13)
Q3	.75 (.09)	.70 (.13)
Q4	.66 (.10)	.61 (.09)
Q5	.64 (.10)	.59 (.07)
Q6	.61 (.09)	.57 (.08)

Note Standard deviations appear in parentheses.

Experiment Six: Dependent variable is proportion correct. "Group" refers to POS versus POSNEG; "Type" refers to type of SLD question (the question probes for allowable letters in the first letter position versus second letter position . . . versus sixth letter position); "Ti" refers to the trend in Type of degree i.

Source	df	SS	MS	F
Group	1	0.06	0.06	2.24
Error	38	1.03	0.03	
T1	1	0.29	0.29	16.33 p=.000
Error	38	0.68	0.02	
T2	1	0.07	0.07	2.87 p=.098
Error	38	0.95	0.02	
T3	1	0.03	0.03	1.85
Error	38	0.64	0.02	
T4	1	0.01	0.01	1.19
Error	38	0.47	0.01	
T5	1	0.07	0.07	8.64 p=.006
Error	38	0.29	0.01	
Type	5	0.47	0.09	5.96 p=.000[1]
Type X Group	5	0.04	0.01	0.52
Error	190	3.02	0.02	

[1] p=.001 with Huyn-Feldt correction.

Experiment Six: Dependent variable is sensitivity (the logistic approximation to d'). "Group" refers to POS versus POSNEG; "Task" refers to SLD versus classification.

Source	df	SS	MS	F
Group	1	0.95	0.95	9.22 p=.004
Error	38	3.92	0.10	
Task	1	0.19	0.19	3.88 p=.056
Task X Group	1	0.01	0.01	0.27
Error	38	1.88	0.05	

Experiment Six: Dependent variable is sensitivity (logistic approximation to d'). "Group" refers to POS versus POSNEG; "Task" refers to SLD (only those questions used to calculate PPSUM) versus classification.

Source	df	SS	MS	F
Group	1	1.37	1.37	12.06 p=.0013
Error	38	4.33	0.11	
Task	1	0.40	0.40	3.55 p=.075
Task X Group	1	0.01	0.01	0.07
Error	38	3.74	0.10	

Experiment Six: Dependent variable is number of correct classifications. "Group" refers to POS versus POSNEG; "Perf" refers to PPSUM versus actual classification performance.

Source	df	SS	MS	F
Group	1	361	361	6.35 p=.016
Error	38	2162	57	
Perf	1	45	45	0.91
Perf X Group	1	3	3	0.06
Error	38	1879	49	

Experiment Six: Dependent variable is number of correct classifications. "Group" refers to POS versus POSNEG; "Perf" refers to PPPROD versus actual classification performance.

Source	df	SS	MS	F
Group	1	466	466	8.51 p=.006
Error	38	2080	55	
Perf	1	38	38	0.86
Perf X Group	1	1	1	0.01
Error	38	1664	44	

Experiment Six: Dependent variable is number of correct classifications. "Group" refers to POS versus POSNEG; "Perf" refers to Free Recall versus actual classification performance.

Source	df	SS	MS	F
Group	1	273	273.8	10.70 p=.002
Error	38	972	25.6	
Perf	1	1479	1479.2	50.02 p=.000
Perf X Group	1	18	18.1	0.61
Error	38	1123	29.6	

Experiment Seven: Dependent variable is sensitivity (the logistic approximation to d'). "Group" refers to POS versus Experiment Seven; "Task" refers to SLD versus classification.

Source	df	SS	MS	F
Group	1	3.67	3.67	27.62 p=.0000
Error	30	3.98	0.13	
Task	1	0.00	0.00	0.01
Task X Group	1	0.07	0.07	1.63
Error	30	1.26	0.04	

Experiment Seven: Dependent variable is number of correct classifications. "Group" refers to POS versus Experiment Seven; "Perf" refers to PPSUM versus actual classification performance.

Source	df	SS	MS	F
Group	1	964	964	12.97 p=.001
Error	30	2229	74	
Perf	1	48	48	1.09
Perf X Group	1	0	0	0.00

Error	30	1309	44	

Experiment Seven: Dependent variable is number of correct classifications. "Group" refers to POS versus Experiment Seven; "Perf" refers to Free Recall versus actual classification performance.

Source	df	SS	MS	F
Group	1	451	451	12.46 p=.001
Error	38	1086	36	
Perf	1	714	714	29.94 p=.000
Perf X Group	1	105	105	4.44 p=.044
Error	38	715	23	

Experiment Eight: Dependent variable is sensitivity (logistic approximation to d'). "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions.

Source	df	SS	MS	F
Cond	2	1.79	0.90	5.69 p=.006
Inst	1	0.23	0.23	1.45
Cond X Inst	2	0.22	0.11	0.69
Error	54	8.51	0.16	

Experiment Eight: Dependent variable is number of correct classifications. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions.

Source	df	SS	MS	F
Cond	2	572	286	5.08 p=.0095
Inst	1	94	94	1.67
Cond X Inst	2	115	58	1.02
Error	54	3039	56	

Experiment Eight: Dependent variable is error proportions. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions; "Type" refers to EE vs Av.

Source	df	SS	MS	F
Cond	2	62.5	31.2	3.92 p=.026
Inst	1	16.9	16.9	2.12
Cond X Inst	2	13.4	6.7	0.84
Error	54	430.1	8.0	
Type	1	258	258	22.80 p=.0000
Type X Cond	2	40	20	1.77
Type X Inst	1	33	33	2.92 p=.093
Type X C. X I.	2	1	1	0.06
Error	54	611	11	

Table 3.
Experiment Eight: Classification performance and position of violation for nongrammatical items

	Group					
	Implicit			Explicit		
	Low	High	Single	Low	High	Single
Position of violation						
PA	.65	.67	.65	.80	.75	.77
	(.21)	(.27)	(.27)	(.15)	(.29)	(.26
P1	.66	.66	.70	.71	.69	.64
	(.12)	(.15)	(.24)	(.22)	(.26)	(.20
P2	.57	.60	.60	.67	.58	.58
	(.21)	(.12)	(.19)	(.31)	(.20)	(.18
P3	.67	.75	.73	.60	.69	.67
	(.23)	(.18)	(.14)	(.15)	(.15)	(.18
P4	.60	.69	.65	.58	.65	.69
	(.22)	(.17)	(.15)	(.18)	(.28)	(.21
P5	.53	.65	.60	.38	.48	.68
	(.22)	(.27)	(.29)	(.24)	(.28)	(.17
P6	.45	.50	.40	.35	.25	.50
	(.37)	(.41)	(.39)	(.41)	(.35)	(.41

Note Standard deviations appear in parentheses.

Experiment Eight: Dependent variable is proportion correct. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions; "Type" refers to type of nongrammatical string (position of violation in first position versus second position . . . versus sixth position); "Ti" refers to the trend in Type of degree i.

Source	df	SS	MS	F
Cond	2	0.19	0.10	1.16
Inst	1	0.11	0.11	1.33
Cond X Inst	2	0.14	0.07	0.85
Error	54	4.55	0.08	
T1	1	2.02	2.02	22.09 p=.0000
Error	54	4.94	0.09	
T2	1	0.76	0.76	10.47 p=.0021
Error	54	3.92	0.07	
T3	1	.22	.22	4.36 p=.042
Error	54	2.77	0.05	
T4	1	0.18	0.18	3.40 p=.071
Error	54	2.83	0.05	
T5	1	0.05	0.05	2.26
Error	54	1.16	0.02	
Type	5	3.23	0.65	11.17 p=.0000[1]
Type X Cond	10	0.40	0.04	0.69
Type X Inst	5	0.16	0.03	0.57
Type X C. X I.	10	0.45	0.05	0.78
Error	270	15.61	0.06	

[1]p=.0000 with Huyn-Feldt correction for sphericity violation.

Experiment Eight: Dependent variable is number of correct SLD responses. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions.

Source	df	SS	MS	F
Cond	2	1355	677	12.83 p=.0000
Inst	1	3	3	0.05
Cond X Inst	2	48	24	0.46
Error	54	2851	53	

Table 4.
Experiment Eight: Proportion correct and position probed on SLD test

Position	Group Implicit Low	Implicit High	Implicit Single	Explicit Low	Explicit High	Explicit Single
Q1	.60	.70	.80	.60	.56	.60
	(.21)	(.24)	(.13)	(.16)	(.16)	(.23
Q2	.63	.54	.67	.58	.54	.69
	(.12)	(.18)	(.16)	(.12)	(.13)	(.10
Q3	.66	.68	.73	.71	.65	.73
	(.15)	(.09)	(.13)	(.10)	(.10)	(.12
Q4	.54	.57	.65	.52	.55	.71
	(.09)	(.12)	(.12)	(.09)	(.12)	(.10
Q5	.53	.55	.64	.55	.54	.68
	(.08)	(.08)	(.10)	(.09)	(.09)	(.09
Q6	.56	.57	.63	.52	.56	.64
	(.08)	(.09)	(.12)	(.07)	(.06)	(.08

Note Standard deviations appear in parentheses.

Experiment Eight: Dependent variable is proportion correct. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructional conditions; "Type" refers to type of SLD question (the question probes for allowable letters in the first letter position versus second letter position . . . versus sixth letter position); "Ti" refers to the trend in Type of degree i.

Source	df	SS	MS	F
Cond	2	7625	3812	10.47 p=.0001
Inst	1	292	292	0.80
Cond X Inst	2	132	132	0.18
Error	54	19662	364	
T1	1	2172	2172	13.67 p=.0005
T1 X Inst	1	699	699	4.40 p=.041
Error	54	8579	159	
T2	1	303	303	2.36
T2 X Inst	1	885	885	6.89 p=.010
Error	54	6937	128	
T3	1	268	268	1.92
T3 X Inst	1	72	72	0.52
Error	54	7527	139	
T4	1	1056	1056	12.13 p=.001
T4 X Inst	1	95	95	1.09
Error	54	4700	87	
T5	1	2127	2127	25.37 p=.0000
T5 X Inst	1	4	4	0.05
Error	54	4529	84	
Type	5	5926	1185	9.92 p=.0000[1]
Type X Cond	10	1242	124	1.04
Type X Inst	5	1756	351	2.94 p=.013[2]
Type X C. X I.	10	1550	155	1.30
Error	270	32272	120	

[1] p=.0000 with Huyn-Feldt correction for sphericity violation

[2] p=.027 with Huyn-Feldt correction for sphericity violation.

Experiment Eight: Dependent variable is sensitivity (logistic approximation to d'). "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructions; "Task" refers to SLD versus classification.

Source	df	SS	MS	F
Cond	2	5.61	2.80	14.74 p=.000
Inst	1	0.09	0.09	0.47
Cond X Inst	2	0.32	0.16	0.83
Error	54	10.27	0.19	
Task	1	0.13	0.13	2.16
Task X Cond	2	0.33	0.17	2.79 p=.070
Task X Inst	1	0.14	0.14	2.39
Task X C. X I.	2	0.02	0.01	0.18
Error	54	3.19	0.06	

Experiment Eight: Dependent variable is number of correct classifications. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructions; "Perf" refers to PPSUM versus actual classification performance.

Source	df	SS	MS	F
Cond	2	1237	619	7.81 p=.001
Inst	1	295	295	3.72 p=.059
Cond X Inst	2	184	92	1.16
Error	54	4277	79	
Perf	1	132	132	2.85 p=.097
Perf X Cond	2	31	15	0.33
Perf X Inst	1	12	12	0.26
Perf X C. X I.	2	156	78	1.68
Error	54	2504	46	

Experiment Eight: Dependent variable is number of correct classifications. "Cond" refers single task versus grammar learning high priority versus grammar learning low priority; "Inst" refers to implicit versus explicit instructions; "Perf" refers to PPFR versus actual classification performance.

Source	df	SS	MS	F
Cond	2	807	403	7.23 p=.002
Inst	1	53	53	0.96
Cond X Inst	2	170	85	1.53
Error	54	3014	56	
Perf	1	3329	3329	160.83 p=.0000
Perf X Cond	2	37	18	0.88
Perf X Inst	1	41	41	1.97
Type X C. X I.	2	45	22	1.08
Error	54	1118	21	

Appendix for Chapter Five

Example of a program used for running the Connectionist models.
This program was used for all the simultaeous single-letter models.

```
58 RULE$ = "DELTA"
59 ALLEIGEN$="NO"
60 REM ****SIMULTANEOUS SINGLE****
62 REM ***LINES 10-56 USED IN SUB7000***
65 LEARNIT = 6
66 LR=.07
67 PRINT "LEARNED FOR ";LEARNIT;" ITERATIONS"
70 REM ***LEARNIT IS THE NUMBER OF LEARNING ITERATIONS****
73 DIM X(21,40),TEST(40),Y(40),W(30,40),CG(25)
74 DIM CNG(50),SEQ(20),CORR(50)
75 COOC$ = "CONT"
77 NUMIT=1
78 WEIGHTSONLY$ = "YES"
79 SEQUENCE$="THESIS"
80 REM **"RANDOM" FOR RANDOM SEQUENCE OF EXEMPLARS, "THESIS" FOR SEQ
USED IN
82 REM ***EXPTS SIX TO EIGHT.  FOR "THESIS", SET LEARNIT=6, FOR THE
6 SEQUENCES
85 PRINT RULE$
86 IF COOC$="COOC" THEN PRINT "COOCCURRENCE" ELSE PRINT
"CONTINGENCY"
87 PRINT "SINGLE SIMULTANEOUS LR= ";LR;" SEQUENCE= ";SEQUENCE$
88 BEGIN=31
89 GLIP=100
90 NLIP=100
91 BEGIN$="YES"
92 FOR I=31 TO BEGIN
93 IF BEGIN$="YES" THEN TEST(I)=1
94  FOR J=1 TO 20
95  IF BEGIN$="YES" THEN X(J,I)=1
96  NEXT J
97 NEXT I
98 REM "BEGIN" PROVIDES PERMAMENTLY ACTIVE "CONTEXT" UNITS
99 REM  (BEGIN-30) OF THEM
100 REM THESE UNITS WERE USED FOR THE SLD TASK
105 REM ****ACQUISITION***
110 FOR A=1 TO 20                          .
120  FOR I=1 TO 30
130  READ X(A,I)
131  NEXT I
132 NEXT A
133 FOR B=1 TO LEARNIT
135 IF SEQUENCE$="RANDOM" THEN GOSUB 6000
140 IF SEQUENCE$="THESIS" THEN GOSUB 7000
165  FOR A=1 TO 20
170  FOR I=1 TO 40
176  IF COOC$="COOC" AND X(SEQ(A),I)=-1 THEN X(SEQ(A),I)=0
180  IF RULE$="HEBB" THEN Y(I)=X(SEQ(A),I)
185  IF RULE$="DELTA" THEN Y(I)=0
190  NEXT I
195 FOR I=1 TO 30
```

```
200  FOR J=1 TO 40
205  IF RULE$="DELTA" THEN Y(I)=Y(I) + W(I,J)*X(SEQ(A),J)
210  NEXT J
220 NEXT I
230  FOR I=1 TO 30
240   FOR J=1 TO 40
245   FLAG=0
250   IF I<>J AND RULE$="HEBB" THEN
W(I,J)=W(I,J)+LR*Y(I)*X(SEQ(A),J)
255   IF I<>J AND RULE$="DELTA" THEN FLAG=1
257    IF FLAG=1 THEN
W(I,J)=W(I,J)+LR*(X(SEQ(A),I)-Y(I))*X(SEQ(A),J)
260   NEXT J
270  NEXT I
760 NEXT A
810 NEXT B
815
817 IF WEIGHTSONLY$="NO" THEN GOTO 1030
820 FOR I=1 TO 30
830  FOR J=1 TO BEGIN
840  PRINT BEGIN*I-BEGIN+J;" DATA ";W(I,J)
850  NEXT J
860 NEXT I
870
880 IF WEIGHTSONLY$="YES" THEN EXIT PROGRAM
1030
1040 REM ****** INPUT ******
1042 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1,-1,-1,-1,1
1043 DATA -1,-1,-1,-1,1,-1,1,-1,-1,-1
1044 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1,1,-1,-1,-1
1045 DATA -1,-1,-1,-1,1,-1,-1,-1,-1,-1
1046 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,-1,-1
1047 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1048 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,1,-1
1049 DATA -1,-1,1,-1,-1,-1,-1,-1,-1,-1
1050 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
1051 DATA -1,-1,1,-1,-1,1,-1,-1,-1,-1
1052 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
1053 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1054 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
1055 DATA -1,-1,-1,1,-1,-1,-1,-1,1,-1
1056 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
1057 DATA -1,-1,-1,-1,1,-1,1,-1,-1,-1
1058 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
1059 DATA -1,1,-1,-1,-1,-1,-1,-1,-1,-1
1060 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
1061 DATA -1,1,-1,-1,-1,-1,-1,-1,-1,1
1062 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1063 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1064 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1
1065 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1066 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1
1067 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,-1
1068 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1
1069 DATA -1,-1,-1,1,-1,-1,-1,-1,1,-1
1070 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1
1071 DATA -1,1,-1,-1,-1,-1,-1,-1,-1,1
1072 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1
```

```
1073 DATA -1,-1,-1,1,-1,-1,-1,1,-1,-1
1074 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1
1075 DATA -1,-1,-1.-1,1,-1,-1.-1,-1,-1
1076 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
1077 DATA -1,-1,1,-1,-1,-1,-1,-1,-1,-1
1078 DATA -1,1,-1.-1.-1,-1,-1,1,-1,-1,-1,1.-1,-1,-1.-1,-1,-1,1,-1
1079 DATA -1,-1,1.-1,-1,-1,1,-1,-1,-1
1080 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,-1,1
1081 DATA -1,-1,-1,-1,-1,-1.-1,-1.-1,-1
1175
1190
4780 REM ***** TEST ******
4781 REM ***GRAMMATICAL ITEMS***
4782 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1
4783 DATA -1,-1.-1,-1,1,-1,1,-1,-1,-1
4784 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,1,-1,-1,-1,-1,1
4785 DATA -1,1,-1,-1,-1,-1,-1,-1,-1,-1
4786 DATA 1,-1,-1,-1.-1.-1,-1.-1.-1,1.-1.-1,-1.-1,1,-1,1.-1.-1,-1
4787 DATA -1,-1,-1,1,-1,-1,-1,1,-1,-1
4788 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
4789 DATA -1,1,-1,-1,-1,-1,-1,-1,-1,1
4790 DATA 1,-1,-1.-1.-1,-1,-1.-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
4791 DATA -1,-1,1,-1,-1,-1,1,-1,-1,-1
4792 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
4793 DATA -1.-1,1,-1,-1,-1,-1,-1,1,-1
4794 DATA 1,-1,-1.-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
4795 DATA 1,-1,-1.-1,-1,-1.-1,-1,-1,-1,-1
4796 DATA -1,1,-1.-1,-1,-1.-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
4797 DATA -1,-1,1,-1,-1,-1,-1.-1,1,-1
4798 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1.-1,-1,-1,-1,1,-1,-1,-1,-1,1
4799 DATA -1,1,-1.-1,-1,-1,-1,-1,-1,1
4800 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1.-1,1,-1.-1,-1,-1
4801 DATA -1,-1,-1,-1,-1,-1.-1,-1,-1,-1,-1
4802 DATA 1,-1,-1,-1,-1,-1,1,-1,-1.-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1
4803 DATA -1,-1.-1.-1,-1,-1,-1,-1,-1,-1,-1
4804 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,-1,1
4805 DATA -1,-1.-1.-1,-1,-1,-1,-1,-1,-1
4806 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1.-1,1,-1,1.-1,-1,-1
4807 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4808 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1
4809 DATA -1,-1,-1,1,-1,-1,-1,-1,-1.-1
4810 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1.-1,-1,-1.-1,-1,1,-1
4811 DATA -1,-1,-1,1,-1,-1.-1,-1,-1,-1
4812 DATA -1,1,-1,-1,-1,-1,-1,1,-1.-1.-1,-1,-1,-1,1,-1,1.-1,-1.-1
4813 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4814 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1
4815 DATA -1,-1,-1,-1,-1,-1.-1,-1,-1,-1
4816 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,-1,1
4817 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4818 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,-1,-1
4819 DATA -1,-1,-1.-1,-1,-1,-1,-1,-1,-1
4820 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1.-1,-1.-1,1,-1.-1,-1,-1,1,-1
4821 DATA -1,-1,-1,1,-1,1,-1,-1,-1,-1
4822 DATA -1,1,-1.-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1.-1,1
4823 DATA -1,1,-1,-1,-1,-1,-1.-1,-1,-1
4824 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1.-1,-1,-1
4825 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4826 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
```

```
4827 DATA -1,-1,1,-1,-1,-1,-1,-1,-1,-1
4828 DATA -1,1,-1,-1,-1,-1,-1,1,-1.-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1
4829 DATA -1,-1,1,-1,-1,-1,1,-1.-1.-1
4830 DATA 1,-1,-1,-1,-1,-1,1,-1.-1.-1,-1,-1,-1,1,-1,-1.-1,1,-1,-1
4831 DATA -1,-1,-1,1,-1,1,-1,-1,-1,-1
4832 REM ***NONGRAMMATICAL ITEMS***
4833 REM ***SINGLE LETTER VIOLATIONS***
4834 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1.-1,-1,-1,-1,1,-1
4835 DATA -1,-1,-1.-1,1,-1.-1,-1,-1,-1
4836 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1,-1.-1,1,-1,-1,-1,-1.-1,-1,-1
4837 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4838 DATA -1,1,-1,-1.-1,-1,-1,1,-1,-1,-1.-1.-1,1,-1,-1,1,-1,-1.-1
4839 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4840 DATA -1,-1,1,-1,-1,-1,1,-1,-1,-1,-1,-1.-1,1,-1,-1,-1,1,-1.-1
4841 DATA -1,-1,-1,1,-1,-1,-1,-1,1,-1
4842 DATA -1,-1,1,-1,-1,-1.-1,-1.-1,1,-1,-1,-1,-1,1,-1.-1.-1,-1,1
4843 DATA -1,-1,-1,-1,1,-1,1,-1,-1,-1
4844 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,1.-1,-1,-1
4845 DATA -1,-1,-1.-1,-1,-1,-1,-1,-1,-1
4846 DATA 1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,1.-1.-1,-1,-1,-1,-1,1,-1
4847 DATA -1,-1,1,-1,-1,-1,-1,-1,-1.-1
4848 DATA 1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1.-1.-1,1
4849 DATA -1,-1.-1,1.-1,-1.-1.-1,-1,-1
4850 DATA 1,-1,-1,-1.-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,1.-1.-1,-1
4851 DATA -1.-1,-1,1,-1,-1,-1,1,-1,-1
4852 DATA -1,-1,-1,-1,1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,-1,-1,-1.-1,1
4853 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4854 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1.-1,-1,-1,1,-1,1,-1,-1,-1
4855 DATA -1,-1,-1.-1,1,-1,-1.-1,1,-1
4856 DATA -1,-1,-1,1,-1,-1,1,-1,-1,-1,-1,-1,-1,1,-1.-1,-1,-1,-1,-1
4857 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4858 DATA 1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1.-1,-1,-1.-1,-1,-1,1
4859 DATA -1.-1.-1,-1.-1,-1,-1,-1,-1
4860 DATA -1,1,-1,-1,-1.-1.-1,-1,1,-1,-1,-1.-1,1,-1,-1,-1,-1,1,-1
4861 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4862 DATA -1,1.-1.-1,-1.-1,1,-1,-1,-1,-1,-1,1.-1.-1,-1,-1,-1,1,-1
4863 DATA 1,-1,-1,-1,-1.-1.-1.-1,-1.-1
4864 DATA -1,1,-1,-1,-1,-1,-1,1,-1,-1.-1.-1,-1,1,-1,-1,-1.-1.-1,1
4865 DATA -1,-1,-1,-1,-1,-1,-1,-1,-1,-1
4866 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,-1,1,-1,-1,1,-1,-1,-1
4867 DATA -1,-1,-1.-1.-1,-1,-1,-1,-1,-1
4878 DATA -1,1,-1,-1,-1,-1,-1,1.-1,-1,1.-1.-1,-1,-1,-1,-1,-1,1,-1
4879 DATA -1.-1,1,-1,-1,-1,1,-1,-1,-1
4880 DATA 1,-1,-1,-1,-1,-1,-1,-1,-1,1,1,-1,-1,-1,-1.-1,-1,-1,-1,-1
4890 DATA -1,-1,-1,-1.-1.-1,-1,-1,-1,-1
4891 DATA -1,-1,-1,-1,1.-1.-1,1,-1,-1,-1,-1,-1,1.-1.-1,-1,-1,1,-1
4892 DATA 1,-1,-1,-1,-1.-1.-1,-1,-1.-1
4893 DATA 1,-1,-1,-1,-1,-1,-1,1,-1.-1,-1,1.-1,-1,-1,-1,-1,-1,1,-1
4894 DATA -1,-1,1,-1,-1,1,-1,-1,-1,-1
4895 DATA 1,-1,-1,-1,-1,-1,-1.-1,-1,1.-1,1,-1,-1,-1,-1.-1,-1,1,-1
4896 DATA -1,-1,-1,-1,1,-1,-1,-1,1,-1
4897 REM ***MULTIPLE VIOLATIONS***
4898 DATA -1.-1,-1,-1,1,-1,1.-1.-1,-1.-1.-1,-1,-1,1,-1.-1,-1,-1,1
4899 DATA -1,-1,1,-1,-1,-1,1,-1,-1,-1
4900 DATA -1,-1,1,-1,-1.-1,-1,-1,1,-1,-1,1,-1.-1,-1,-1,-1,1.-1,-1
4901 DATA -1,1,-1,-1,-1,-1,-1,-1.-1,-1
4902 DATA -1,-1,-1,1,-1.-1,-1,-1,1,-1,-1.-1,-1,1.-1,-1,-1,1.-1,-1
4903 DATA -1,1,-1,-1,-1,-1.-1,-1.-1,-1
```

```
4904 FOR Y=1 TO 50
4910  FOR I=1 TO 30
4930  READ X(21,I)
4935  IF COOC$="COOC" AND X(21,I)=-1 THEN X(21,I)=0
4937  TEST(I) = X(21,I)
4940 NEXT I
4970 GOSUB 5200
4980 NEXT Y
4985 GOSUB 5660
4987 PRINT "SETTLED FOR ";NUMIT;" ITERATIONS"
4990 PRINT "PCG ";PCG;"PCC ";PCC;"PCE ";PCE;"PEE ";PEE
5000 PRINT "PCNG ";PCNG;"PCCN ";PCCN;"PCEN ";PCEN;"PEEN ";PEEN
5005 PRINT "GIP ";GIP;" NIP ";NIP
5007 PRINT "GBIP ";GBIP;GB;" GLIP ";GLIP;GL;"NBIP ";NBIP;
5007 PRINT NB;"NLIP ";NLIP;NL
5008 PRINT "ORDER OF DIFFICULTY OF GRAMMATICAL ITEMS"
5009 FOR J=1 TO 25
5010 FOR I=1 TO 25
5011 IF CG(I)>XYZ THEN XY=I
5012 IF CG(I)>XYZ THEN XYZ=CG(I)
5013 NEXT I
5014 CG(XY)=CG(XY)-10
5015 PRINT XYZ;" EXEMPLAR NUMBER  ";XY;" RANK ";26 - J
5016 XYZ=0
5017 NEXT J
5018 PRINT "ORDER OF DIFFICULTY OF NONGRAMMATICAL ITEMS"
5019 XYZ=0
5020 FOR J=26 TO 50
5021 FOR I=26 TO 50
5022 IF CNG(I)>XYZ THEN XY=I
5023 IF CNG(I)>XYZ THEN XYZ=CNG(I)
5024 NEXT I
5025 CNG(XY)=CNG(XY)-10
5026 PRINT XYZ;" EXEMPLAR NUMBER  ";XY;" RANK ";51 - J
5027 XYZ=0
5028 NEXT J
5029 PRINT "EIGENVALUE ";LTEST.5
5030 LTEST= 0
5032 FOR I=1 TO 30
5034 LTEST = LTEST + TEST(I)2
5036 NEXT I
5038 PRINT LTEST.5
5040 PRINT "EIGENVECTOR "
5042 FOR I= 1 TO 30
5044 PRINT TEST(I)
5046 NEXT I
5048 EXIT PROGRAM
5180
5200 REM ***** CATEGORY OUTPUT ******
5203 FOR SETTLE = 1 TO NUMIT
5210 FOR I=1 TO 30
5220 Y(I)=0
5230 NEXT I
5265 FOR I=1 TO 30
5270  FOR J=1 TO BEGIN
5275  IF I<>J THEN Y(I)=Y(I)+W(I,J)*TEST(J)
5280  NEXT J
5290 NEXT I
```

```
5391 LTEST=0
5392 FOR I= 1 TO 30
5393 LTEST = LTEST + Y(I)2
5394 NEXT I
5395 FOR I=1 TO 30
5396 TEST(I) = Y(I)/LTEST.5
5397 NEXT I
5398 NEXT SETTLE
5405 IF ALLEIGEN$="NO" THEN GOTO 5460
5410 PRINT "EXEMPLAR ";Y;" EST EIGENVALE ";LTEST.5
5415 PRINT "EIGENVECTOR"
5420 FOR I=1 TO 30
5425 IF MOD(I,5)=1 THEN LETTER$="M"
5430 IF MOD(I,5)=2 THEN LETTER$="V"
5435 IF MOD(I,5)=3 THEN LETTER$="X"
5440 IF MOD(I,5)=4 THEN LETTER$="R"
5445 IF MOD(I,5)=0 THEN LETTER$="T"
5450 PRINT INT(I/5 - .00001) + 1;" ";LETTER$;" ";TEST(I)
5455 NEXT I
5460 IP=0
5465 LY=0
5470 LX=0
5480 FOR I=1 TO 30
5490 IP=IP+Y(I)*X(21,I)
5500 LY=LY+Y(I)2
5510 LX=LX+X(21,I)2
5520 NEXT I
5530 IP=IP/(LY*LX).5
5540 CORR(Y)=IP
5550 IF Y<26 AND IP>GBIP THEN GB=Y
5560 IF Y<26 AND IP>GBIP THEN GBIP=IP
5570 IF Y>25 AND IP>NBIP THEN NB=Y
5580 IF Y>25 AND IP>NBIP THEN NBIP=IP
5590 IF Y<26 AND IP<GLIP THEN GL=Y
5600 IF Y<26 AND IP<GLIP THEN GLIP=IP
5610 IF Y>25 AND IP<NLIP THEN NL=Y
5620 IF Y>25 AND IP<NLIP THEN NLIP=IP
5630 RETURN
5650
5660 REM ***ADJUSTING CLASSIFIFCATION CRITERIA***
5665 FOR I=1 TO 25
5670 GIP=GIP + CORR(I)/25
5680 NIP=NIP + CORR(I+25)/25
5690 NEXT I
5700 T=15
5705 THRESH=(NIP + GIP)/2
5710
5715 FOR J=1 TO 100
5720 FOR I=1 TO 50
5725 PCG=0
5730 PCC=0
5735 PCE=0
5740 PEE=0
5745 PCNG=0
5750 PCCN=0
5755 PCEN=0
5760 PEEN=0
5770  FOR Y=1 TO 50
```

```
5780  PG=1/(1+EXP(-T*CORR(Y)+T*THRESH))
5790  IF Y<26 THEN PCG=PCG+PG/25
5800  IF Y<26 THEN PCC=PCC+PG2/25
5810  IF Y<26 THEN PCE=PCE+PG*(1-PG)/25
5820  IF Y<26 THEN PEE=PEE+(1-PG)*(1-PG)/25
5830  IF Y>25 THEN PCNG=PCNG+(1-PG)/25
5840  IF Y>25 THEN PCCN=PCCN+(1-PG)2/25
5850  IF Y>25 THEN PCEN=PCEN+PG*(1-PG)/25
5860  IF Y>25 THEN PEEN=PEEN+PG2/25
5870  IF Y<26 THEN CG(Y)=PG
5880  IF Y>25 THEN CNG(Y)=1-PG
5890  NEXT Y
5900 IF PCG > PCNG THEN THRESH = THRESH + .01
5910 IF PCG < PCNG THEN THRESH = THRESH - .01
5920 NEXT I
5930 IF (PEE + PEEN)/2 > (PCE + PCEN)/2 + .05 THEN T=T-1
5940 IF (PEE + PEEN)/2 < (PCE + PCEN)/2 + .05 THEN T=T+1
5950 NEXT J
5960 PRINT "SCALING CONSTANT ";T;" THRESHOLD ";THRESH
5970 RETURN
5990
6000 REM ***ROUTINE TO PRESENT A RANDOM SEQUENCE OF LEARNING
6001 REM ***EXEMPLARS***
6100 FOR I=1 TO 20
6110  SEQ(I)=I
6120 NEXT I
6155 FOR I=1 TO 10
6160  RAN1 = INT(RND*20 + 1)
6165  RAN2 = INT(RND*20 + 1)
6170  DUMMY1 = SEQ(RAN1)
6180  SEQ(RAN1) = SEQ(RAN2)
6190  SEQ(RAN2) = DUMMY1
6200 NEXT I
6300 RETURN
6500
6600
7000 REM ***ROUTINE TO PRESENT LEARNING EXEMPLARS IN SAME ORDER AS
IN THESIS***
  10 DATA 1,2,9,16,6,5,14,19,12,20,4,15,18,13,7,11,8,3,10,17
  20 DATA 2,18,20,19,15,8,5,9,14,4,16,13,10,3,7,11,12,1,6,17
  30 DATA 20,5,10,16,17,3,12,8,2,18,9,13,11,15,6,14,19,1,4,7
  40 DATA 19,3,6,14,16,20,7,12,18,10,15,2,8,17,4,11,1,13,9,5
  45 DATA 15,4,7,20,3,9,12,5,8,19,2,13,16,6,11,1,18,17,10,14
  50 DATA 3,18,11,2,15,12,10,17,1,14,6,16,7,19,5,13,9,20,4,8
  51 DIM D(6,20)
  52 FOR I=1 TO 6
  53  FOR J=1 TO 20
  54  READ D(I,J)
  55  NEXT J
  56 NEXT I
7010 FOR I=1 TO 20
7020 SEQ(I) = D(MOD(B + 5,6) + 1,I)
7030 NEXT I
7040 RETURN
```

Appendix for Chapter Six

Example of a program used in simulating learning the dynamic control tasks. This program was used to run the first auto associator model.

```
30 rem ***this is the first auto associator model in the thesis***
40  N=28
45  T=9
60  DIM W(N,N),I(N),D(N),Y(N),B(3),REC(60),OLD(12,2)
62 B(1)=-1
64 B(3)=1
65 INPUT "NOISE SCALING (=SD/10)";NOSC
66  INPUT "NUMBER OF ACTIVE REINFORCEMENT UNITS FOR NEAR TARGET
(25-28)";JN
67  INPUT "TEST FOR KNOWLEDGE IN SPECIFIC SITUATIONS ";MARESCAUX$
70  INPUT "LEARNING RATE ";LR
75  INPUT "PRINT DETAILS ";DET$
80  INPUT "TEST EXPLICIT KNOWLEDGE ";EXPLICIT$
85  INPUT "PERSON (C,S, OR U) ";PERSON$
90  IF PERSON$="C" THEN H=60
95  IF PERSON$="S" THEN H=30
100 IF PERSON$="U" THEN H=50
105 FOR SUBJECT=1 TO 10
106 PRINT "SUBJECT ";SUBJECT;
110 FOR I=1 TO N
116  OLD(I,1)=0 IF I<13
117  OLD(I,2)=0 IF I<13
118  FOR J=1 TO N
120  W(I,J)=0
130  NEXT J
140 NEXT I
150 I(6)=1
155 FOR V=25 TO N
156 I(V)=1
157 NEXT V
160 FOR X=1 TO H
170 GOSUB 1000
180 G=-1000000
185 Q=0
190  FOR I=13 TO 24
192  GOSUB 3000
195  Y(I)=Y(I)+NOSC*NZ
200  IF Y(I)>G THEN Q=I
210  IF Y(I)>G THEN G=Y(I)
220  NEXT I
260 IF Q>0 THEN YB=Q-12
265 IF Q=0 THEN YB=INT(RND*12+1)
267 IF Q=0 THEN Q=YB+12
269 PQ=Q
270 IF PERSON$="C" THEN GB=2*YB-PGB+B((INT(RND*3+1)))
271 IF PERSON$="S" THEN GB=YB + 2 +B((INT(RND*3+1)))
272 IF PERSON$="U" THEN GB=PYB + 2 +B((INT(RND*3+1)))
273 IF GB>12 THEN GB=12
275 IF GB<1 THEN GB=1
276 FOR I=1 TO N
277 I(I)=0
278 Y(I)=0
```

```
279 NEXT I
280  FOR V=25 TO N
281  IF GB=T THEN I(V)=1
282  IF GB<>T THEN I(V)=-1
283  NEXT V
284  FOR V=25 TO JN
286  IF GB>(T-2) AND GB<(T+2) THEN I(V)=1
287  NEXT V
288  IF GB>(T-2) AND GB<(T+2) AND X>30 THEN OLD(PGB,2)=1 ELSE
OLD(PGB,2)=0
289  OLD(PGB,1)=YB IF X>30
295 PGB=GB
296 PYB=YB
297 I(GB)=1
300 I(Q)=1
301 REC(X)=GB
310 GOSUB 1000
320 GOSUB 2000
330 PRINT "GREEN'S BEHAVIOUR ";GB;" YOUR BEHAVIOUR ";YB;" TRIAL ";X
 IF DET$="Y"
340 I(Q)=I(Q)
349 FOR V=25 TO N
350 I(V)=1
351 NEXT V
360 NEXT X
365 GOSUB 5000
370 GOTO 380 IF EXPLICIT$<>"Y"
375 GOSUB 4000
380 FOR I=1 TO N
381 I(I)=0
382 NEXT I
384 IF MARESCAUX$="Y" THEN GOSUB 30000
390 NEXT SUBJECT
500 EXIT PROGRAM
750
1000 REM ***** ACTIVATIONS ******
1180 FOR Z=1 TO N
1190 Y(Z)=0
1195 NEXT Z
1210  FOR J=1 TO N
1220   FOR B=1 TO N
1230   Y(J)=Y(J)+I(B)*W(J,B)
1240   NEXT B
1280  NEXT J
1330 RETURN
1500
2000 REM ***** WEIGHTS *****
2350 FOR J=1 TO N
2360 D(J)=I(J)-Y(J)
2370  FOR I=1 TO N
2380  W(J,I)=W(J,I) + LR*D(J)*I(I) IF I<>J
2390  NEXT I
2400 NEXT J
2410 RETURN
2500
3000 REM *****NOISE*****
3010 NZ=0
3015 DIM R(8)
```

```
3020 FOR UW=1 TO 8
3030 R(UW)=(INT(RND*10+1))/10
3040 NZ=NZ+R(UW)/8
3050 NEXT UW
3060 NZ=NZ-.5
3070 RETURN
3080
4000 REM ***EXPLICIT KNOWLEDGE TEST***
4005 FOR X=1 TO N
4010 I(X)=0
4020 NEXT X
4025 EXPLKN=0
4030 FOR X=13 TO 24
4050 I(X)=1
4055  FOR LOOP=1 TO 12
4056  I(LOOP)=1
4060  GOSUB 1000
4070  Q=0
4080  G=-10000000
4090   FOR J=1 TO 12
4100   IF Y(J)>G THEN Q=J
4110   IF Y(J)>G THEN G=Y(J)
4120   NEXT J
4125   CGB = 2*(X-12) - LOOP IF PERSON$="C"
4126   CGB = (X-12) + 2 IF PERSON$="S"
4127   IF Q=CGB OR Q=CGB+1 OR Q=CGB-1 THEN EXPLKN=EXPLKN+1/144
4130   I(LOOP)=0
4135   NEXT LOOP
4140  I(X)=0
4150 NEXT X
4155 PRINT "EXPLICIT KNOWLEDGE ";EXPLKN
4160 RETURN
4500
5000 REM ***NUMBER OF TRIALS ON TARGET***
5010 IF PERSON$<>"C" THEN GOTO 5300
5020 ONTARGET=0
5030 FOR I=1 TO 30
5040 IF REC(I)=T-1 OR REC(I)=T OR REC(I)=T+1 THEN
 ONTARGET=ONTARGET+1
5045 NEXT I
5050 PRINT "NUMBER OF TRIALS ON TARGET ";ONTARGET
5060 ONTARGET=0
5070 FOR I=31 TO 60
5080 IF REC(I)=T-1 OR REC(I)=T OR REC(I)=T+1 THEN
 ONTARGET=ONTARGET+1
5090 NEXT I
5100 PRINT "NUMBER OF TRIALS ON TARGET ";ONTARGET
5200 GOTO 5360
5300 IF PERSON$="S" THEN START=21 ELSE START=41
5310 ONTARGET=0
5320 FOR I=START TO H
5330 IF REC(I)=T-1 OR REC(I)=T OR REC(I)=T+1 THEN
 ONTARGET=ONTARGET+1
5340 NEXT I
5350 PRINT "NUMBER OF TRIALS ON TARGET ";ONTARGET
5360 RETURN
25000
30000 REM ***MARESCAUX***
```

```
30001 OLDU=0
30002 OLDR=0
30020 CORRRE=0
30030 CORUNR=0
30035 CONRE=0
30037 CONUN=0
30038 CORRNEW=0
30040 FOR I=1 TO 12
30060  IF OLD(I,2)>0 THEN OLDR=OLDR+1
30070  IF OLD(I,1)>0 THEN OLDU=OLDU+1
30080 NEXT I
30085 OLDU=OLDU-OLDR
30090 FOR I=1 TO 12
30110   FOR K=13 TO 24
30120   Y(K)=W(K,I)
30280   GOSUB 3000
30290   Y(K)=Y(K)+NOSC*NZ
30300   NEXT K
30310  G=-100000000
30320   FOR K=13 TO 24
30330   IF Y(K)>G THEN Q=K-12
30340   IF Y(K)>G THEN G=Y(K)
30350   NEXT K
30360   FOR K=13 TO 24
30370   IF K=Q+12 THEN Y(K)=1
30380   IF K<>Q THEN Y(K)=0
30390   NEXT K
30395   print Q
30405  AO=2*Q-I
30407  OK=0
30408  NEW=12-OLDU-OLDR
30410  IF AO=T OR AO=T-1 OR AO=T+1 THEN OK=1
30420  IF OK=1 AND OLD(I,2)>0 THEN CORRRE=CORRRE+1/OLDR
30430  IF OK=1 AND OLD(I,2)=0 AND OLD(I,1)>0 THEN
 CORUNR=CORUNR+1/OLDU
30440  IF OK=1 AND OLD(I,1)=0 AND NEW>0 THEN CORRNEW=CORRNEW+1/NEW
30450  IF OLD(I,2)>0 AND Q=OLD(I,1) THEN CONRE=CONRE+1/OLDR
30460  IF OLD(I,2)=0 AND Q=OLD(I,1) THEN CONUN=CONUN+1/OLDU
30480 NEXT I
30485 PRINT "OLDR ";OLDR;" OLDU ";OLDU;" NEW ";NEW
30490 PRINT "PROPORTION CORRECT, OLD REINFORCED SITUATION:  ";CORRRE
30500 PRINT "                    OLD UNREINFORCED        :  ";CORUNR
30510 PRINT "                    NEW                     :
 ";CORRNEW
30520 PRINT "CONCORDANCE, REINFORCED                     :  ";CONRE
30530 PRINT "             UNREINFORCED                   :  ";CONUN
30540 RETURN
```