UNIVERSITY OF SOUTHAMPTON

FACULTY OF ELECTRONICS, SCIENCE AND MATHEMATICS

School of Electronics and Computer Science

LEARNABLE ARTIFICIAL GRAMMAR RULES ARE ONLY LEARNED EXPLICITLY

by

Martina Treacy Johnson

Thesis for the degree of Doctor of Philosophy (PhD)

Cognitive Science

August 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

LEARNABLE ARTIFICIAL GRAMMAR RULES ARE ONLY LEARNT EXPLICITLY

by Martina Treacy Johnson

In most rule-learning experiments subjects (Ss) are trained with both positive and negative instances of the rule. However, in most traditional artificial grammar learning (AGL) experiments Ss are trained with positive instances only and using very complex rules. In a typical training phase Ss are unaware of the underlying rules governing the stimuli, and are instead instructed to do an irrelevant task. After training they are told about the rules, and then have to differentiate between positive and negative stimuli. Ss' typical performance is significantly better than chance, although Ss are unable to verbalise the rules and think they are guessing. This dissociation of performance and verbalisation led Reber (e.g. 1967) to conclude that Ss are acting on "implicit" (i.e. unconscious), abstract knowledge. However, it has also been argued that Ss are not learning the abstract rules but are basing their classification on memorised fragments of the stimuli. In two experiments in this thesis, it was shown that Ss seem to be memorising fragments of the stimuli, rather than learning the underlying rules of the stimuli. It was further shown that presenting Ss with positive and negative evidence in the training phase was detrimental to subsequent test performance. If Ss were simply memorising fragments of stimuli, negative evidence would simply add to the memory load, and performance would thus decrease. In the main experimental series in this thesis, the critical antecedent step of testing the learnability of rules was taken, and an easy, medium and hard set of rules were constructed. It was seen that only very simple rules could be mastered to 100% during an experimental session. In several web-based experiments, groups of Ss were trained (1) with and without corrective feedback, (2) with an active response or mere passive exposure to the stimuli, (3) with forewarning about the existence of rules, forearmed with the actual rules, or with no prior knowledge of the existence of rules, and (4) with positive and negative stimuli, positive stimuli only, or negative stimuli only. It was found that Ss with high performance rates could also verbalise the rules, i.e. learning was explicit, and that passively being exposed to stimuli resulted in better performance than actively responding to the stimuli. Established AGL effects may merely be artefacts of the fact that traditional artificial grammars were too complex and unlearnable within the scope of an experimental session, leaving the memorisation of fragments as the only basis for any improvement.

Table of Contents

Acknowledgements

First of all, I would like to thank my supervisor, Stevan Harnad, for his support, motivational abilities, constructive criticism, and patience.

I am grateful to all members of the IAM Group for their support, both academic and technical. Particular thanks go to Les Carr for standing in as supervisor, and Jon Hallett without whose technical advice and support I would not have been able to work from abroad.

The creation of the web-based experiments was started by Jorge Louis de Castro as a third-year project. The code was further extended and modified with the ever-patient help of Michael O. Jewell, who was indeed a 'jewel'! Many, many thanks to both of them for their support and advice and for helping me with all the little Java code problems (and there were many!).

I could not have had better bay-mates than Yioula Roidouli and Vijay Dialani and I thank them for enjoyable times and accommodating my many desk rearrangements and complaints about the air conditioning! I particularly wish to thank Yioula for always being willing to accompany me outside, rain or shine, when the inevitable craving for nicotine and/or caffeine kicked in.

Most thanks go to my family for their continuous love and support and to Pavlos for everything we have together.

Martina Johnson

1 Introduction

One of the most important capacities of the mind is its ability to sort the inputs it receives into categories on which it can act differentially. The world consists of an infinite number of potentially different things and kinds of things, a "blooming, buzzing confusion." In order to survive, behave adaptively in their environment, and reproduce, all organisms must partition things into kinds, so that non-identical stimuli of the same kind that *need* to be treated in the same way *can* be. Correct categorisation is based on the detection of a set of features or rules that is sufficient to determine which category the thing in question belongs to. This set of features or rules can only be detected if one has sampled things belonging to the category in question (positive instances) as well as things that do not belong to the category (negative instances). If one were to sample only positive instances of a category one would have no way of knowing whether a new instance belonged to the same category or to a different one. To find a set of features sufficient to serve as a basis for discriminating positive from negative instances, one must sample both their presence and their absence.

The early research on concept learning illustrates this need for both positive and negative instances. In the typical concept learning experiment conducted by Bruner, Goodnow, & Austin (1956), subjects (Ss) are presented with a number of pictures varying in several features: shape, colour, number of borders, and number of figures depicted. Ss are told that the experimenter has a "concept" in mind (e.g. any red circles, or three blue squares) and that some of the pictures are instances of that concept while others are not. Ss' task is to discover what the concept is.

The experiment begins with Ss being shown a picture that is a positive instance of the concept. Ss are then shown a sequence of cards one by one, and for each one they are told whether it is a positive or a negative instance of the concept. After seeing each card, Ss are required to write down their hypothesis as to what they think the concept is. They are not able to refer back to previously seen cards, or to previous hypotheses. If their final answer at the end of the training is the correct concept, they are considered to have attained the concept. Ss are given just enough training instances to provide a sufficient database for attaining the concept. Ss typically proceed by trying and testing hypotheses (e.g. "it's anything red", or "anything red and square" etc), and revising their hypotheses based on the feedback they receive. Ss are in effect trying to find the set of features and "rules" that will enable them to determine whether or not any further instance falls under the concept (i.e., belongs in the category).

In such experiments, successful learning depends on several factors: The learner needs to sample both positive and negative instances, and to ascertain which instances are positive and which are negative; in other words, he needs feedback on the success or failure of his hypotheses in order to be able to revise and correct them. (And that feedback obviously has to be reliable, for if the feedback to a correct categorisation were too often "wrong", and vice versa, learning could not occur.)

Although it is common practice in the concept learning literature to use positive and negative instances (indeed, when no innate or prior knowledge exists, learning depends critically on sampling both), in another area of rule learning, the typical study has used only positive instances: the area of artificial grammar learning (AGL). Reber first started experimenting with artificial "grammars" (AGs) in the late 1960s (e.g. Reber, 1967; Reber, 1969) to investigate how Ss respond to strings of letters when their sequence follows a "grammatical" rule. In the training phase, Reber presented letter strings that were all positive instances of a rule without telling Ss that they followed a rule. Ss were instructed to do a mental task that had nothing to do with trying to find rules, for example, memorising the string of letters.

The training phase was followed by a test phase in which Ss were told that all the previous strings had followed a rule. They were then asked to sort a further set of strings into those they thought followed that same rule, and those that did not. (This was similar to Bruner et al.'s concept-attainment task, except that the training strings had all been positive instances only, and Ss had been unaware while sampling them that there was any underlying rule; the test strings – positive and negative – were presented for sorting without the help of any corrective feedback.) Reber found that Ss were nevertheless able to sort the test strings at a level better than chance under these conditions, without being able to verbalise any rule. Reber dubbed this "implicit learning", that is, learning without awareness, in contrast to the kind of learning that had occurred in Bruner et al.'s studies, which would have been called "explicit learning" (because most Ss could verbalise the rule).

The research area of AGL and implicit learning was initially motivated by a phenomenon in natural language learning. One of the early observations in natural language grammar learning (e.g. Chomsky, 1980) had been that young children "attain" the rules of Universal Grammar (UG) even though they never sample negative instances (they never hear "ungrammatical" sentences, i.e. those violating the rules of UG). Linguists concluded that the rules of UG must therefore be innate, because positive instances alone are not a sufficient basis for learning them (the "Poverty of the Stimulus", see also Chapter 2). Reber and others, however, wanted to show that positive instances alone are sufficient for learning rules after all, so they began to study Artificial Grammar (AG) strings. Even though it soon became clear that AG bore little or no relation to UG, the vast majority of AGL experiments use only positive instances.

There is much debate in the AGL literature about the nature of the knowledge that Ss acquire, with many researchers (e.g. Reber, 1967) convinced that Ss are learning the rules of the AG, while others (e.g. Perruchet & Pacteau, 1990; Johnstone & Shanks, 2001) argue that Ss are merely memorising fragments of the stimuli and using that rote memorisation to perform at above chance levels on new strings that contain the familiar fragments. The results of the first two experiments in this thesis support the latter interpretation: that traditional AGL experiments are really just memorisation experiments, and that Ss do not learn the underlying rule(s). This is partly because in the AGL training phase Ss saw only positive instances, and partly because the AG rules used were so complex that Ss had no chance of learning them within the short training time allocated for an experiment. In fact, no researcher had ever even checked whether or not the AGs used were learnable at all, given enough positive and negative instances, corrective feedback, and time. If the AG is so complex that it cannot be learnt at all, perhaps it is unreasonable to expect that any kind of learning, explicit or implicit, is actually taking place.

In the main experimental series in this thesis, the critical antecedent step was taken of ensuring that the rules (the AG) were learnable within the time of the single-session short-term training period of the typical AGL experiment. (Long-term, multi-session learning and over-learning were not investigated.) Three learnable short-term learning rules – an easy, medium and difficult one – were devised. It was shown that only very simple rules are learnable in the number of trials (between 20 and 80 learning trials and 40 test trials) and time (maximum 45 minutes) of a typical experiment. Several experiments were conducted in which various critical factors were manipulated, such as: (1) presence or absence of corrective feedback, (2) type of response (active or passive), (3) instructions (presence of absence of prior knowledge that there was an underlying rule), and (4) training sample (positive instances only, negative only, or both). An attempt was then made to relate the outcomes to the findings from past AGL experiments and thereby reassess the field of AG implicit learning.

1.1 Chapter Outline

In Chapter 2 of this thesis, the emergence of AGL experiments is reviewed, from their origins in natural language learning to Project Grammarama (Miller, 1958), the grandmother of all AGL experiments.

In Chapter 3, the standard AGL paradigm is described, beginning with the Reber studies in the late 1960s that created the implicit learning paradigm. One of the most controversial issues in the AGL community concerns what kind of knowledge people actually derive from AGL experiments, and this debate is also reviewed in Chapter 3.

Chapter 4 addresses the debate and reports the first experiment, which shows that Ss may not really be learning the rules of the AG but merely remembering parts of strings that tend to recur.

Then the few AGL studies in which both positive and negative instances were used are reviewed in detail in Chapter 5.

Chapter 6 reports the second experiment, which shows that negative instances actually interfere with learning when instances are just presented passively for memorisation, with no corrective feedback, because all they do is add to S's memory load. This too supports the conclusion that many AGL experiments may simply be rote memorisation tasks rather than rule learning.

Chapter 7 considers the many different factors that need to be controlled and tested in an experiment in order to be able to draw any precise conclusions.

Chapter 8 firstly outlines the general method used in the main experimental series of this thesis. Because traditional AGL experiments have mainly used very complex rules and researchers have not taken the critical prior step of testing whether those rules were learnable at all, Chapter 8 also describes several pilot studies. These were conducted in order to select an easy, medium and hard rule that was learnable within a single experimental session of 20, 40, or 80 learning trials respectively (lasting about 15, 25 and 40 minutes respectively, but varying according to the subject's speed). Training in these pilot studies was with both positive and negative instances, along with information about which was which (error-corrective feedback). Only very simple rules proved to be learnable within the number of trials used.

Chapters 9 to 13 consist of the main experimental series of this thesis. Using the very simple rules constructed in the pilot studies (Chapter 8), several experiments tested the effects of training with *presence versus absence of corrective feedback* (Chapter 9); *active versus passive responding* (Chapter 10); *presence versus absence of prior knowledge that there was an underlying rule* ("forewarned" about the existence of rules, "forearmed" with the actual rules themselves, or "rule-blind" with no prior knowledge at all of the existence of rules) (Chapter 11); and *presence or absence of negative instances* (positive instances only, negative instances only, or both positive and negative instances) (Chapter 12).

Chapter 9 analyses the effect of corrective feedback. One group of Ss received immediate error-correcting feedback on their responses, from trial to trial; another group received delayed feedback about their overall performance following an entire block of 20 trials.

Chapter 10 analyses the effect of active responding. One group of Ss was required to make an active response to the stimulus after each presentation (i.e. to indicate whether they thought it had or had not followed the underlying rule). The passive group was not asked to make any response and merely told after each presentation whether or not the stimulus had followed the rule.

Chapter 11 analyses the effect of prior knowledge that there is an underlying rule. One group was told what the underlying rule was in advance; the second group was told there was a rule, but not what it was; the third group was not told anything about any rules and were instead told they were being tested on how fast they could reproduce words in a foreign language.

In Chapter 12 the effect of positive and negative instances was analysed. One group of Ss was trained with positive instances only, a second group with negative instances only, and these were compared to the group that received both positive and negative instances.

Chapter 13 summarises the main findings and concludes with an attempt to relate these findings to the origins of the AGL paradigm and implicit learning in general.

2 Background

2.1 Natural language learning

One of the first things a child has to learn is his native language. In order to speak a natural language one must learn to distinguish between grammatical and ungrammatical sentences. However, infants almost never encounter ungrammatical sentences[1] (negative evidence); the only sentences they hear (or produce) are grammatical ones (positive evidence; see e.g. Brown & Hanlon, 1970, for relevant empirical work). A child hears a finite number of sentences from his parents and elsewhere, and from this input must generalise to an infinite set of sentences (i.e., the whole language) that includes the input sample but goes beyond it. According to Chomsky's (1980) *"poverty of the stimulus"* argument, there are too many possible grammars that are compatible with the actual input data the child samples. Many of these grammars are simpler or more probable on the child's evidence than the correct grammar, yet the child does not "choose" an incorrect grammar. Since the child receives only positive instances, there is no way for him to discover that his hypothesised grammar is wrong; for that, his input data would also have to include negative instances, or he would have to utter enough ungrammatical sentences and be corrected on them. There is, however, considerable evidence that negative instances are not and cannot be necessary for first language acquisition (e.g. Brown & Hanlon, 1970). The knowledge that infants acquire when learning their first language is far greater than the information that is made available to them in the environment. As philosophers sometimes put it, the output of the language learning process is *underdetermined* by the input (e.g. Laurence & Margolis, 2001). In philosophy, *underdetermination* occurs when all data

[1] *Here, ungrammatical means 'violating Universal Grammar (UG)'. UG is a complex implicit set of rules gradually discovered by syntacticians and made explicit by them, but already "known" implicitly by the child. Children do sometimes hear, produce, and get corrected on ungrammatical examples in the sense of stylistic grammar (e.g. *ain't, etc.), but those are not violations of the rules of UG.*

can be explained by more than one theory and there is no way of determining which of the alternative theories is the correct one.

The underdetermination of linguistic input has led many researchers, beginning with Chomsky (e.g. 1968), to conclude that infants must have a pre-wired disposition to pick up the correct set of rules (grammar) on the basis of their impoverished training sample, because already "have" an inborn Universal Grammar (UG) at birth. Chomsky and many other linguists have come to the conclusion that every natural language has grammatical properties which are common to all human languages and are innately coded in the brain, giving infants the ability to learn grammar from positive instances only. This theory of the existence and innateness of UG is the prevailing one in linguistics today.

To illustrate, consider the following sentences (Chomsky, 1968):
1(a) They intercepted a message to *the boy*.
1(b) Who did they intercept a message to?

2(a) They intercepted John's message to *the boy*.
2(b) * Who did they intercept John's message to?[2]

If we transform the phrase in 1(a) to a question, we get the grammatical sentence 1(b). However, if we apply exactly the same process to the very similar sentence 2(a), we get the ungrammatical sentence 2(b). The speaker of English implicitly has somehow learned the rules for creating such sentences, and knows under which formal conditions these rules are applicable; i.e. speakers know that 1(a) is acceptable (grammatical), but 1(b) is not; they do not make this type of mistake. This is a big problem for learning machines (see Section 2.2), but not for children. Since children do not make mistakes that violate UG during the course of language acquisition, the principles of UG *must* already be built into the brain.

[2] *The asterisk is used to indicate a sentence that deviates in some respect from a grammatical rule.*

Chomsky (1987) likens UG to a complex structure with a switch box that contains a finite number of switches. To acquire a language the child's mind learns how the switches are set from the simple input it receives (this is called "parameter-setting"), but it need not learn the rules of the structure itself, because it has them already.

2.2 The Credit/Blame Assignment Problem

Related to the problem of underdetermination is the *"credit/blame assignment problem"* in machine learning. When learning to categorise objects, a machine (or a human) may be categorising several objects in a row successfully, basing the categorisation on a hypothesis it forms about certain underlying features or rules. However, the hypothesis may be wrong, because some or all of the rules it is based on are wrong; so eventually the learner categorises wrongly. It is often difficult to determine at the point of failure *which* features or rules are to blame for the error. The same thing applies to the opposite case, when the learner is repeatedly categorising incorrectly and at some point there occurs a correct categorisation: which feature is to be credited with the successful categorisation?

Researchers in machine learning try to design algorithms for solving the credit/blame assignment problem (e.g. Stroulia & Goel, 1996). For an algorithm to be successful it has to be able to generalise from the set of instances it was trained on, to all possible cases. When a machine makes a mistake, the algorithms are designed to detect that their performance has gone awry, and to adjust themselves. However, if the objects in the training set contain many features, determining which features to credit and which to blame is difficult; indeed, the credit/blame assignment problem is one of the hardest problems in Artificial Intelligence. The number of potential features that can be credited/blamed is sometimes simply too large (i.e., as in the problem of under-determination mentioned earlier).

2.3 Gold's Theorem

Results in formal learning theory also support the theory of UG, suggesting that the learners of natural language must be predisposed to choose the correct set of rules. In Gold's (1967) study of language learnability, learners are given strings of symbols and must say whether or not each string is correct on the basis of the information received so far. A "language" is taken to be a subset of the possible symbols from some finite alphabet of symbols. Gold investigated two basic ways to present the strings: "positive" and "positive/negative" presentation[3]. In the "positive" presentation Ss see only strings that are grammatical. In the "positive/negative" presentation Ss see both grammatical and ungrammatical strings and are informed which is which.

Gold studied several different types of "grammars" and their learnability. He found that only the most trivial of grammars, namely, those consisting of only a finite number of sentences, are learnable from positive instances alone. All other languages are learnable only if negative instances are also sampled.

2.4 Project Grammarama

In the late 1950s Miller started "Project Grammarama" on how infants learn a natural language. A natural language is based on a set of complex rules. The aim was to discover how people learn the pattern of a "Finite-State Grammar" (FSG). FSGs are complex rules for generating strings of letters; they can be depicted graphically as a finite number of states (see Figure 2.4.1 taken from Reber, 1967). The transition from one state (represented in the figure as a circle) to the next state produces a letter. A string is generated by entering at state 0; then each move to another state produces a letter until the final state is reached, and the string is complete. Any strings

[3] *Gold called positive presentation "text presentation" and positive/negative presentation "informant presentation", but for the present purposes I find it clearer to use positive and positive/negative.*

which can be generated in this way are *grammatical* or positive strings, and those which cannot are *ungrammatical* or negative strings. For example, in Figure 2.4.1 the positive string VVPS can be generated, whereas the string VXVPTS is negative as it cannot be produced by this FSG. A finite-state language is the infinite number of strings which can be generated by a FSG.

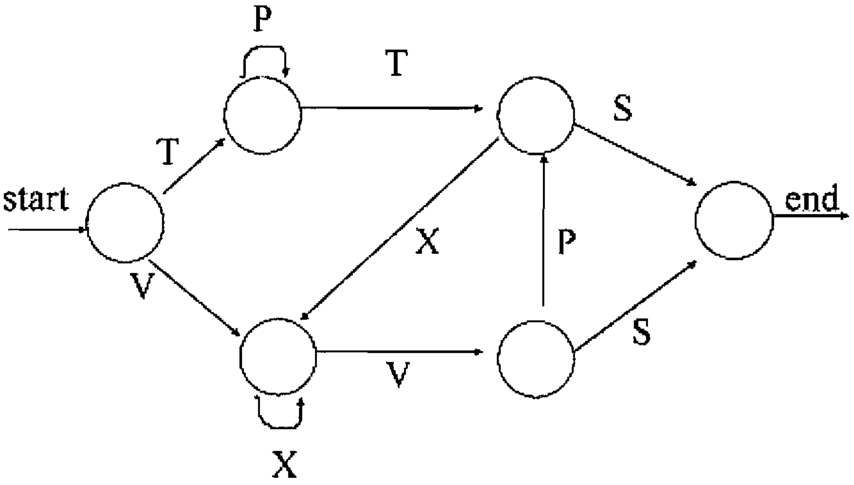


Figure 2.1: The finite-state artificial grammar created by Reber (1967)

Miller (1958) exposed people to strings generated by a simple FSG to see what they would do with them. To induce them to pay close attention to each string, Miller asked Ss to try to *memorise* each string when it was presented[4]. Miller used 18 positive strings (of no more than seven letters) generated by a FSG using the letters N, S, X, and G, and 18 negative strings of the same letters constructed from a table of random numbers. The strings were divided into two lists of nine positive strings (positive list) and two lists of nine negative strings (negative list). There were four pairs of lists: positive/positive, positive/negative, negative/positive, negative/negative. Ss were shown the first list of nine strings at a rate of one string every five seconds and were then asked to write down, in any order, all the strings they could recall. Then their response sheet was removed and the next trial started with the strings in a different order. Ss knew nothing about the rules governing the strings, and were just asked to read the nine strings and write down as many as they could remember on

[4] Miller used memorisation because previous work with AGs had also used memorisation (Aborn & Rubenstein, 1952, 1954).

each trial. Each subject had ten such trials for the first list and ten more for the second. After the tenth trial on the second list, Ss were asked to write down all the strings they could remember from the *first* list.

Ss who had studied a positive list first could remember more strings after the first ten trials than those who studied a negative list first, with an average of 8.9 compared to 3.79. Ss who learnt the positive/positive lists could remember most strings from the second list on the tenth trial, while the Ss who learnt the positive/negative, negative/positive, or negative/negative hardly differed in their results and were worse than the positive/positive Ss. In addition, when Ss studied a negative list prior to a positive list, the variance on the positive list was much greater than when they studied a positive list first. (The average number of strings recalled correctly from the first list after studying the second tended to be less than the number recalled after the first ten trials.)

In general, the more similar strings are to each other, the more Ss are to mix them up (interference) and to make errors. Miller found, however, that Ss were better able to memorise the positive strings, which were more similar to one another than the negative strings. Thus, Miller concluded that Ss must have been grouping and recoding the strings and thereby avoiding the interference effects one would have expected due to the similarity of the strings to each other. For example, one of the lists contained the strings: NNSG, NNSXG, NNXSXG, NNXXSG, NNXXSXG, and NNXXXSG. Given NNSG, these strings can be recoded as 00, 01, 11, 20, 21, and 30, where the two digits indicate the number of X's preceding and following the letter S. If one groups and recodes the stimuli in such a way, the similarity of the strings to each other does not interfere with performance.

Although when Miller first conceived of Project Grammarama he hoped that these experiments would cast light on the way infants acquire the rules of their native language, he soon came to realise that there are too many differences between this laboratory experiment and natural language learning. First, Ss are adults, not infants. They already know one or more languages, and a second language is never learnt the same way as the first. The artificial “language” is visual, not auditory, which accentuates a different kind of patterning. There is no meaning (no semantics) in FSGs and there is no use for the language once it has been learnt. There is also the time difference within which the language is learnt: infants take about 2 years, whereas the Ss in Project Grammarama had only one to two hours to learn. The infant acquires a practical sensorimotor skill whereas the Ss are puzzling over an abstract cognitive pattern. Most important, like Chomsky, Miller noted that the infants’ task is much more complicated than the Ss’. Since UG is much more complicated than artificial grammar (AG) rules, the child must already “know” UG innately and implicitly. These differences led Miller to conclude that there is almost nothing in common between natural language learning and artificial language learning (1958).

2.5 Conclusion

Although experiments like Miller’s, which are now collectively called artificial grammar learning (AGL) experiments, were originally motivated by the desire to gain insight into natural language learning (e.g. Miller, 1958; Reber, 1967), UG is so much more complex than any of the AGs used that AGL data proved unable to shed much light on that topic. However, even now that its original ties to the problem of UG have been dropped, the question of how people learn rules and under which conditions the rules can be learnt remains an object of research interest.

3 Artificial Grammar Learning

3.1 The Artificial Grammar Learning Paradigm

The artificial grammar learning (AGL) paradigm was further developed by Reber in 1967. He was interested in the process by which subjects (Ss) respond to the statistical nature of a stimulus array. To induce Ss to attend to the stimuli, Reber, like Miller, asked them to memorise the letter strings. Several subsequent researchers have continued to use the task of memorising to ensure that Ss are paying attention to the stimuli. Other researchers have used other attention-focussing tasks and other forms of presentation which have yielded similar results, for example sequential observation of instances (Lewis, 1975), anagram solving (Reber & Lewis, 1977), similarity matching (Mathews, Buss, Stanley, Blanchard-Fields, Cho, & Druhan, 1989), recognition (Vokey & Brooks, 1992), simultaneous scanning of large numbers of letter strings (Kassin & Reber, 1979; Reber, Kassin, Lewis, & Cantor, 1980) and paired-associate learning (Brooks, 1978).

Many AGL experiments use finite-state grammars (FSGs) such as the one depicted in Figure 2.4.1 (from Reber, 1967). In a typical AGL experiment, Ss are asked in a training phase to memorise strings of letters such as XMXRVM. These letter strings appear to be arbitrary but actually follow a set of rules. Ss are told they are taking part in a short-term memory experiment; they are told nothing about the underlying rules governing the structure of the strings. After the training phase, Ss are informed that the strings they have been memorising follow rules; they are then shown new strings. Some of these test strings follow the rules and some do not: S's task in the test phase is to classify which are which. Classification performance of Ss has been found to be significantly better than chance, indicating that Ss have acquired some knowledge. Most, however, are unaware of the knowledge they have gained and are unable to verbalise the rules.

This dissociation between performance and reportability led Reber (1967, 1989) to conclude that Ss are acting on “implicit”[5] (i.e. unconscious), abstract knowledge. Reber came to this conclusion by analogy with natural language learning. As with Miller, Reber’s motivation for AGL studies arose from the inability of the traditional learning paradigms (e.g. the stimulus-response approach[6]) to explain the way a child learns a natural language (UG). In Reber's view, what learning theories could not account for was the fact that learning seemed to occur implicitly, without conscious awareness. That is, a child learns to recognise and generate what does conform to UG and reject what does not, but children (and adults) are not explicitly aware of UG, hence cannot state the rules they are using (Chomsky, 1968).

Reber initially believed that implicit learning was "a rudimentary inductive process which is intrinsic in such phenomena as language learning and pattern perception" (1967). However, although Reber's interest in AGL was initially motivated by natural language learning, with the advances in linguistics, and for reasons Miller had already noted (see section 2.4), Reber later wrote, "I view the synthetic language as a convenient forum for examining implicit cognitive processes. If our explorations of this implicit domain of mentation turn out to provide insights into the acquisition of natural languages, that would be most pleasing" (p.30, Reber, 1993).

3.2 Rule or Fragment Knowledge?

In the AGL literature there has been much debate about the nature of the knowledge people gain: is it knowledge of the abstract rules of the grammar or is it just rote memory for recurring fragments of letter strings? Reber initially concluded that people are picking up the abstract rules of the grammar. However, there is considerable evidence (e.g. Perruchet & Pacteau, 1990; Brooks & Vokey, 1991) that Ss are merely memorising (explicitly) short

[5] *The term “implicit” was also used in the field of language learning: Chomsky argued that knowledge of UG was “implicit” or “tacit” knowledge, i.e. that UG was both innate and unconscious*

[6] *Stimulus-response learning is associative learning governed by the forming of a link or bond between a particular stimulus and a specific response*

fragments (chunks) of the whole letter string, i.e. two letter bigrams or three letter trigrams, and classifying novel positive and negative strings based on the familiarity of these novel strings to the previously seen strings. Perruchet and Pacteau (1990) trained one group of Ss on positive (grammatical) letter strings and another group only on the bigram fragments used in the strings, and both groups were able to distinguish between positive and negative test strings indicating that fragment memory is enough to account for the above-chance performance.

3.2.1 Transfer studies – Evidence for Abstract Knowledge?

Several researchers have claimed that transfer studies in which the letter-set or modality of the artificial grammar (AG) are changed at test provide evidence of abstract rule acquisition (e.g. Altmann, Dienes, & Goode, 1995; Knowlton & Squire, 1996). Although the only common factor between the training and test items is the abstract structure of the rules, it has been shown that Ss can still classify test items better than chance. For example, Altmann et al. (1995) trained one group of Ss on letter strings and a second on tone sequences, both conforming to the same grammar. At test Ss were required to classify strings that had been presented either in the same modality they had been trained on or strings presented in the other modality. Trained Ss could perform better than chance whether the classification test was in the same or different modality, whereas untrained control Ss only performed at chance levels. These findings suggest that Ss had learnt the abstract structure of the rules of the grammar.

However, Redington & Chater (1996) demonstrated in several models (simulations) that remembering fragments of two or three letters is enough to perform at similar levels at test without the need for abstract knowledge of the rules. Brooks & Vokey (1991) suggested that performance in transfer tests can be explained by the formal similarity between training and test items. For example, MXVVVM can be seen as similar to BDCCCB. Subjects could be learning the abstraction of a pattern, which could explain their performance, and need not be learning the actual abstract rules underlying the AG.

3.2.2 Problem with FSGs

One of the problems with this debate over what is and is not learned in AGL is that most AGL experiments have used FSGs. Shanks, Johnstone, & Staggs (1997) questioned the use of FSGs to investigate implicit learning, since they do not provide a convincing way of determining the contributions of rule and fragment knowledge in classification performance. The problem with FSGs is that the rule structure dictates legal sequences of letters tied to particular locations in the string. For example, in the grammar created by Reber (1967), all grammatical strings start with either TP, TT, VX, or VV. If Ss are classifying test strings because they know which letters are allowed in the first two positions, it is not clear what type of knowledge they are using to make this decision. They could be using rules, such as “all grammatical strings must begin with T or V”, “an initial T must be followed by P or T”, and “an initial V must be followed by X or V”. However, they could also be using knowledge about bigrams, such as “all training strings must begin with either TP, TT, VX, or VV”. In principle, a FSG could be created which produces strings that are so complex that they would appear random (e.g. a sufficiently complex FSG could produce strings which start with any of the letters of the alphabet). However, the FSGs used in AGL experiments are mostly not this complex. Shanks et al therefore suggest using a different type of grammar, namely a biconditional grammar, that allows rule and fragment knowledge to be disentangled.

3.2.3 Biconditional Grammars

Mathews et al (1989, Exp. 4) created a biconditional grammar (BCG), which is based on biconditional rules. A biconditional rule is a relationship between two propositions when one is true only if the other is true, i.e. 'p if and only if q'. From prior research (e.g. Bourne, 1970) it is known that people are capable of explicitly generating simple logical rules such as these biconditional rules. The BCG Mathews et al created was based on three letter-correspondence rules specifying which letters must occur in corresponding positions in the left and right halves of the string. The grammar generates strings of eight letters separated by a dot. The three rules in Mathews et al's study are T with X, P with C, and V with S. An example of a correct string would be TPPV.XCCS, whereas VSXX.SVCT would be incorrect because to be correct it would have to have a T in the seventh position.

3.2.3.1 St. John's study

St. John (1996) tested whether Ss are limited to learning adjacent (neighbouring) regularities (as in strings created by FSGs) or whether they can also learn non-adjacent (distant) regularities (as in strings created by BCGs) in implicit learning conditions (i.e. without knowing that there are rules to be learnt). In his experiment there were three different rules: (1) The intervening-0 rule placed contingent letter pairs in adjacent positions (i.e. 1a,1b,2a,2b,3a,3b,4a,4b). (2) The intervening-1 rule interleaved the contingent letter pairs so there was one irrelevant letter between each contingent pair (i.e. 1a,2a,1b,2b,3a,4a,3b,4b). (3) The intervening-3 rule placed the first letter of each contingent pair in the first half of the string, and the second letter of each pair in the second half (i.e. 1a,2a,3a,4a,1b,2b, 3b,4b). This intervening-3 rule was very similar to the BCG created by Mathews et al. The only difference is that in Mathews et al's study the letters could appear in all eight positions, whereas in this study the letters were permitted only in one half. For example, with the rule S <-> B, only S could be in the first half, and B in the second half, not the other way around.

Three groups of Ss were trained on the three rules. They were told to read the letters silently from left to right. Once the string disappeared from view they were to repeat the string again silently from memory. They were told that they would later be tested on their memory for the strings. There were three additional groups of Ss (one for each rule) who received no training but only did the classification test. These Ss acted as controls for any learning that might occur during the classification test itself. In the classification test, Ss were told about the existence of underlying rules, but not what the rules were, and asked to classify new strings according to whether or not they were 'acceptable'. All Ss received corrective feedback on their responses.

St. John found that the intervening-0 rule was learnt well, but the intervening-1 and intervening-3 rules showed only marginal learning. Since the intervening-1 rule was almost as difficult to learn as the intervening-3 rule, St. John concluded that Ss cannot learn about non-adjacent dependencies in implicit learning conditions. He supported this conclusion with a second experiment in which Ss were instructed to encode (memorise) the strings in two different ways, either normal left to right encoding, or alternating back and forth through the string: from first letter to fifth letter to second letter to sixth letter etc. He reasoned that if it is the left to right linguistic properties or the physical order of the letters in the stimuli that determine adjacency, the intervening-3 grammar should be difficult to learn under either encoding scheme. If it is the encoding order that determines adjacency, then the alternation encoding scheme will make the pairs adjacent and thereby permit good learning. He found that when Ss memorised the strings by alternating back and forth through the string, they were able to learn the intervening-3 grammar, confirming that adjacency is determined by the order in which the letters are encoded rather than by the physical order of the stimuli. In other words, it is unlikely that Ss will implicitly learn the rules of a BCG that has at least one intervening letter between the rule-related positions if they encode it from left to right.

3.2.3.2 Advantages of the BCG over FSGs

The BCG created by Mathews et al has several advantages over FSGs. The strings created by a BCG are less similar to one another than FSG strings because there is greater variation among letters across different positions in correct strings. There are, for example, no constraints on which letters can occur at the beginning or end of a string, and the letters can occur in any of the eight positions (as long as the corresponding letter is in the appropriate position). There are three intervening letters between the rule-related positions, which make it possible to disentangle rule and fragment knowledge by creating four different types of test strings: Grammatical and high similarity (GH), grammatical and low similarity (GL), ungrammatical and high similarity (UH) and ungrammatical and low similarity (UL). Grammatical strings are those that follow the same rules as the training strings. Ungrammatical strings violate these rules in either one or two letters. Test strings with high similarity have many fragments (or "chunks") that have already appeared in the training strings, whereas test strings with low similarity have few or none.

This measure of similarity is based on the competitive chunking model of Servan-Schreiber and Anderson (1990, see Appendix A) and is generally referred to as '*associative chunk strength*' (ACS). ACS is a measure of the frequency with which bigrams and trigrams in the test strings have appeared in the training strings. For example, the trigram GKX could appear for the first time in the test strings and never have appeared in the training strings, whereas the trigram KDF could have appeared three times in the training strings already, in which case a string containing KDF would be more similar than a string containing GKX (but see section 4.1.2.4 for exact details on how to calculate ACS).

3.2.4 Associative Chunk Strength (ACS)

Servan-Schreiber and Anderson (1990) assumed that the way we learn about regular stimulus fields like AGs is by sorting the material into some sort of hierarchy of "*chunks*", i.e. into contiguous substrings. There is abundant evidence that people faced with the task of memorising meaningless strings of letters will break the stimuli into chunks. For example, the 26 letters of the alphabet seem to be encoded into seven chunks: abcd, efg, hijk, lmnop, qrst, uvw, and xyz (Klahr, Chase, & Lovelace, 1983)[7].

Servan-Schreiber and Anderson tested whether the learning process in AGL experiments is chunking and whether the grammatical knowledge is implicitly encoded in a hierarchical network of chunks. They considered a chunk to be a single letter, a pair (bigram) or a triplet (trigram) of adjacent letters and

[7] *However, since mobile phones are widely used by the public nowadays and the letters of the alphabet are divided into different chunks on mobile phones (the letters 'abc' are on one button,' def' on the next, etc.), the alphabet may now be encoded into different chunks.*

predicted that the probability that a string would be classified as grammatical increases with the familiarity of a string, given the network of chunks acquired during the memorisation task. Thus, the more existing chunks it consists of, the more likely it will be judged grammatical.

3.2.4.1 Servan-Schreiber & Anderson's study

Servan-Schreiber and Anderson controlled which chunks Ss created during the memorisation task (training phase) by inserting spaces between chunks, e.g. T PP TX X VS. There were four groups of Ss: two saw well structured strings, one saw badly structured strings, and one saw unstructured strings. The training task was to memorise the strings. After the training task, Ss were told that the strings they had just memorised were all examples of “good strings”, and that they would now be asked to judge whether the next new strings were “good” or “bad”. The test strings were presented in an unstructured way. There were five different types of bad strings: two preserved the well-structured chunks, and three did not. Those bad strings that preserved the well-structured chunks were bad because they violated the chunk order constraints of the grammar. Those strings that did not preserve the chunks were either randomly generated strings using the same letters, or they were made ungrammatical by changing one of letters. Ss in the well-structured condition rejected strings that did not preserve the chunks more often (88.73%) than strings that did preserve the chunks (64.3%). This suggested that Ss do make use of chunking, and Servan-Schreiber and Anderson concluded that Ss were basing their classification on an overall familiarity (in terms of recurrent chunks) rather than on abstract knowledge about the rules of the grammar.

Based on the results of their study, Servan-Schreiber and Anderson formulated a theory of competitive chunking (CC). They also reported two simulations of experimental data providing evidence for their theory of competitive chunking (see Appendix A for detailed description of theory and simulations).

3.3 Summary

The major controversy in the AGL field about whether Ss learn the abstract rules or just recognise recurrent fragments has generated a good deal of research. However, most of the studies used FSGs, the nature of which makes it difficult to disentangle the role of abstract rule knowledge and fragment knowledge. Using a grammar based on biconditional rules makes it possible to distinguish between people who are classifying novel stimuli in the test phase according to the abstract rules of the grammar (grammaticality), and those who are classifying them according to the ACS of test exemplars to training exemplars (similarity). Very simple biconditional rules are used in this thesis.

4 Rule or Fragment knowledge?

The first experiment was designed to test whether subjects (Ss) in a standard artificial grammar learning (AGL) experiment (i.e. without being told in advance that there are rules to be learnt) are classifying test strings according to their rulefulness (i.e. whether or not they obey the rules) or according to their similarity to previously seen examples (i.e. as a result of their associative chunk strength, see section 3.2.4). Using the biconditional grammar (BCG) created by Johnstone and Shanks (2001), the role of abstract rule knowledge versus the mere memory for familiar fragments (2- or 3-letter chunks) can be analysed separately (see section 3.2.3).

Another aim of this experiment was to examine whether the knowledge people gain in AGL is implicit or explicit. This was tested by examining the relationship between the correctness of Ss' performance and their degree of confidence in their performance. A major controversy in the area of implicit learning is whether the knowledge of AGs – be it abstract knowledge or knowledge of letter chunks – is implicit (unconscious) or explicit (conscious and verbalisable; e.g., Perruchet & Pacteau, 1990; Shanks & St. John, 1994; Dienes & Perner, 1996).

The controversy is partly based on what is taken as the dividing line between what is conscious and what is unconscious. Cheesman & Merikle (1984) suggested that knowledge could be defined as unconscious if it was below what they called the "subjective threshold", which is the threshold below which Ss are unaware that they have any knowledge. They suggested that two criteria determine a lack of this 'metaknowledge' (knowledge of having knowledge): (1) better than chance performance when Ss think they are *guessing* (Cheesman & Merikle, 1984, 1986; Dienes & Altmann, 1997) and (2) *zero correlation* between performance correctness and confidence (Chan, 1992). Chan (1992) observed that Ss were just as confident in their incorrect responses as they were in their correct responses, although their performance showed high levels of correctness. This indicated that they had no metaknowledge. In a subliminal perception experiment, Cheesman & Merikle (1984, 1986) found that although Ss claimed they were guessing, they were actually performing at a level above chance. This again suggests that they weren't aware of what they were doing.

Dienes & Altmann (1997) used Reber's (1969) finite-state grammar (FSG) to test for the zero correlation and the guessing criteria. They found evidence for the guessing criterion, but none for the zero-correlation criterion. When their Ss were performing accurately they were more confident, indicating that they had some awareness of their knowledge (metaknowledge). However, their results showed a mean of 63% correct when Ss believed they were guessing, indicating that there was some implicit knowledge involved. It is not clear whether Ss learnt the abstract rules of the grammar and were responding according to those rules, or they were just classifying strings according to the presence of recurring fragments.

4.1 Method

4.1.1 Hypotheses

Our first null hypothesis is that Ss are classifying the test strings according to whether or not they follow the rules of the AG (grammaticality). Our second null hypothesis is that Ss are explicitly aware of how they are classifying the strings and that we will see no evidence for the guessing criterion and zero correlation criterion.

4.1.2 Subjects

16 voluntary Ss from the University of Southampton took part in the experiment; eight were assigned to the experimental group, and eight to the control group. There were eight males and eight females with ages ranging from 20 to 50 (mean 28.31).

4.1.3 Design

There was an experimental group and a control group. The training strings for the experimental group followed certain rules, while the training strings for the control group did not follow rules (see below for rules and creation of strings). The experimental group and the control group were presented with the same test strings.

This experiment used a mixed design with 2 x 2 x 2 levels. The dependent variable is test performance and the independent variables are rulefulness and similarity. The between-Ss factor is group (experimental and control), and the 2 within-Ss variables are rulefulness (ruleful and unruleful) and similarity (similar and dissimilar).

4.1.3.1 Rules

All strings were created from the letters D, F, G, L, K, and X. Strings of eight letters were produced. Each string was governed by three rules controlling the relationship between letters in positions 1 and 5, 2 and 6, 3 and 7, and 4 and 8. One possible set of rules could be D↔F , G↔L , K↔X. Hence, for the string to be ruleful (positive), when the letter D occupies one position, the letter F must appear in the corresponding position, where the letter G appears, the letter L must appear in the other position, and when a position contains a K, the corresponding position must be an X. See Figure 4.1.

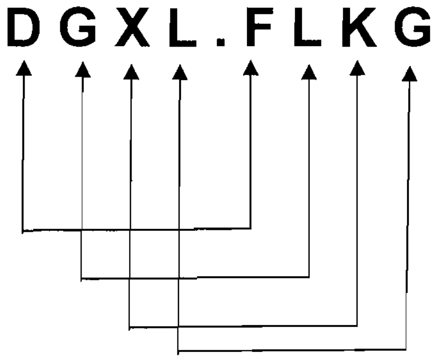


Figure 4.1: An example of a letter string using the bi-conditional rules D<->F , G<->L, K<->X

Each letter could appear in any of the eight positions. Ruleful strings consisted of four valid linkages, while unruleful strings consisted of three valid linkages and one invalid linkage. That means that the unruleful strings differed from the ruleful ones in one letter.

15 possible sets of three letter rules can be created from the letters D, F, G, L, K, and X. Each subject in each group was presented with a different version of the 15 possible rules.

4.1.3.2 Training strings

36 ruleful training strings were created for the experimental group using a subset of 18 of the 36 possible two-letter fragments (bigrams) that could be generated from the letters D, F, G, L, K, and X. The 36 unruleful training strings created for the control group contained all possible bigrams of the six letters without using double letter bigrams (e.g. LL). Two other strings, which we will call violation strings, were created for each training string. The violation strings were presented in the choice of three strings in the training phase and differed from the training string (i.e. the string that was to be memorised and chosen from the choice of three) in one and two letters. The two violation strings for the experimental group were constructed so that they comprised the same subset of bigrams and trigrams as the ruleful training strings. Each letter was evenly distributed across each of the eight possible positions in the training set for both the control and the experimental group.

4.1.3.3 Test strings

The set of test strings consisted of 48 new ruleful strings and 48 new unruleful strings. Within these test strings, half of them were similar to the experimental group's training strings and half were dissimilar. The similarity measure used was associative chunk strength (ACS) as defined by Servan-Schreiber & Andersen (see section 3.2.4. Thus four different types of test string (12 strings for each type) were generated: ruleful and high similarity (rh), ruleful and low similarity (rl), unruleful and high similarity (uh), and unruleful and low similarity (ul). There was no difference in similarity for ruleful versus unruleful test strings for either group.

The control group was presented with the same test strings as the experimental group. However, the control groups' training strings and the test strings had no rulefulness or similarity relationship.

4.1.3.4 Associative chunk strength (ACS)

Associative chunk strength (ACS) was calculated for each test string on the basis of the theoretical perspective on chunking presented by Servan-Schreiber and Anderson (1990, see Appendix A). Two measures of ACS were calculated: *Global ACS* for all fragments in a test string, and *Anchor ACS* for the initial and terminal fragments within each test string.

Global ACS was calculated by breaking down each test string into its constituent bigrams and trigrams and then calculating how many times each fragment had occurred in any location in the training items, and then dividing the totals by the number of fragments (i.e. 7 bigrams and 6 trigrams) (Johnstone & Shanks, 2001). For example, the test string LFGK.GDLX can be broken down into the bigrams LF, FG, GK, KG, GD, DL, and LX, and the trigrams LFG, FGK, GKG, KGD, GDL, and DLX. For example, the bigram LF could have appeared 4 times in the training strings, and the bigram FG 3 times, etc., so the global ACS would be calculated as follows: ((4 + 3 + 5 + 3 + 4 +2 + 5) / 7) + ((0 + 1 + 0 + 0 + 0 + 1) / 6) / 2 = 2.02.

Anchor ACS was calculated because prior research has shown that Ss are especially sensitive to 'extremities', i.e. beginnings and ends of strings (e.g. Reber, 1967; Servan-Schreiber & Anderson, 1990). Anchor ACS was calculated by adding the initial and terminal bigrams and trigrams and dividing by 4. For example, the anchor ACS for the test string LFGK.GDLX would be (1 + 1 + 0 + 1)/4 = 0.75.

Although the actual letters comprising the test string may be different for each subject (as 15 different sets of rules could be created, see section 4.1.2.1), the ACS scores would be the same, as the letters within the string always varied in the same way. For instance, the test string LFGK.GDLX (an example of a test string in the rule set D<->F, L<->G K<->X) with a global ACS of 2.02 and a anchor ACS of 0.75, would have the same ACS scores as the corresponding test string created using a different set of rules (e.g. the test string FLXG.XDFK from the rule set D<->L, F<->X, G<->K), as each fragment would have appeared the same number of times for each rule set.

Figure 4.2 shows the global ACS of each of the 48 test strings compared to the experimental groups' training strings from the least similar to the most similar, i.e. from the lowest ACS to the highest ACS score.

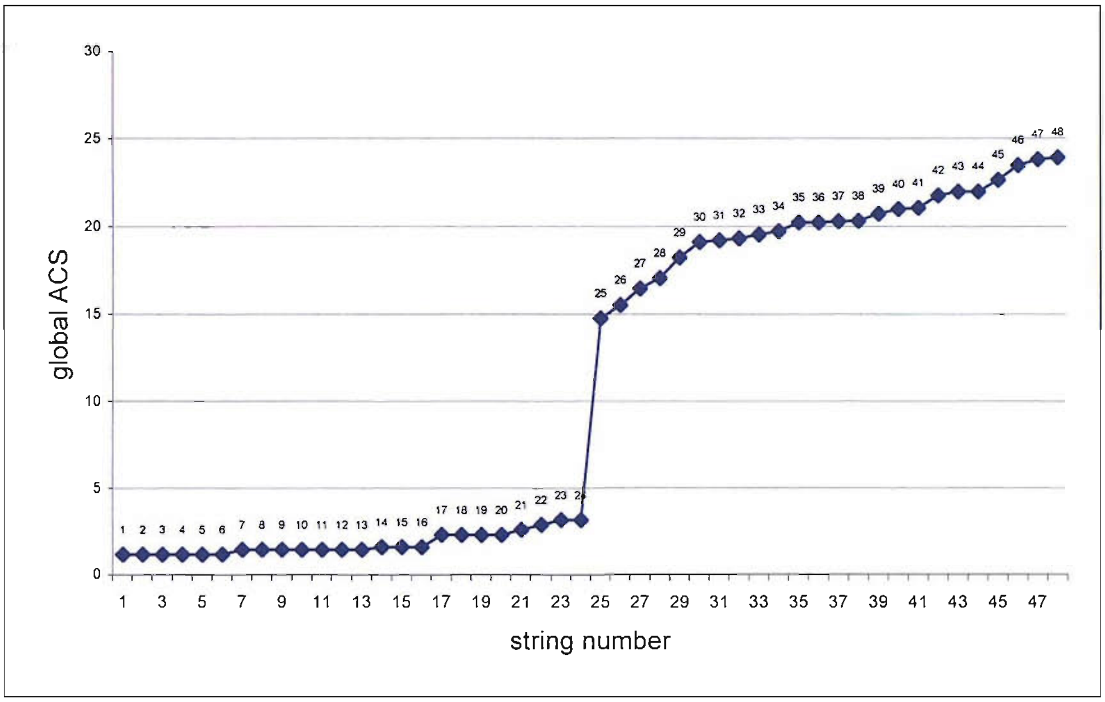


Figure 4.2: Global ACS of each test string for the experimental group

In Figure 4.2 it can be seen that half of the test strings are similar (the high similarity strings) to the training strings of the experimental group, i.e. have a large global ACS score (the right part of the figure), and half are not similar (dissimilar) to the training strings (low similarity strings), i.e. have a low global ACS score (the left part of the figure).

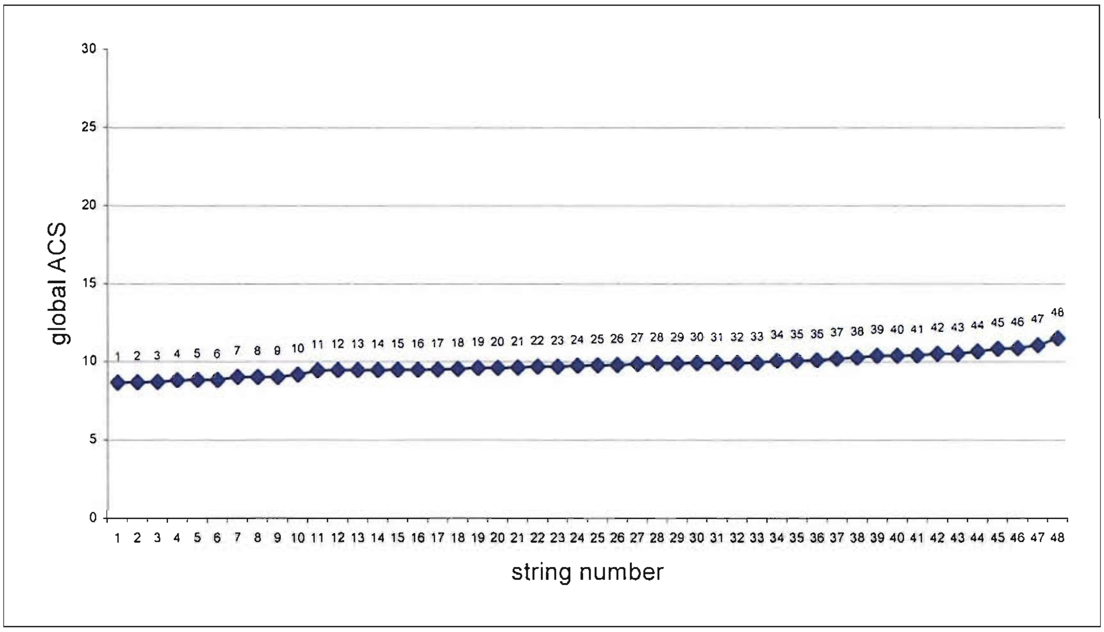


Figure 4.3: Global ACS of each string for the control group

In Figure 4.3 the global ACS scores for each test string compared to the control groups' training strings are shown. The ACS scores are approximately equal to each other at around 10, i.e. there is no similarity relationship between the control group's training strings and the test strings.

4.1.3 Apparatus

A computer program written in Java by Jorge Louis de Castro as a third-year undergraduate project at the University of Southampton was used for this experiment. The experiment ran on a Web browser (http://www.ecs.soton.ac.uk/~mtj00r/ApplicationClass.php) and everyone with Internet access could take part in the experiment. This allowed the potential subject sample to be distributed geographically, making it more representative of the general population than traditional local psychology experiments before the Web era (Krantz & Dalal, 2000). The program recorded all responses made by each subject in a cumulative database.

4.1.4 Procedure

Both the experimental and the control group were told that the experiment was testing their short-term memory for letter strings such as GFXK.LDKX. The training phase consisted of 72 trials (each set of 36 strings presented twice). On each trial a string appeared on the screen for five seconds, which the subject was asked to memorise. Then there was an interval of two seconds after which a list of three strings appeared. One of the strings was the string presented before; the other two strings were violation strings and differed from the training string in either one or two letters. Ss task was to choose the same string as they had just seen. They were told whether their choice was correct, and if they did not choose correctly, the correct string was shown again. Then the next trial started.

After this training phase, all Ss were told that the strings they had been memorising followed a particular set of rules. They were reassured that they probably hadn't noticed this, but that they may have picked something up without realising. In reality, only the experimental group had seen strings that followed rules; the control group had seen random strings (though they were told the same thing).

The test phase consisted of 96 trials in which Ss had to decide whether new strings that they had not seen before followed the same rules or not by choosing 'Yes, it follows the same rules' or 'No, it doesn't follow the same rules'. They were not told whether their answers were right or not.

They were then asked to indicate how confident they were that their response had been correct on a scale from 50% (complete guess) to 100% (completely sure)[8].

[8] *In hindsight, the scale may have been more intuitive if between 0 and 100, rather than 50 and 100. It was used as 50% indicates a 50-50 chance of getting it correct.*

4.2 Results and Discussion

4.2.1 Training phase

First, the data from the training phase were analysed. The mean reaction time (RT) of the control group was 577 msec, and the mean RT of the experimental group was 431 msec. RTs of longer than 4000 msec were excluded, as well as RTs of 0 msec, on the assumption that RTs of over 4 seconds implied that the subject was temporarily distracted, or that in the case of 0 msec the program had failed to record the data[9].

The experimental group's RTs were slightly faster than the control group's. An independent samples t-test for this group difference in mean RT was not significant, t(14)= -2.086, with a borderline probability of 0.056. Prior studies (e.g. Miller, 1958) have shown that structured stimuli (e.g. those following rules) are easier to learn than unstructured stimuli (e.g. random stimuli), which may account for the slightly faster responses.

In the training phase, Ss were presented with a letter string, and after an interval of two seconds they were asked to choose the string they had seen before from a list of three strings. Responses were counted as correct if Ss chose the same string as the one they had seen before. The mean percentage of correct responses was 90.8% for the control group and 91.15% for the experimental group. An independent samples t-test, t(14)=0.127, $p<0.05$, indicated that the control and experimental group were not performing significantly differently on the memory task.

[9] *4 seconds (4000msec) was chosen as the cut-off point as most (95%+) RTs were either below 4 seconds or much longer than 4 seconds.*

4.2.2 Data analyses

The data were analysed in different ways:

(a) A*ccording to correctness*: Ss classified ruleful (rh and rl) strings as ruleful, and unruleful (uh and ug) strings as unruleful.

(b) *According to similarity*: Ss mistakenly thought similar strings were ruleful and thus classified similar (rh and uh) strings as ruleful, and dissimilar (rl and ul) strings as unruleful, regardless of their rulefulness.

In the following sections and from here forth, the different ways of analysing the data will be referred to as 'analysed according to rulefulness' and 'analysed according to similarity'. 'Correct according to rulefulness' will refer to case (a), i.e. when the stimulus was classified correctly according to rulefulness; 'correct according to similarity' will refer to case (b), i.e. when the stimulus was classified according to similarity.

4.2.3 Test phase

The following analyses focus on the responses made in the test phase. The mean RT of the control group was 612.5 msec, and the mean RT of the experimental group was 544.6 msec (again excluding RTs of 0 msec and those over 4000 msec). An independent samples t-test showed no significant difference, t(14)= -0.689, p>0.05, indicating that the two groups were not performing differently on RT.

The mean percentages of test strings classified as 'Yes, it follows the same rules' were analysed for each of the four types of test items: ruleful and high similarity (rh), ruleful and low similarity (rl), unruleful and high similarity (uh) and unruleful and low similarity (ul). These mean percentages are shown in Figure 4.4.

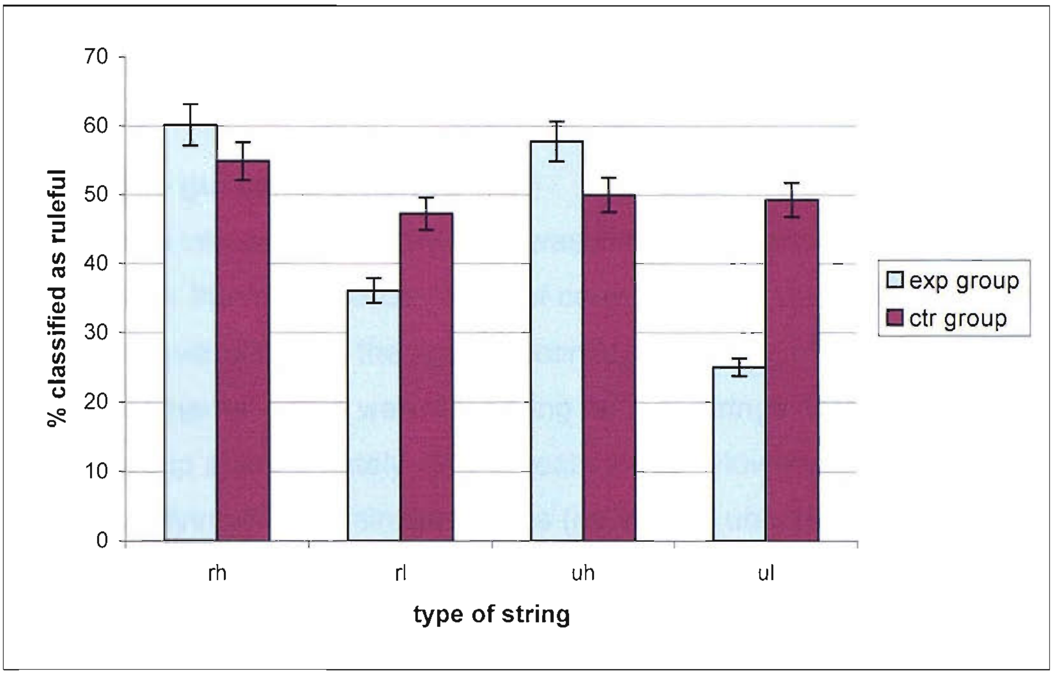


Figure 4.4: Mean percentage of test strings classified as ruleful ("Yes, it follows the rules")

To test whether Ss were classifying test strings based on their rulefulness or on their similarity to the training strings, a repeated measures ANOVA with group as between-Ss variable, and rulefulness and similarity as within-Ss variables was carried out. There was a significant effect of similarity, F(1,14)=10.719, $p<0.05$. This suggests that Ss were classifying the test strings based on their similarity to the training strings. There was also a significant interaction between similarity and group, F(1,14)=5.924, $p<0.05$, showing that the two groups were responding differently as a function of similarity.

There was no significant effect of rulefulness F(1, 14)=1.247, $p<0.05$, indicating that Ss were not basing their classification on the rulefulness of the strings. Neither was there a significant effect of group, F(1, 14)=1.095, $p<0.05$, showing that the experimental and the control group were not classifying differently according to rulefulness.

These results suggest that Ss in the experimental group were classifying the test strings according to their similarity to the training strings, and not according to whether or not they follow the rules (rulefulness). This finding supports the supposition that people are simply memorising fragments of the strings, not learning abstract rules.

4.2.3.1 The guessing criterion

To examine whether Ss' knowledge was implicit (unconscious) or explicit (conscious), the mean percentages of correct classifications when Ss thought they were guessing (i.e. they gave a confidence rating of 50) were analysed. The experimental group was classifying 46% of strings correctly, and the control group approximately 42% of test strings. However, experimental Ss were classifying 57% of similar strings (i.e. rh and uh strings) as ruleful and dissimilar strings (i.e. rl and ul strings) as unruleful when they indicated that they were only guessing. In other words, experimental Ss thought that 57% of similar strings followed rules. Actually, these strings did not follow the rules, but were just more similar (i.e. had a higher associative chunk strength) to the training strings. Control Ss were classifying 52% of similar strings as ruleful. These data are shown in Table 4.2.3.1.1.

	Rulefulness	**Similarity**
Experimental group	46.00	56.82
Control group	41.55	52.26

Table 4.1: Mean percentages of classifications when Ss thought they were guessing (the guessing criterion)

However, an independent samples t-test showed no significant differences, t(12)=1.223, p>0.05, indicating that statistically, the two groups were not performing differently. One-sample t-tests revealed that neither group was performing significantly differently from chance. In other words, when Ss thought they were guessing and gave a confidence value of 50, they really were guessing.

4.2.3.2 The zero correlation criterion

The confidence ratings were divided into 6 groups: the first included all responses of 50% (i.e. when the Ss thought they were guessing), the second those between 51 and 60, the third those between 61 and 70 and so on.

Figure 4.5 shows the mean percentage of correct responses for experimental and control group for each level of confidence. Accuracy and confidence were not correlated. To test the zero correlation criterion, a repeated measures ANOVA with group as between-Ss variable and the six levels of confidence as within-Ss variable was conducted as a function of accuracy. There were no significant effects.

Next the data were analysed according to similarity (see section 4.2.2). Figure 4.6 shows that for similar and dissimilar strings the experimental group was more confident about their responses.

A repeated-measures ANOVA with group as between-Ss variable and the 6 levels of confidence as within-Ss variable revealed no significant effects for responses according to similarity indicating that the two groups were not performing differently.

It can be seen in figure 4.6 that the experimental group was performing at a higher level than the control group at each level of confidence. Percentage of responses according to similarity and confidence were positively correlated for the experimental group ($r = 0.410$, $p<0.05$) indicating that experimental Ss were more confident on the similarity measure (i.e. Ss thought similar strings were ruleful and dissimilar strings were unruleful and were confident about their answers). When Ss in the experimental group indicated that they were most confident (confidence level of 91-100) they were classifying 86.5% of the strings according to their similarity (i.e. Ss classified similar (rh and uh) strings as ruleful and dissimilar (rl and ul) strings as unruleful).

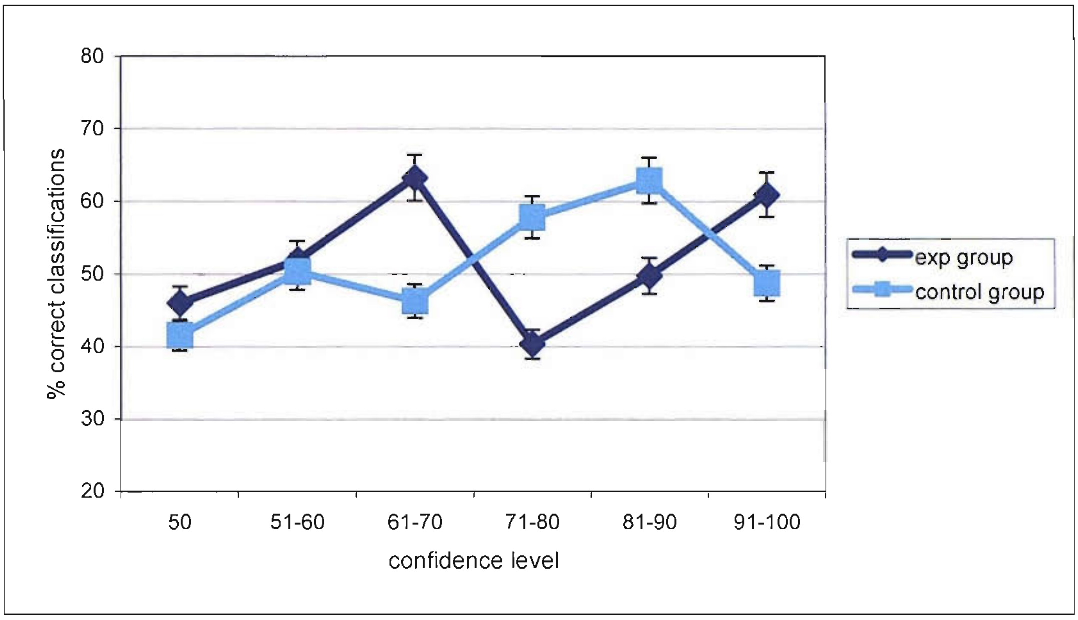


Figure 4.5: Mean percentages of correct responses according to rulefulness

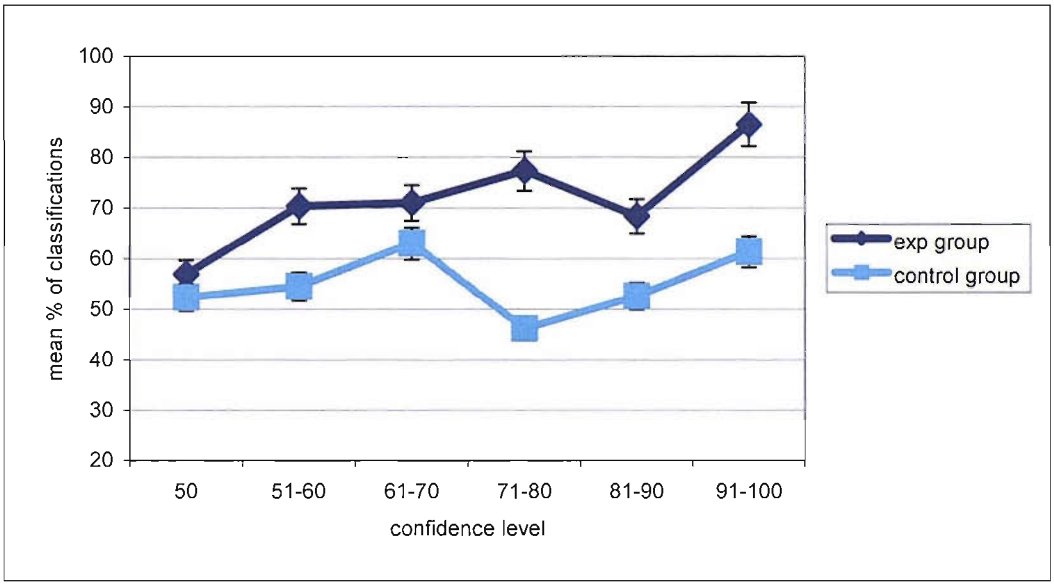


Figure 4.6: Mean percentage of responses according to similarity

If one understands this zero correlation criterion as a measure of conscious knowledge, these results suggest that Ss in the experimental group had some meta-knowledge[10] about what they were doing. In other words, Ss seemed to be somewhat aware of doing something correctly (indicated by higher confidence), but they did not seem to know how they were doing it.

[10] Metaknowledge = Knowledge about knowledge. Knowing you have some knowledge, but not necessarily what the knowledge is.

Further results confirm that there was a certain degree of metaknowledge: When analysing subjects' classifications according to similarity there was a significant interaction between confidence and linear trend, F=9.473, $p<0.05$ and a significant interaction between confidence, group and linear trend, F=18.566, $p<0.01$. This indicates that the experimental and control groups were performing differently, and that there was a linear trend between the mean percentages of classifications (when analysing according to similarity) and confidence level. In other words, when Ss were less confident and gave a lower confidence rating, they were also performing at a lower level (according to similarity), and when they were more confident and gave higher confidence ratings, the mean percentage of responses analysed according to similarity was also higher. This linear trend can be seen in Figure 4.6 in the experimental group's data.

Correlational analyses showed that global ACS and anchor ACS were both positively correlated with confidence for the experimental group: global ACS $r = 0.112$ and anchor ACS $r = 0.102$, $p<0.01$. When the global and anchor ACS scores were both high, i.e. the test string was very similar to the training strings, Ss' confidence about whether their responses were correct or not was also high (and the same respectively for low similarity and low confidence). This again suggests that Ss were classifying the test strings according to how similar they were to the training strings, and were confident that they were classifying "correctly" on these strings. In reality, subjects were not classifying the strings according to the underlying rules as the task required them to, but rather, they were incorrectly thinking that similar (ruleful and unruleful) strings followed rules, and the dissimilar (ruleful and unruleful) strings did not follow rules.

4.3 Conclusions

Ss seemed to possess virtually no knowledge of the rules of the artificial grammar, and there was no correlation between their performance and their confidence ratings. Their accuracy in detecting whether strings followed rules was at chance level. Instead, Ss seemed to be classifying the test strings according to their similarity to the training strings.

Furthermore, Ss in the experimental group were more confident of their accuracy when they were classifying strings according to their similarity to the training strings. Although Ss were not classifying correctly according to the rules, they were more confident that they were performing accurately when classifying similar strings as ruleful and dissimilar strings as unruleful. Ss in the experimental group felt more confident that the test string followed rules when it was similar to the training stimuli, i.e., when they looked more *familiar.* Conversely, they also felt more confident that the string didn't follow any rules when it was less similar to the training strings, i.e. when they looked less familiar.

5 Positive (Ruleful) and Negative (Unruleful) Instances in Artificial Grammar Learning

As noted earlier in Chapter 2, most artificial grammar learning (AGL) experiments were originally motivated by natural language learning, and have therefore trained subjects (Ss) with only positive (ruleful) strings. In the subsequent test phase, Ss have to distinguish between positive (ruleful) and negative (unruleful) strings, although they have never seen negative strings until that point. Unless the rule is clearly obvious, it is crucial -- if successful *learning* of the artificial grammar is to occur-- that the training includes both positive and negative instances. Learners need to sample stimuli in which the relevant features are present (positive) as well as stimuli in which they are absent (negative) if they are to discover and then respond selectively to the relevant features only, ignoring the irrelevant features. Otherwise there is no empirical way to find the basis for a distinction between positive and negative.

Only very few AGL studies have used both positive (ruleful) and negative (unruleful) instances (Brooks, 1978; Whittlesea & Dorken, 1993; Dienes, Broadbent, & Berry, 1991; Dienes, Altmann, Kwan, & Goode, 1995). Moreover, because the interest of even these few was not in the role of negative instances per se, but in whether Ss could distinguish between grammars, two different grammars tended to be used, rather than positive and negative instances of one grammar. A positive instance of one grammar can be seen as a negative instance of the other grammar and vice versa. But this use of two different grarnmars creates several problems, which will be discussed later (see sections 7.1.1 and 7.1.2).

5.1 Brooks' (1978) series of studies

Brooks (1978) wanted to demonstrate that when Ss are encouraged to use information about individual items, they draw analogies between individual items. He used AGs in his experiments to emphasise "the learning of specifics as a basis for the later generalisation of an extremely complicated concept" (Brooks, p. 170, 1978). Brooks trained Ss on three paired-associate lists: Each letter string was associated with a particular word; for example, the letter string VVTRXRR was paired with the word 'Paris'. Half the strings were generated from grammar A, half from grammar B. Those from grammar A were paired with 'Old-World' items (e.g. Paris, elephant, Rome), those from grammar B with 'New-World' items (e.g. Montreal, possum, Detroit).

In the first experiment Ss knew nothing about the existence of two grammars, nor the relevant categories during training, nor that there were any rules to be learnt. The training task was to respond to the letter string with the correct word, i.e. to memorise which word was paired with which letter string. The responses were divided into two categories, animals (e.g. elephant, possum) and cities (e.g. Paris, Montreal), varying *independently* of the Old-World and New-World distinction. The training phase ended when Ss completed one trial without error for each of the three lists. Ss were then asked if they had noticed that there were two different types of letter strings: None of the Ss had noticed. They were then asked if they had noticed two different types of response; most Ss answered that the responses were all either cities or animals.

Ss were then told that there was a distinction between the strings that had been paired with Old-World items and New-World items. They were given a test stack of 30 cards with novel letter strings printed on them. Ten of the strings were generated from grammar A (i.e. the one that had been paired with Old-World items), ten from grammar B (i.e. New-World items), and ten were random letter strings using the same letters. Ss were required to sort the 30 cards into three piles and were told that 10 were Old-World items, 10 were New-World items and 10 did not belong to either category. Although Ss indicated that they did not know what they were doing and were just guessing, they were able to distinguish the categories from one another at a level significantly above chance (60-64.4% correct performance, Brooks takes chance to be 33%[11]).

Brooks conducted a further experiment in which he made the presence of two different types of stimuli quite obvious by pairing the letter strings from grammar A with the word "city" and the strings from grammar B with the word "animal". The experiment was again presented as a memory test, and training finished when all three lists had been learnt. In the test phase, Ss were presented with the same test stack of 30 new letter strings as the previous group, and were asked to sort them into three piles: animal, city, or neither. The results of this group (60.4%-65.8% correct performance) are very similar to the results of the previous group. So making the presence of two different categories apparent in the training phase does not affect the way Ss perform at test.

In a further experiment, Brooks explicitly told Ss that they were to learn to distinguish letter strings that were generated by two different sets of rules ("grammars"). They were shown the same training strings as the previous groups and were required to categorise each item as following grammar A or grammar B. After each trial, Ss were told whether they were right or wrong. In the previous experiments the responses were passively paired with the letter strings, while in this experiment Ss received active feedback after each response. They continued until they could categorise all three lists correctly (which, by chance, averaged about the same number of trials as the previous groups). They were then given the same test stack of 30 new letter strings as the previous groups and asked to sort them into three piles (corresponding to grammar A, grammar B, or neither).

Their success rate was between 45.4% and 50.8%. These Ss did not generalise as well to new items as the previous Ss who were not told about the category distinction. In fact, they performed at almost the same level as a group who received no training at all, i.e. who only did the test phase, and a

[11] *See Section 7.1.4 for discussion*

group in which the New-World and Old-World pairings were randomly assigned to the stimuli in training. The group which had received no training were categorising between 37.1% and 51.1% correctly, and the group which had received random-pair training were performing at between 40.4% and 47.9%. These Ss were apparently sorting the cards according to the structural similarity of items to one another, i.e. items that looked similar to each other were put into the same pile[12].

At first glance these experiments seem to suggest that Ss are able to distinguish between the two grammars at levels better than chance only when they are *not* told about the existence of two different sets of rules during training. However, as Brooks suggests, the regularities in the material are so complex that it is difficult for Ss to grasp them in any reasonable amount of time, and "certainly not under the conditions of successive presentation" (p.175, 1978). He also observes that the stimuli are sufficiently similar to one another, so that those Ss who know about the existence of rules would have poor incidental memory for them, i.e. would not be storing individual stimuli, since they would be concentrating on finding rules. Those Ss who were not informed of the existence of rules, however, were doing a paired-associate task, which consists of memorising the strings, so by virtue of their task they would have good incidental memory for them.

Brooks thus concludes that Ss were "drawing analogies" between previously memorised individual training stimuli and test stimuli. He suggests that Ss might simply be thinking that, for example, MRMRV looks similar to MRRMRV and since they had learnt that MRRMRV is associated with Vancouver which is New-World, they would also call MRMRV a New-World item "by simple analogy". In other words, Ss who had been told to look for rules were in a

[12] *In the previous experiments, Brooks takes 33% to be chance level performance, since there are three categories (Grammar A, Grammar B, and neither/random). However, he didn't establish a baseline chance level. Ss who received no training and were just required to sort the 30 stimuli into three piles, were apparently basing their classifications on the structural similarities, and were classifying at a level above 33%, namely between 37.1% and 51.1%. This can and should be taken as baseline chance level performance. (See Section 7.1.4)*

worse position than Ss who hadn't for two reasons; (1) the rules were much too complicated for them to learn under those circumstances, and (2) because they were looking for rules, they had not been memorising individual items, which could have later helped them in drawing analogies.

5.2 Whittlesea & Dorken's (1993) studies

Whittlesea & Dorken (1993) were interested in investigating whether performance in implicit learning experiments (e.g. Reber, 1976) reflects automatic abstraction of the structure underlying stimuli or whether it instead reflects task-dependent encoding of particular experiences of stimuli.

They showed Ss instances from two different finite-state grammars (FSGs) and instructed them on how they were to process the stimuli. The strings were designed to be pronounceable and Ss were to speak the strings from one grammar (A) and spell the strings from the other (B) during training. Examples of strings are ENOBGAD and OLFELID. First, all 40 training instances of one grammar were presented in random order. Then all 40 instances of the other grammar were presented in random order. This procedure was repeated using a different random sequence (within the 40 instances of one grammar), for a total of 160 training trials.

Just before the test phase Ss were informed that the items they had spelled earlier were generated by one set of rules, while the items they had spoken were generated by a second set of rules. They were then shown novel items which they again had either to spell or speak, and then to classify as either 'Spell' or 'Speak' (Whittlesea & Dorken used the term 'Say' but for this thesis I have used 'Speak'). The critical manipulation was that Ss were now required to speak half of the test items belonging to the Spell grammar, and spell the rest, and to spell half the Speak items and speak the rest. It was carefully explained to the Ss that the spelling or speaking of the words in the test phase was completely unrelated to whether the item belonged to the 'Speak grammar' or the 'Spell grammar', and that, in fact, half the items actually belonging to the Speak grammar would have to be spelled now, and vice versa.

In the test phase, Ss were able to discriminate members of the Speak grammar from members of the Spell grammar at levels above chance. When the processing contexts were matched (e.g. items from the Spell grammar were spelled) Ss were classifying at a level of 66% correct, and when the processing contexts were mismatched (e.g. items from the Speak grammar were spelled) they were getting 61% of trials correct. A third of the stimuli in the test phase were common to both grammars, i.e. they could be created by both grammars. These items were more likely to be assigned to the grammar that matched the processing context, i.e. when they were required to spell the item, they were more likely to assign it to the Spell grammar.

Whittlesea & Dorken conclude that Ss were able to discriminate the two grammars by “feelings of familiarity” induced by task context. Because processing is more “fluent” and thus seems more familiar when a test experience resembles prior experiences (Jacoby, Kelley, Brown & Jasechko, 1989), this feeling of familiarity could be used to discriminate the grammars. A test string from the Spell grammar would seem familiar if spelled at test, and would be correctly classified as belonging to the Spell grammar, whereas if it were pronounced at test it would feel unfamiliar and would be correctly categorised as not belonging to the Speak grammar.

In the test phase of a further experiment, Whittlesea & Dorken presented Ss with negative strings as well as positive strings from grammars A and B. The negative test strings were created using the same letters as positive strings, but violated both grammars in at least one sequence rule. After training on the same training set and under the same conditions as in the first experiment, Ss were told that some items conformed to the same rules as the items they had just seen, and others didn’t. Their task was to say whether the novel strings were ‘good’ or ‘bad’. As in the previous experiment, half of the test items were spelled and half were spoken.

Positive items were classified as good with a higher probability (p = 0.64) than negative items (p = 0.27), which Whittlesea & Dorken take as a demonstration of sensitivity to the “Goodness” (i.e. rulefulness). Positive items were judged good with greater probability when the processing contexts were the same (p = 0.66) than when they were different (p = 0.49). For example, strings from grammar A that had previously been spelled were considered good more often when they were again spelled, than when they were spoken.

Whittlesea & Dorken argue that this effect of task context (spell or speak) demonstrates a sensitivity to rulefulness that is mediated by representations of the training items, which have preserved highly specific information. In other words, they suggest that the rulefulness effect is due to Ss’ memory for specific information about how they experienced the stimuli in the training phase. This is in agreement with Brooks's interpretation of his data. It seems that here, too, Ss are simply remembering specific instances of the training items and basing their classifications of novel stimuli in the test phase on those remembered instances.

5.3 Dienes, Altmann, Kwan, & Goode's (1995) study

Dienes, Altmann, Kwan, & Goode (1995) also used two different grammars. They were investigating how conscious Ss are of what they have learned in an AGL experiment. They trained Ss first on one grammar and then on a different grammar. Ss were given a sheet of paper with strings from grammar A and asked “to study the strings as carefully as possible”[13]. This sheet was replaced by another sheet with training strings from grammar B. In the subsequent test phase Ss were told that the letter strings conformed to complex rules: one set of rules for the first sheet and a different set of rules for the second sheet. Ss then received a test sheet on which they were told that a third of the strings were like the strings on the first sheet, a third like the

[13] *This task is quite vague, and it cannot be controlled what Ss would have understood from these instructions, but for purposes of this thesis the experiment has been included.*

strings on the second sheet, and a third were like neither. Half the Ss were asked to tick only the strings that were like those on the first sheet, and half the Ss were asked to tick those strings that were like those on the second sheet.

With this procedure Dienes et al wanted to test whether Ss have intentional control over the knowledge they have gained in the training phase. Knowledge of an AG could be unconscious in the sense that it is applied in an *obligatory* way, regardless of whether Ss *want* to apply it, i.e. no intentional control. Dienes et al. reason that if the knowledge is unconscious in this sense, Ss should have difficulties differentiating the two grammars in the test phase. So if Ss were required to tick strings from the first grammar, the familiarity of those from the other grammar may intrude and they may also be ticked. This would indicate that Ss are using the acquired knowledge contrary to their intentions.

Dienes et al. found that Ss performed the task quite successfully, with means of 0.53 to 0.59 correct. The proportion of strings ticked incorrectly was subtracted from the proportion of strings ticked correctly to give a measure of intentional control. A score of zero would then indicate no intentional control and a score of 1 would indicate complete intentional control. They found that the scores were significantly greater than zero, indicating that Ss had a significant amount of intentional control.
Dienes et al also noted that Ss had virtually no obligatory knowledge (knowledge not under intentional control). They compared the performance of the experimental group with a control group, in which Ss had received no training and were simply told that the order of letters in some but not all of the strings obeyed a complex set of rules, and that they should tick only those strings that obeyed the rules. A baseline measure was calculated for the control group for each grammar: the number of ticks to positive strings of one grammar divided by {the number of ticks to positive strings of that grammar plus number of ticks to negative strings}. The means were 0.43 for grammar A and 0.50 for grammar B. The comparison of these baselines to the experimental group's obligatory knowledge revealed no significant differences.

Dienes et al conclude that the knowledge their Ss gained was largely under intentional control, i.e. available to consciousness.

However, the results can also be explained as memory for fragments: Ss could simply be remembering fragments of letters from the training strings. Since the strings from each grammar were studied simultaneously (i.e. a list of strings from one grammar first, then a list of strings from the other grammar), Ss could quite easily also be remembering which sheet of paper (the first or the second) the item was on.

5.4 Dienes, Broadbent, & Berry's (1991) study

Dienes, Broadbent, & Berry (1991) showed Ss negative as well as positive strings of a FSG grammar. Theirs is the only experiment which used positive and negative instances of one grammar instead of two different grammars. They hypothesised that providing a distinction between positive and negative instances may induce a strategy that inhibits implicit learning and promotes explicit learning.

Half their Ss saw only positive examples, the other half saw positive and negative instances. In the training phase the total set of strings was shown six times in a different random order each time. The positive group saw 20 positive instances in black, the positive/negative group saw 20 positive instances in black and 20 negative instances in red without being told what the colour distinction meant. They were told that it was a simple memory experiment, and that their task was to “learn and remember as much as possible about all 20 (40) items”.

After the training phase, the positive group was informed that the strings followed a complex set of rules; the positive/negative group was told that the black (positive) strings followed a complex set of rules, while the red (negative) strings violated those rules in some way. In the test phase Ss were presented one by one with black instances of 25 positive and 25 negative strings and asked to classify them as following the rules (positive, ruleful) or not following the rules (negative, unruleful). The 50 instances were repeated once in a different random order. They were also given a free report test, asking how they had decided whether an item followed the rules and what strategies they had used. The results showed that the two groups were performing significantly differently, with the positive group classifying at 65% correct and the positive/negative group classifying at 60% correct, which is still above chance-level of 50% correct.

Those Ss who received only positive instances were performing better than those who received positive and negative instances. Despite their lack of confidence, even the positive/negative group were classifying the test items at a level above chance. Dienes et al. concluded that the presence of negative instances interfered with Ss' performance and impaired both implicit and explicit learning (if performance is taken as a measure of implicit learning and free report as a measure of explicit learning). Ss reported that they couldn't remember which instances had appeared in black and which had appeared in red in the training phase[14].

If Ss were memorising specific instances from the training set, as also suggested by Brooks's experiments and Whittlesea & Dorken's experiments, then the presence of negative instances in this experiment would interfere with performance at test. Ss in the positive/negative group had memorised both positive and negative instances, but not their differentiation (red or black), since they had not been told what the colours signified. Ss in the positive group would have an advantage over the positive/negative group, since they would know that all instances that they remembered were positive instances, and wouldn't have to also remember whether an instance had been red or black. In effect, the presence of colours is just an additional bit of information to remember.

[14] *This also seems to suggest that Ss were trying to remember specific instances from the training phase*

5.5 Conclusions

Initially, Reber concluded that Ss were implicitly learning the abstract rules of the grammar. If Ss were acquiring abstract knowledge, the presence of negative instances in the training phase ought to aid learning, as it would provide an empirical basis on which to define rules. However, in the above experiments, in which both positive and negative instances were used, negative instances seem to interfere with learning. This can be explained if AGL experiments are just fragment memorisation tasks not involving rule learning at all. The presence of negative instances – whether from a different grammar, or from the same grammar – interferes with this rote memorisation, and consequently performance levels drop. In the next chapter an experiment is conducted similar to Dienes et al's (1991) study where Ss are trained on both positive and negative strings of the same grammar.

6 The effect of negative instances in a standard AGL experiment

In the few artificial grammar learning (AGL) experiments where Ss have been presented with positive and negative instances of a grammar (or instances from two different grammars) the results are rather varied. Under certain circumstances, Ss perform well above chance, whereas under other circumstances negative instances are detrimental to learning. In this experiment, Ss were again told that they were taking part in a short-term memory experiment and initially not told anything about the underlying rules. The experimental group was presented with positive and negative strings (differentiated by different background colours) while the control group only saw negative strings. If Ss are trying to complete the test phase by memorising fragments from the training strings, we should find similar results in this experiment to the results Dienes et al (1991) found in their experiment (see Section 5.4), namely that negative strings interfere with learning.

6.1 Method

6.1.1 Subjects

22 voluntary Ss took part in the experiment. 5 were female, 17 were male. This experiment was run on the Internet. An email informing people where the experiment could be found on the Internet was sent to several departments in the University of Southampton, UK. 14 Ss were assigned to the experimental group, and 8 to the control group.

6.1.2 Design

The same biconditional grammar (BCG) was used as in the experiment detailed in Chapter 4.

6.1.2.1 Training strings

The *experimental group* was presented with 36 positive strings and 36 negative strings. The positive training strings were the same as the strings created for the experimental group in The experiment described in Chapter 4, the negative training strings were the same as the strings created for the control group in The experiment described in Chapter 4. The positive strings were displayed on a green background, the negative strings were displayed on a red background.

The *control group* saw the same strings as the control group the experiment described in Chapter 4 (i.e. negative strings) on random green and red backgrounds.

6.1.2.2 Test strings

The 96 test strings consisted of 48 new positive and 48 new negative strings and were the same as those used in The experiment described in Chapter 4. They were all displayed in black on a white background.

6.1.3 Apparatus

The computer program was the same computer programme written in Java by Jorge Louis de Castro as described in Section 4.1.3. It was modified by Michael O. Jewell to allow the use of different colour background screens for different stimuli (which was not necessary in the experiment described in Chapter 4). The program recorded all stimuli and all responses made by each subject.

6.1.4 Procedure

All Ss were told they were taking part in a short-term memory experiment. They were also told that the screen would change colours, but that they need not worry about it for this part of the experiment. Subjects were presented with a string which remained on the screen for 5 seconds. For the experimental group, the positive strings were displayed on a green background and the negative strings were displayed on a red background. For the control group all (negative) strings were displayed on random green and red backgrounds. Then three strings were presented on a white background, one of which was identical to the string they had just seen. Their task was to choose the string they had just seen. Ss were told whether their choice was correct or not, and if incorrect, they were shown the correct string again. The training phase was 72 trials. For the experimental group this consisted of 36 positive and 36 negative strings. The control group saw the set of 36 negative strings presented twice.

After the training phase, all Ss were informed that the green strings all followed a set of rules, whereas the red strings did not. In reality, only the experimental group's green strings followed rules, whereas the control group's green strings were random. Their task then was to decide whether or not the following 96 test strings (48 new positive and 48 new negative strings) followed the same rules as the green strings they had previously seen. Ss were asked whether they thought the test string followed the rules and were required to click on Yes or No. They were reassured that they needn't worry if they couldn't remember or felt that they didn't know, as this was normal.

Ss were not told whether their answers were correct or not in the test phase. After each trial Ss were also required to give a confidence rating of between 50% sure and 100% sure that their answer was correct.

At the end of the experiment Ss were asked to write down any strategies they were using, or any other tactics they were using to decide how to classify the stimuli, and any other comments they might have no matter how irrelevant they thought their comments might be.

6.2 Results and Discussion

6.2.1 Training phase

The mean reaction times (RTs) in the training phase were 506 msec for the experimental group and 476 msec for the control group. The mean percentage correct was 90% for the experimental group and 87% for the control group. T-tests show no significant difference in RT, t(20) = 0.486, p>0.05, or correctness, t(20) = 1.247, p>0.05, for experimental and control group, indicating that the two groups were not performing differently in the training phase.

6.2.2 Test phase

The mean RTs in the test phase were 678 msec for the experimental group and 863 msec for the control group; the difference was not significant, t(20) = -1.014, p>0.05.

The experimental group had a mean percentage correct of 50% at test and the control group had a mean percentage correct of 53%. The experimental group classified 58% of similar strings as ruleful or dissimilar strings as unruleful, while the control group classified 54% of similar strings as ruleful and dissimilar strings as unruleful. These percentages are shown in Table 6.1.

	Rulefulness	**Similarity**
Experimental group	50.00	57.96
Control group	52.86	53.65

Table 6.1: Mean percentages of classifications

One-sample t-tests showed that the mean percentages were not significantly different from chance, indicating that neither group was performing better than chance. An independent measures t-test showed no significant difference in correctness for control and experimental group, t(20) = -1.063, nor for when the data was analysed according to similarity, t(20) = 1.368. This indicates that the experimental and control group were not doing the task differently.

The mean percentages of test strings that Ss classified as ruleful (i.e. "Yes, it follows the rules") for each type of string are shown in numerical form in Table 6.2. The four string types are ruleful (positive) and high similarity (rh), ruleful (positive) and low similarity (rl), unruleful (negative) and high similarity (uh), and unruleful (negative) and low similarity (ul) (for similarity measure see Sections 3.2.4.3. and Appendix A).

	rh	rl	uh	ul	average
Experimental group	59.88	32.34	46.53	46.89	47.25
Control group	42.47	36.13	39.68	31.15	38.15

Table 6.2: Mean percentages of each type of test string classified as ruleful ("Yes, it follows the rules")

Figure 6.1 shows these mean percentages in graphical form.

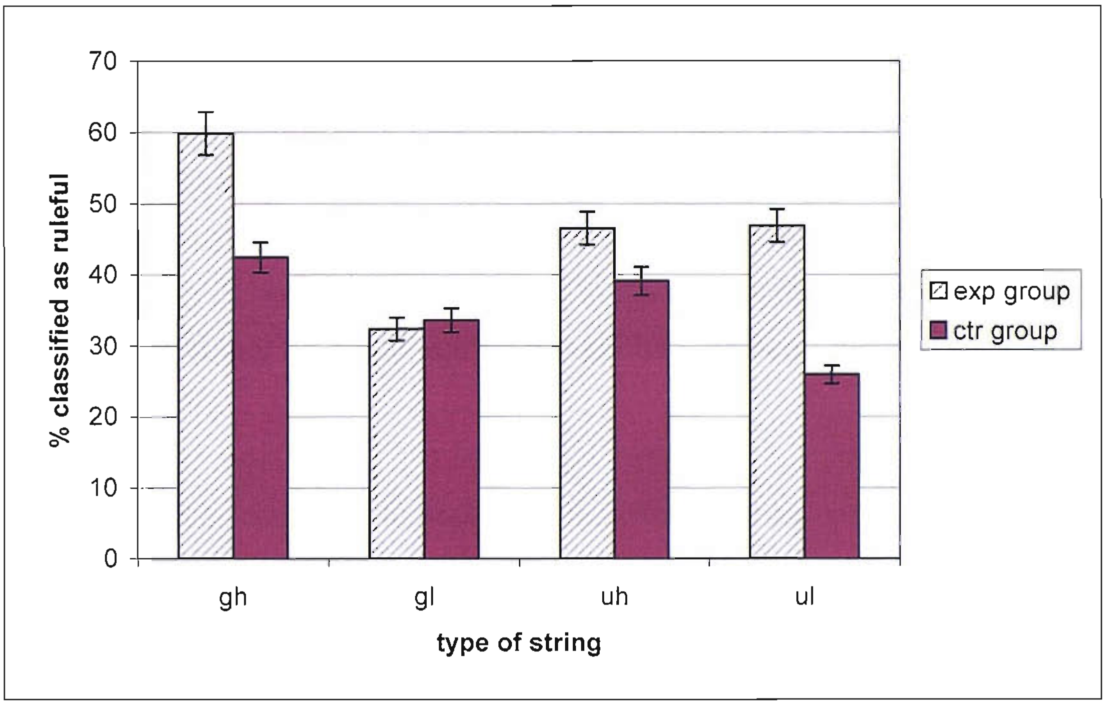


Figure 6.1: Mean percentages of test strings classified as ruleful ("Yes, it follows the rules")

As in the experiment described in Chapter 4, the data was analysed in two ways, according to the rulefulness and according to similarity, see section 4.2.2. In order to see whether Ss were classifying test strings correctly, i.e. according to their rulefulness, or whether they were instead basing their classifications on the similarity of the test strings to the training strings. A repeated measures ANOVA with group as between-Ss variable and rulefulness and similarity as within-Ss variables revealed a significant effect of rulefulness, F = 10.551, p<0.01. This suggests that Ss were classifying test strings according to their rulefulness.

There was no significant effect of group, which indicates that Ss in the experimental and control groups were not performing differently. However, this may also reflect the fact that more strings were classified as unruleful ("No, it doesn't follow the rules") than ruleful; thus, more unruleful strings would have been correctly classified than ruleful strings. This interpretation also corresponds to the mean percentage of correct classifications (see Table 6.1), which do not show a rulefulness effect.

The experimental group selected “Yes, it follows the rules” on an average of 47% of trials, and “No, it doesn't follow the rules” on an average of 53% of the trials, while the control group selected “Yes” on 38% of trials, and “No” on 62% of the trials. There seems to have been a tendency to call more strings unruleful (“No”) than ruleful (“Yes”), although the difference is not significant. This tendency is especially pronounced in the control group. This is probably due to the fact that Ss in the control group have not seen ruleful strings until this point.

Signal detection analyses indicate that the experimental group has a d’=-0.20 (where FA=0.5796 and H=0.50) and a response bias of -0.20. The control had a d’=-0.02 and response bias of -0.09. This also indicates that neither the experimental nor the control group had any sensitivity to the rule, and the control group were much less willing to call strings ruleful (i.e. answer Yes) than the experimental group.

6.2.2.1 The guessing criterion

The mean percentage correct when Ss thought they were guessing was 59% for the experimental group and 34% for the control group. The mean percentage of similar strings that Ss thought were ruleful when they thought they were guessing was 61.02% for the experimental group and 48% for the control group. These data are also shown in Table 6.2.2.1.1.

	Rulefulness	Similarity
Experimental group	58.62	61.02
Control group	34.1	47.63

Table 6.3: Mean percentages of classifications when Ss thought they were guessing

A chi square test conducted on the percentages showed that the means were significantly different from chance, $\chi^2 = 9.0845$ (critical value 3.84). This indicates that Ss in the experimental group were performing significantly better than the control group and also knew more than they thought they knew; i.e. it suggests that Ss were classifying the test strings implicitly (if "implicitly" means being right without being sure you are right).

6.2.2.2 The zero correlation criterion

The confidence ratings were divided into six groups: the first when they thought they were guessing, the second contained the ratings between 51 and 60, the third all ratings between 61 and 70, and so on, as in the experiment described in Chapter 4. Figure 6.2. shows the mean percentages of correct classifications for each level of confidence.

Accuracy and confidence were not correlated. A repeated measures ANOVA with group as between-Ss variable and the six levels of confidence as within-Ss variables revealed a significant main effect of confidence, $F = 2.915$, $p<0.05$, and a significant interaction between confidence and group, $F = 2.922$, $p<0.05$. There was also a significant linear trend of confidence, $F = 5.907$, $p<0.05$, and a significant linear interaction between confidence and group, $F = 8.152$, $p<0.05$.

These findings indicate that Ss were more confident when they were responding correctly. According to the zero correlation criterion, this result suggests that Ss were to some extent conscious of the knowledge they possessed, i.e. they were not responding implicitly.

The results further show that the experimental and control group were responding differently. Figure 6.2 suggests that it was the control group that was more confident on correct answers and less confident on incorrect answers. The experimental group showed virtually no correspondence between correct responses and confidence level. The linear trend can also be seen in Figure 6.2, in the experimental group as a horizontal line around 55%, and in the control group as a diagonal line (indicating the correspondence between confidence level and performance) from low confidence/low correctness to high confidence/high correctness.

Ss in this experiment were not asked whether they could describe (verbalise) the rule, which would be overt proof of explicitness. The zero correlation criterion solely shows that subjects felt they were doing it right, and thus felt more confident that they were getting it right. It cannot tell us whether Ss knew *how* they were getting it right.

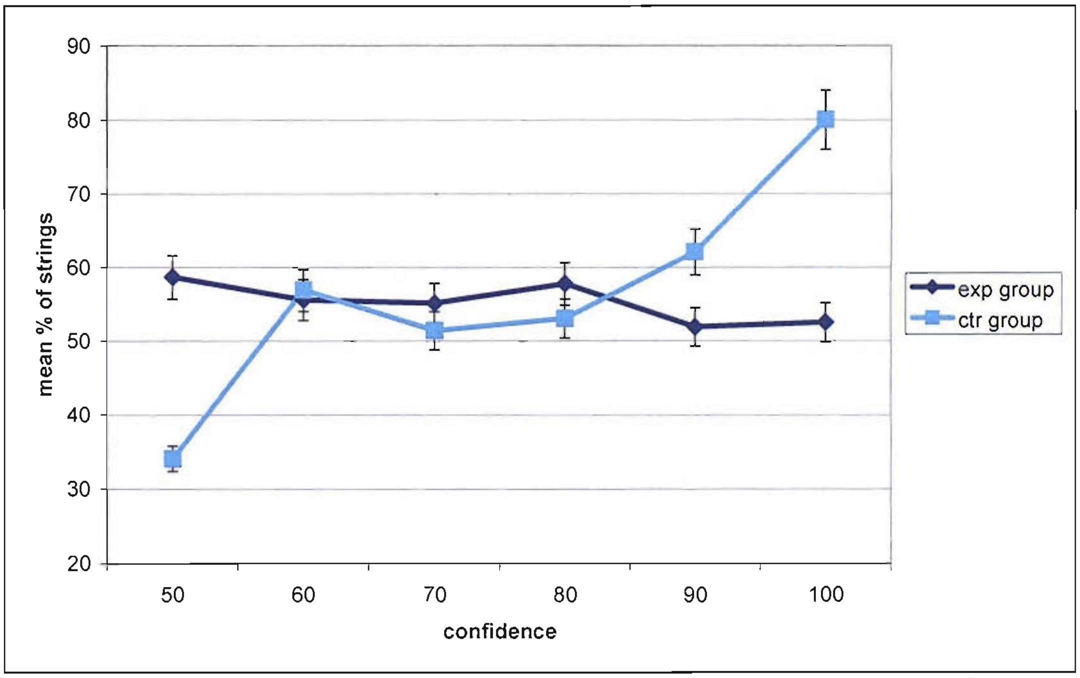


Figure 6.2: Mean percentage of correct classifications for each level of confidence

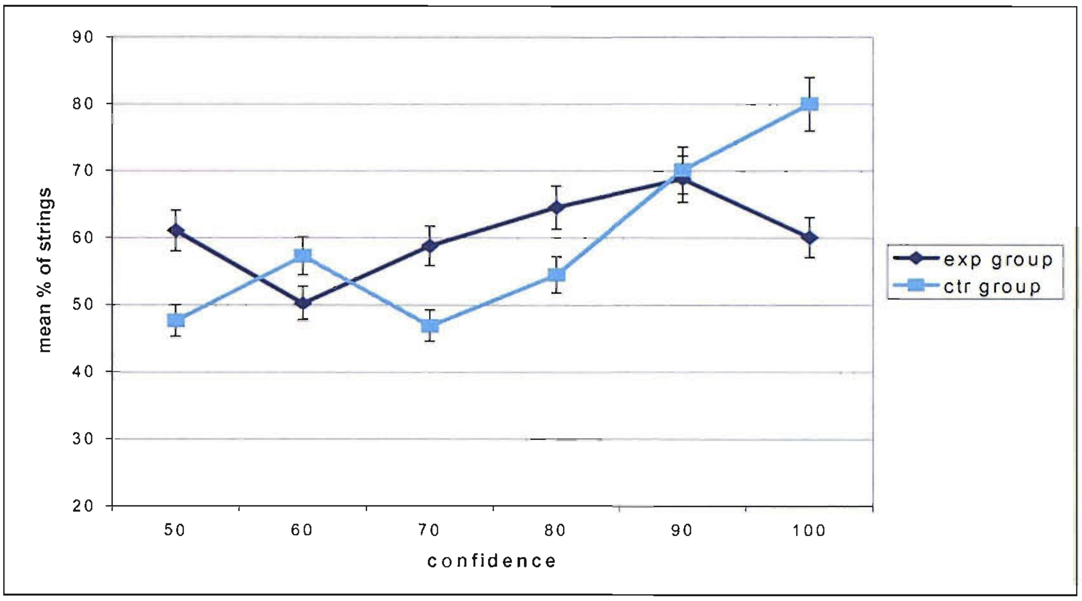


Figure 6.3: Mean percentage of classifications when data were analysed in terms of similarity for each confidence level

Figure 6.3 shows the mean percentages of responses when the data were analysed in terms of similarity for each level of confidence. Confidence and percentage of strings called similarity were significantly correlated in the control group, r = 0.112, p<0.05, but not in the experimental group.

A repeated measures ANOVA with group as between-Ss variable and the six levels of confidence when the data were analysed in terms of similarity as within-Ss variables revealed a significant main effect of confidence, F = 17.920, p<0.05. This suggests that when their responses were analysed in terms of similarity, they were more confident about the "correct" similarity-based responses than about the incorrect similarity-based ones. Figure 6.3 suggests that here, too, the control group was more confident on the similarity-based responses than the experimental group. However, there was no significant effect of the interaction between group and confidence, which indicates that the two groups were not responding significantly differently on the confidence measure.

6.3 Conclusions

The inclusion of unruleful instances in this experiment interfered with learning. Ss were classifying test strings at chance. If an instance memory interpretation is applied, this finding makes sense. With the inclusion of unruleful instances, the memory load was merely increased, and performance understandably worsened.

There was a tendency to call more strings unruleful (i.e. "No, it doesn't follow the rules") than ruleful, especially in the control group. Ss in the control group seemed to be more confident on correct responses than on incorrect responses. However, when experimental Ss thought they were guessing, they were performing at a level above chance, whereas control Ss were performing at a level below chance. Since control Ss received unruleful instances in the training phase but were told that the ones shown on a green background were ruleful instances, they could have been remembering specific green instances (or chunks) from the training phase, incorrectly thinking that these were ruleful instances and thus categorising unruleful instances as ruleful. (Instead of random feedback, control Ss were receiving incorrect feedback, which contaminated their results. This control group hence cannot be treated as a reliable control.) The main conclusion of this experiment, however, is that including unruleful instances in the training phase did not help learning; it hindered it. If Ss were learning rules, the presence of unruleful instances in the training phase should aid learning.

Cheesman & Merikle (see Chapter 4) suggested that knowledge could be considered unconscious if it was below the threshold at which Ss are aware they have some knowledge (the "subjective threshold"). They suggested the zero correlation criterion as a measure which would indicate metaknowledge, which would be above the subjective threshold. However, Ss could be confident about their answers without really knowing *why* they are confident, which could in turn be seen as implicit knowledge. Having explicit knowledge would require Ss to actually *know* the rule, i.e. know *how* they are classifying the test strings and be able to verbalise how they are doing it. Whether they are confident about their answers is relatively irrelevant to being able to verbalise the rule(s). In order to judge whether Ss have explicit knowledge of the rules, Ss in the subsequent experiments in this thesis were simply asked whether they knew the rule(s) or any features of the rules and what these rule(s) were, and to explicitly write down any strategies they were using no matter how trivial they thought they might be.

7 Variables that need to be controlled and tested

The methodologies and designs of the studies that have been reviewed differ considerably in several respects because of the different questions the experimenters were addressing. These differences make it difficult to make any conclusive comparisons among the studies. In order to say something meaningful about the Ss' learning performance, several critical variables need to be controlled:

- The grammar used (see section 7.1);
- The degree of difference among the strings (see section 7.2)
- The number of trials and presentation style (see section 7.3);
- The chance baseline (see section 7.4).

Several further variables differ across studies, and likewise need to be systematically tested and manipulated, as they may affect learning in different ways.

- The task instructions (see section 7.5).
- The nature of the training task and the feedback received (see section 7.6);
- The percentages of positive vs. negative instances (see section 7.7)

7.1 The variables that need to be controlled

7.1.1 Grammar

A major difference among the studies described is the artificial grammars (AGs) used. Whittlesea & Dorken created two finite-state grammars (FSGs) consisting of vowels as well as consonants, since their strings had to be pronounceable. The grammars in all other studies consisted only of consonants.

Brooks created strings from two FSGs similar to the ones Reber used in 1969[15]. Dienes et al (1991) used the same grammar as Dulany et al (1984), Perruchet & Pacteau (1990) and Reber & Allen (1978). Dienes et al (1995) used the two grammars created by Reber (1969). These grammars all consist of different letters of the alphabet and have different numbers of rules. They accordingly differ in difficulty, making some of the AGs used more difficult to learn than others. In order to assess the degree of difficulty it is necessary to know how many trials are needed for 100% correct performance. In this thesis, the number of trials needed for 100% correct performance on a particular set of rules is referred to as the learnability of the rules.

7.1.2 Degree of difference between strings

Three of the four studies reviewed used two different AGs. A ruleful (grammatical/positive) instance of one grammar is then also an unruleful (ungrammatical/negative) instance of a second grammar. The only study in which only one grammar was used was Dienes et al (1991). The use of one versus two grammars changes the degree of difference between the positive and negative strings. In the Dienes et al (1991) study, the negative strings were created by substituting a ruleful element with an unruleful element, so the negative strings differed from the positive strings in only one letter. In the studies using two grammars, the negative strings were created from an entirely different set of rules, which made the positive and negative strings more different from one another.

[15] *Both of the grammars that Brooks used are the same as the ones Reber (1969) used, except that Brooks's are missing the final node. Brooks deleted this final node because he did not want all the instances from each grammar ending in the same letter, as was the case in Reber's grammars. One side-effect of this deletion is that the letter X is associated only with Grammar A. Brooks mentioned, however, that none of his Ss reported having noticed this, even when, in an experiment not reported here, they were asked for their reasons for categorisation after every test trial.*

In addition, the degree of difference among positive strings had not been analysed; nor had the degree of difference among negative strings. In some studies negative strings were created from random letters; in other studies negative strings were created by substituting one or two correct letters with incorrect letters. Random negative strings are more different from the positive strings and more different among themselves than negative strings that differ from positive strings in only one or two letters. The degree of difference among all strings (between positive and negative, between positive and positive, and between negative and negative) accordingly needs to be equated.

7.1.3 Number of trials and presentation method

The number of trials in the training phase differs from study to study. Moreover, not only does the number of times and the length of time Ss see a string vary, but also the number of strings Ss see in total, and the way they are presented (e.g., if they are presented simultaneously or one after the other, and if all positive strings are shown first and then all negative strings, or presentation is random).

In Dienes et al's (1995) and Whittlesea and Dorken's experiments the strings from one grammar were presented first and then the strings from the other grammar were presented, whereas in Brooks and Dienes et al's (1991) experiments strings from both grammars were mixed and shown randomly.

Ss in Brooks' study saw 15 strings from each grammar; Whittlesea & Dorken's Ss saw 40 strings from each grammar; Dienes et al's (1995) Ss saw 32 strings from each grammar; and Dienes et al's (1991) Ss saw 20 positive and 20 negative strings. The strings were shown for varying amounts of time across studies: in some studies for a fixed amount of time, in others until Ss reached a specified performance level.

These variations in presentation time, style, and number may well have affected performance in the various studies. The number of trials and the length of exposure should be systematically related to the learnability of the AG, i.e. the number of trials it takes to learn the grammar to 100%.

7.1.4 Chance baseline

Many of the researchers (e.g. Brooks, Dienes et al, 1991) did not establish a chance baseline for their experiments. Brooks, for example, assumed chance to be 33% in his experiment, because the stimuli were to be sorted into three different categories. However, he also gave the task to a group of control Ss who were asked to sort the stimuli into three piles with no prior training; these Ss were sorting at a level *above* 33% (between 37.1 and 51.1%). Accordingly, the mean performance level of this control group should be treated as chance. There are many ways in which stimuli can be sorted into categories, e.g. based on their structural similarities, their colour etc. Chance performance does not necessarily always correspond to the mathematical chance level. The chance baseline needs to be explicitly ascertained using an untrained control group.

7.2 The variables that need to be tested

7.2.1 Instructions

One variable that must be systematically varied and tested is the instructions Ss receive, i.e. whether they receive instructions that (1) *forewarn* them about the existence of rules generating the stimuli, (2) *forearm* them with the rules from the beginning, or (3) *rule-blind* instructions, in which Ss are not even told that there are rules to be learnt.

Several authors have thoroughly studied the effect of different types of instructions. In this thesis, instructions in which Ss are told about the existence of rules in advance will be termed '*forewarning instructions*', instructions in which Ss are not told about the existence of rules will be termed '*rule-blind instructions'* and instructions in which Ss are told the actual rules which govern the stimuli will be termed *'forearming instructions'*.

7.2.1.1 'Forewarning' vs. 'Rule-blind' Instructions

Mathews et al (1989, Exp. 3) used a FSG and trained their Ss with forewarning and rule-blind instructions. Their Ss who received *rule-blind instructions* were told that on each trial they would see a string of letters which they should try to memorise. Then five choices would appear. Their task was to select the string that was identical to the one they had previously seen. After each response they were informed which was the correct choice and the next trial began.

Their Ss with *forewarning instructions* were told that each string was a flawed example of a string generated by a set of rules. Each string would have from one to four letters that were incorrect. Their task was to mark up to four of the letters as incorrect and to try to figure out the underlying rules. After each trial the incorrect letters were indicated and the correct string was displayed.

After completing the training phase (with either rule-blind or forewarning instructions) all Ss were told that the letter strings they had seen were generated by a complex set of rules. They were told that some of the strings they were about to see in the following test phase were generated by the same set of rules. Their task on each test trial was to pick out – from a choice of five strings – the string that was generated by the same rules to which they had been exposed before. During testing, they were not told whether they had responded correctly. There were 100 multiple-choice trials in this test phase, divided into blocks of 10[16]. After each block of 10 trials

[16] *To measure the generalisability of knowledge acquired during the test phase, the letter set was changed from trial 51 onwards, and from trial 71 onwards, Ss were*

Ss were requested to pause and verbalise instructions on how to perform the task to an "unseen partner" (i.e. so that the unseen partner would be able to perform the task "just [as] you did")[17].

Ss in both groups performed better than chance, with no difference between the groups. This indicates that neither being forewarned in advance that there is an underlying rule (or rules) nor explicitly generating and testing the rules of the grammar (as in the rule-blind condition) necessarily helps Ss learn the rule(s).

Most Ss' verbal instructions for their "unseen partners" consisted of letter patterns to select. Ss in both groups failed to learn the rules and were just remembering fragments from the training strings. The verbal instructions were analysed in terms of the specific trigrams (fragments of three adjacent letters) they told their partners to select. The trigrams that were noticed most frequently, the 'salient trigrams', were the initial (first three letters), terminal (last three letters), and repetition trigrams (e.g. SSS). The mean proportion of trigrams mentioned for the first five blocks (when the letter set was the same) was similar for both groups (0.25 for salient trigrams, and 0.05 for nonsalient trigrams in the rule-blind group, and 0.19 and 0.03 respectively in the forewarned group). This indicates that the verbalisable knowledge of the grammar acquired in the two different training conditions was quite similar, and again, that initial knowledge about the existence of rules does not help performance.

In a further experiment, Mathews et al (1989, Exp. 4) devised a different type of AG based on simple biconditional rules (see Section 3.2.4 for detailed description of this grammar). This biconditional grammar (BCG)

receiving feedback about their choices. These manipulations are not relevant to the issue here, so they (and the results) are not mentioned.

[17] *These instructions were given to a group of yoked Ss, who attempted to perform the same task without the benefit of any prior experience or feedback. The relative performance of yoked Ss versus their experimental partners then provided a direct measure of the extent to which knowledge of the grammar was communicated verbally to another person.*

reduces the level of resemblance between different strings, i.e. the strings are less similar to one another than strings created by a FSG. The design of the experiment was identical to that of their previous experiment (detailed above) but using the BCG instead of the FSG. The forewarned group performed significantly better than the rule-blind group, which was performing at a level very close to chance. The mean proportion of Ss who could verbalise the rules was 0.05 for the group with rule-blind instructions and 0.2 for the group with forewarning instructions.

Mathews et al interpreted these findings as evidence of two distinct learning processes, one explicit and one implicit. They concluded that implicit learning processes are only capable of identifying common patterns of resemblance among strings. The BCG was designed to have a limited degree of resemblance among strings, so that high levels of performance would require going beyond patterns of familiarity. Thus, rote memorisation strategies are less effective with BCGs than with FSGs.

Shanks and his co-workers (1997, 1999, 2000) conducted some similar experiments. Their conclusions were that different instructions induce separate learning processes. In their FSG experiments there was no difference in performance between forewarned and rule-blind Ss, but in their BCG experiments several forewarned Ss were performing perfectly, while both rule-blind Ss and those forewarned Ss who did not learn the rules (non-learners), were performing at chance levels. Shanks et al suggested that forewarning instructions induce a hypothesis-testing strategy for learning the rules and that the complexity of the rules is an important factor. The rules of typical FSGs are too complex to be learnt by hypothesis testing (e.g. Brooks, 1978, Reber, 1976, Reber et al, 1980 etc) whereas the rules of a BCG are learnable (Mathews et al, 1989, Exp. 4; Shanks et al, 1997, Exp. 4). There are large individual differences in hypothesis-testing ability. Shanks et al suggested that some Ss may need preliminary training in hypothesis testing before they can benefit from forewarning instructions.

The contingencies generated by a rule might be learnt with rule-blind instructions, but only very slowly, over a large number of trials.

In sum, performance in implicit learning experiments depends largely on the complexity, and consequently on the learnability of the grammar (which needs to be tested explicitly).

7.2.1.2 'Forearming' instructions

If Ss are forearmed with the rule(s) in advance, one would expect their performance to be 100% accurate (or very close to 100%). Their response times are likely to be slow at the beginning, when they are applying the unfamiliar rule(s) in a conscious, controlled way, but should increase with time (i.e. trials), as the rule execution becomes automatised.

7.2.2 Task and Feedback

Whether Ss are (1) merely passively exposed to the stimuli during the training phase or (2) required to give an active response followed by corrective feedback[18], is also likely to affect test performance. Active responding with feedback occurs when Ss see a stimulus, respond by saying (for example) "Unruleful" and are told "Yes, that was correct". Passive exposure is when Ss are merely shown the strings (possibly repeating or memorising them), but are not asked to judge whether or not they are ruleful. When Ss are responding actively, they are likely to be paying more attention, hypothesis-testing (whether implicitly or explicitly), and learning (whether implicitly or explicitly) more about the correctness/incorrectness of their hypotheses than when they are merely exposed to the stimuli passively.

However, passive exposure can be combined with a cue (discriminative stimulus). For example, in Dienes et al's (1991) study the positive stimuli were presented in black and the negative stimuli were presented in red.

[18] *Reber defines feedback in learning as "any information about the correctness or appropriateness of a response" (Reber, Penguin Dictionary of Psychology, 1985)*

Here the colour was a (passive) cue. This extra bit of information may make the differentiation a bit more salient, especially as colours are quite obvious cues. The cue can also be more "active", for example in the Whittlesea & Dorken study, Ss had to *spell* stimuli from one grammar and *say* stimuli from the other grammar. Ss did not know that the spelled stimuli were from one grammar and the spoken stimuli from a different grammar, but the cue made the two stimuli differ from one another even under these passive conditions. One would predict that the more salient the difference between the positive and negative strings, the more likely that Ss will be able to detect, respond to, and even verbalise the difference.

The most salient condition for differentiating positive and negative strings would be when Ss are explicitly told what the difference is in advance, i.e. when they are told the rule(s), as in the forearmed condition. Then no hypothesis-testing is needed, only the application of the known rule(s). (It is unclear whether such a task is really usefully described as a rule-learning task rather than just a rule-application and automatisation task.) Passive presentation would be least salient. A differentiating passive cue, such as positive and negative strings coded in a different colour would increase the salience of the difference, and being informed in advance of the existence of an underlying rule (forewarned condition) would make it more salient still. Trial-and-error responding on each trial with corrective feedback would make it most salient, short of actually giving hints about the hypothesis itself. All these factors need to be explicitly tested.

7.2.3 Positive and negative instantiation

The proportion of positive and negative instances in the training phase needs to be systematically varied and tested. Optimal learning conditions (assuming the population frequency of positive instances to be equal) would be those in which Ss were trained with 50% positive and 50% negative instances. Changing the proportion of positive and negative instances is likely to affect learning. If Ss receive more of one type of instance than the other, one would expect learning and performance to be less successful.

8 Experimental Series

In general, artificial grammar learning (AGL) research has been concerned with establishing the existence of implicit learning, i.e. learning without conscious awareness. However, no investigator has yet taken the critical antecedent step of testing whether the artificial grammars (AGs) are *learnable at all*, and if learnable, how many trials are needed until near error-free performance is reached. This is an important issue, because if the AG is not learnable in the first place, then neither explicit nor implicit learning can occur. Or if the AG is so complex that it would take Ss an unreasonable number of trials to learn it – as seems to be the case with most of the AGs used in previous experiments –, it seems unreasonable to assume that either implicit or explicit learning will occur in the short time available for an experiment (usually an hour-long session). In the case of an AG that is too complex, Ss have no prospect of ever learning the rules, implicitly or explicitly , and hence have no other option than to try to memorise fragments of the letter strings and to base their subsequent categorisations on this rote memorisation of fragments. There is no rule learning, only memorisation.

We have conducted a series of experiments that extrapolated backwards from the classical AGL tasks (in which many of the AGs had proved unlearnable or nearly unlearnable) to very simple category learning tasks. In the classical AGL experiments, Ss were not told about the existence of rules ("rule-blind"), presented with only positive instances of the grammar, and given no feedback on the correctness of their responses. In the present series of experiments, Ss were presented with optimal learning conditions in order to test whether or not a rule was learnable. Ss were told that their task was to try to learn the rules. They were presented with both positive and negative instances, and on each trial Ss received corrective feedback on their answers.

8.1 Method

8.1.1 Stimuli

All stimuli consisted of strings of eight pronounceable syllables separated by a hyphen in the middle. All syllables took the form *xa* where *x* may be any of the 17 uppercase consonants B, D, G, H, J, K, L, M, N, P, R, S, T, V, W, Y, Z, and *a* was the lowercase vowel a in all strings (to make the string pronounceable). The consonants C, F, Q, and X were not used, because C and F could produce impolite (English) combinations and Q and X were considered unpronounceable combined with the vowel a. An example of a string is LaGaTaRa-MaNaLaVa.

8.1.1.1 Positive strings

All positive strings consisted of six random syllables and two syllables relevant to the rule. The two relevant syllables appeared in all positions across the whole string in all experiments (dependent also on the particular rule used).

40 positive training strings and 20 positive test strings were created. Ten of the test strings were similar to the training strings, i.e. they differed from the training strings in only one syllable. The other ten test strings were dissimilar to the training strings, i.e. the syllables were randomly chosen and created in the same way as the training strings.

8.1.1.2 Negative strings

All negative strings consisted of eight random syllables, that do not conform to any rules.
40 negative training strings and 20 negative test strings were created. Ten of the test strings were similar to the training strings and differed in only one syllable. The other ten test strings were dissimilar and were created randomly to not conform to the rules.

8.1.2 Subjects

These experiments were made available on the World Wide Web, so the potential subject database was much larger than when recruiting Ss by ‘traditional’ means, i.e. conducting the experiment in an experimental laboratory. In addition, "[w]eb experiments provide the researcher with easy access to a much wider and geographically diverse participant population" (Reips, 2000).

Ss were recruited via posters, emails, international mailing lists, Web experimental lists, and word of mouth. The need to visit an experimental lab was eliminated, since Ss could access and complete the experiment from any computer connected to the Internet. However, a disadvantage of web-based experiments is that there is no control over Ss dropping out of the experiment, and the dropout rate can thus be high. Reips (2000) provides a list of recommendations for Web experiments, of which several were used in the current experimental series. Ss were offered the chance of winning a prize for full participation (i.e. only if they completed the whole experiment); Ss were given feedback on their performance and were informed about their current position in the time sequence of the experiment. Ss were also told that participation was a serious matter, and were given informative details on the nature and trustworthiness (e.g. name of institution, contact information, scientific purpose etc) of the experiment. These factors were all intended to reduce drop-out during the experiment.

In an early study by Reips (1995) it was shown that duplicate participation in web-based experiments (i.e. a subject taking part more than once) was uncommon, and he considers it "safe to assume that 'cheating behaviour' is rare". He speculates that this may also be partly due to the fact that his experiment took 45 minutes to complete, and that there may be more duplicate cases in experiments of shorter duration. In the present experimental series, most of the experiments took around 30 minutes or more to complete, so the risk of multiple participation also seemed quite low. Nevertheless, the potential danger of duplicate participation was controlled for by checking dates and time of data submission, the personal identification data. In addition, the "Back" button on the Web browser was disabled, so participants could not go back and change their data.

8.1.3 Apparatus

This experiment was conducted using a program written in Java by Michael O. Jewell and the author. The experiment was made available on the Internet and was accessible at URL http://www.ecs.soton.ac.uk/~mtj00r/webpage/experiment.htm. The experiment consisted of several pages with the instructions for the experiment (actual number of pages varying between one and three depending on the particular experiment), a form, in which Ss entered their personal details, the Java applet, and finally the debriefing page with the possibility of obtaining a personal ID code for entry in the prize draw, of linking to other Web experiment lists, of accessing the experimenter's homepage, and of contacting the experimenter by email. The Java applet compiled each S's descriptive data and all responses in a logfile which was then saved on the server with a unique name comprised of date and time of submission. If the subject decided to obtain a personal ID code, the code was sent via email to the experimenter separately from the logfile.

Various experimental parameters (e.g. number of learning and test trials, difficulty of rule, feedback on or off, etc) could be manipulated as desired by the experimenter.

A disadvantage encountered relatively often was that Ss needed to have at least Java version 1.4 installed on their computer to take part in the experiment. Although Ss were informed in advance whether or not they needed to download Java, several Ss would not (time and/or bandwidth constraints) or could not (security and/or permissions constraints) install Java, and thus could not take part in the experiment.

In order to ascertain whether a S conducting an internet experiment would behave differently to a S conducting the experiment in the same room as the researcher one of the pilot studies (see section 8.2) was also conducted in the traditional way of inviting participants to the lab. The results of the web-based experiment and the traditional experiment were compared and found to be not significantly different. This suggests that web-based experiments are suitable for collecting the type of data described in this thesis.

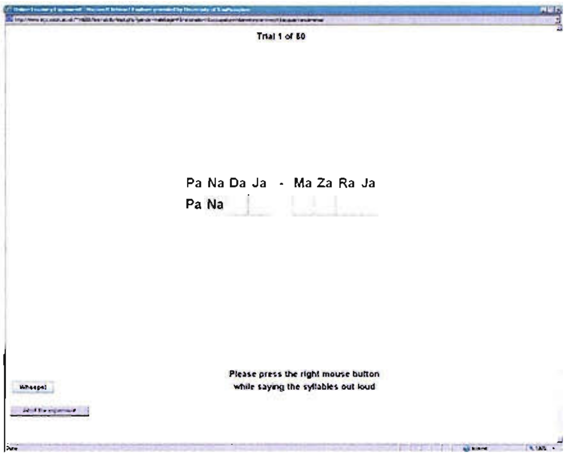
8.1.4 General Procedure

Instructions and Descriptive Data. To participate in this series of experiments, Ss had to have Java 1.4 (or later version) installed on their computer. Ss were notified on the Instructions page whether or not Java was installed on their machine. If Java was present, a green box was visible; if Java was not present, a grey box was visible. Ss who did not have Java installed on their machine, were asked to follow a link www.java.sun.com, via which they could download Java for free. After reading the instructions, Ss were required to fill in a form with their gender, age, occupation, environment, and where they had heard about the experiment, before proceeding to the actual experiment. The environment variable consisted of a dropdown menu with the choices ‘At Home’, ‘At Work’, ‘Internet Café’, ‘Public area’, or ‘Other’. This environment variable was used to investigate if there were any differences in performance depending on the environment the subject was in while completing the experiment. Ss were also requested to indicate where they had heard about the experiment, from a dropdown menu with the choices ‘Email’, ‘Poster’, ‘Mailing list’, ‘Internet Search’, ‘List of online experiments’, ‘Friend’, or ‘Other’. This enabled the experimenter to evaluate from where Ss had heard about the experiment, and which was the most effective method for recruiting Ss.

Practice Trials. Once Ss filled in the form, they were first presented with five practice trials (consistent with the rule(s) of the particular experiment), in order to familiarise them with the stimuli and inputs, and to make sure they understood their task. After the practice trials, Ss could read the instructions again if they wished, or could proceed directly to the experimental trials. No data were recorded in the practice trials.

Learning Trials. Learning trials consisted of either positive and negative stimuli, positive stimuli only, or negative stimuli only, depending on the experiment. All Ss were presented with the same set of learning trials in a different random order.

Stimulus and Repetition. On each trial a stimulus consisting of eight syllables separated by a central hyphen, appeared in the centre of the screen. After 1000 msec (adjustable as desired by the experimenter, but remaining at a constant 1000 msec for this series of experiments), a set of eight boxes appeared beneath the stimulus, one box corresponding to each syllable and appearing directly beneath the syllable. The subject was requested to click on the right mouse button while saying each syllable out loud. With every click of the right mouse button, the syllable, which the subject was requested to say out loud, appeared in the appropriate box. When all 8 syllables had been repeated, the repetition boxes disappeared while the stimulus remained in the middle of the screen. See Figure 8.1 for a screenshot of a learning trial, where two of the syllables have been repeated and appeared in the appropriate box.



Trial 1 of 80

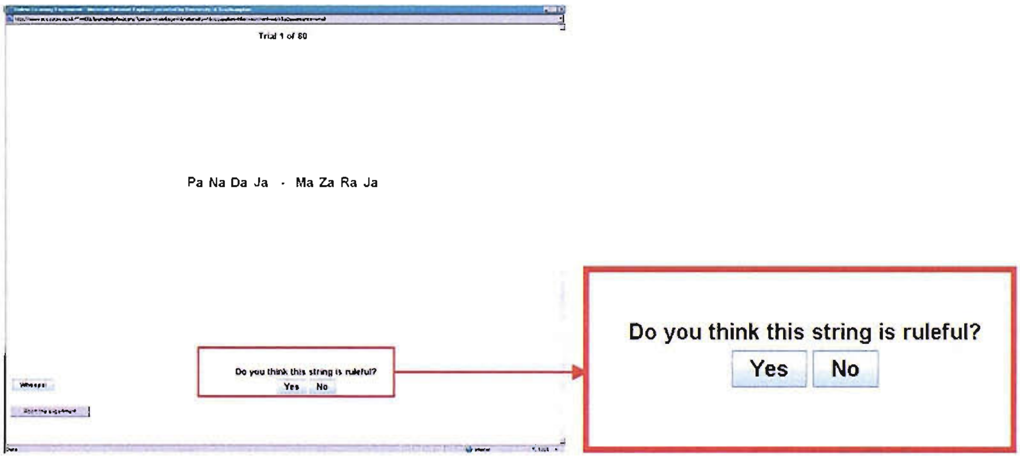
Pa Na Da Ja - Ma Za Ra Ja

Pa Na

Please press the right mouse button while saying the syllables out loud

Figure 8.1 Screenshot of a learning trial

Response. Depending on the experiment, Ss were asked whether or not they thought the stimulus followed the rule(s). Two buttons inscribed with 'Yes' and 'No' appeared at the bottom of the screen. Ss were requested to click on 'Yes' if they thought the stimulus followed the rule(s), or on 'No' if they thought the stimulus did not follow the rule(s).



Trial 1 of 80

Pa Na Da Ja - Ma Za Ra Ja

Do you think this string is ruleful?

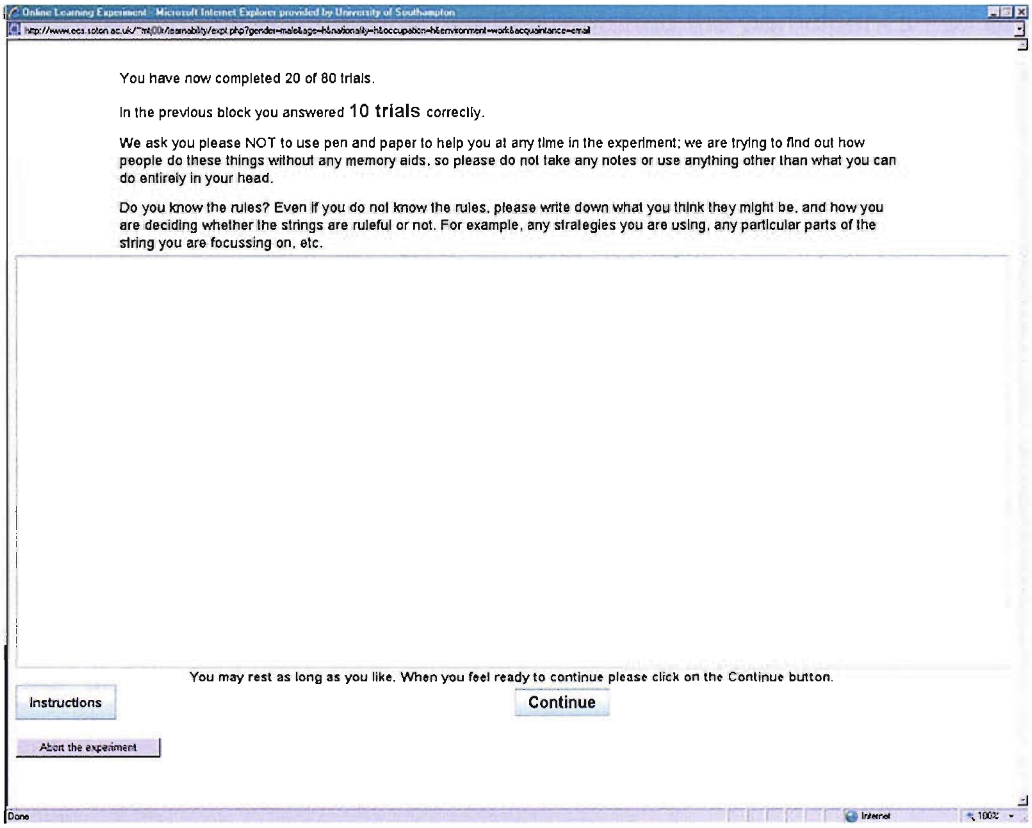
Figure 8.2 Screenshot of the response Ss were required to make

Feedback. When Ss clicked on either of the response buttons, depending on the experiment, the program informed Ss whether their response was correct or not. If the answer was correct, the feedback was shown in the centre of a cyan-blue screen. If the stimulus was ruleful, Ss were told "Correct, *xa xa xa xa – xa xa xa xa* is ruleful", if the stimulus was unruleful, Ss were told "Correct, *xa xa xa xa – xa xa xa xa* is unruleful". If the S's answer was incorrect, the feedback was shown in the centre of a magenta-purple screen. If the stimulus was ruleful (but S thought it was unruleful), Ss were shown "Incorrect, *xa xa xa xa – xa xa xa xa* is ruleful", and if the stimulus was unruleful (but S considered it ruleful), they were shown "Incorrect, *xa xa xa xa – xa xa xa xa* is unruleful". At the bottom of the screen, there was a Continue button which Ss had to click when they were ready to continue to the next trial.

Test Trials. The test trials consisted of new positive and negative stimuli. On each trial, the stimulus appeared in the centre of the screen with the repetition boxes corresponding to each syllable placed directly underneath each syllable. Ss again had to repeat the stimulus out loud and click on the right mouse button for each syllable to appear in the corresponding box. After the repetition, Ss were asked whether or not they thought the stimulus followed the rule(s), and had to click on the 'Yes' or 'No' button. Ss were not given any feedback in the test phase. All Ss were presented with the same set of test trials in a different random order.

Whoops button. At the left hand side of the screen there was a "Whoops" button. Ss had been instructed only to click on the Whoops button, when they had made a mistake (e.g. when they had clicked on 'Yes' button when meaning to signal 'No'), and that the data of that trial would then be cancelled.

Breaks. In the experiments described in Chapters 4 and 6 Ss' confidence ratings indicated that they were more confident when they were classifying those test stimuli that had chunks similar to learning stimuli. This may indicate that Ss might have some awareness of their means of classifying (i.e. according to the similarity to learning strings rather than according to their adherence to the rules). In the subsequent studies conducted for this thesis, Ss were asked at intermittent stages (after every 20 trials) in the experiment whether they knew the rules, and what they thought they were; Ss were also asked to write down any strategies they were using to decide whether or not the stimuli were ruleful, or any other ideas they may have had during the previous 20 trials, no matter how irrelevant they may have seemed. In other words, Ss were asked to report everything they were thinking while participating in the experiment in order to test Ss' explicit knowledge of the rules. It was hoped that this information would give us a better idea of what Ss were doing, and would track whether Ss could verbalise the rules, whether their performance corresponded to their comments, etc. The data gathered from these breaks were later analysed as the verbalisation data. There was a Continue button at the bottom of the screen, which Ss had to click when they were ready to continue the experiment. They were encouraged to rest as long as they liked until they felt ready to continue. Ss also had the possibility of reading the instructions again during the breaks. See Figure 8.3 for a screenshot of a break in an experiment.


Online Learning Experiment - Microsoft Internet Explorer provided by University of Southampton
http://www.ecs.soton.ac.uk/~mj00r/learnability/expt.php?gender=male&age=h&nationality=h&occupation=h&environment=work&acquaintance=email
Instructions
Continue
Abort the experiment
Done
Internet
100%

You have now completed 20 of 80 trials.

In the previous block you answered **10 trials** correctly.

We ask you please NOT to use pen and paper to help you at any time in the experiment; we are trying to find out how people do these things without any memory aids, so please do not take any notes or use anything other than what you can do entirely in your head.

Do you know the rules? Even if you do not know the rules, please write down what you think they might be, and how you are deciding whether the strings are ruleful or not. For example, any strategies you are using, any particular parts of the string you are focussing on, etc.

You may rest as long as you like. When you feel ready to continue please click on the Continue button.

Figure 8.3 Screenshot of a break in an experiment

Debriefing. When Ss had completed the experiment, they were thoroughly debriefed about the nature of the experiment and told what the rule(s) were, and given examples of both positive and negative stimuli. Ss were also encouraged to contact the experimenter if they had encountered any problems or had any questions. There was a link to the experimenter's email and homepage, and also links to other online experiment lists. Very few emails were received from Ss who participated in the experiment. This is probably due to the fact that Ss had already written everything they wanted to say in the space provided them in the breaks and at the end of the experiment (as was also requested in the experiment).

***Compensation**.* If the experiment included the possibility of winning a prize[19], Ss had the chance to obtain a randomly generated ID code. A deadline for participating in the experiment was given and Ss were told to check a certain webpage at a certain pre-set date (e.g. beginning of a certain month[20]). Half of the randomly selected winning ID code was displayed on this webpage. If this half matched the participant's half of the ID code, the participant was asked to send an email to the experimenter with the other half of the winning ID code to claim the prize.

8.2 Learnable Rules

In order to study the effects of the degree of learnability, several pilot experiments were conducted to find an easy, medium and a hard grammar. Easy was arbitrarily defined as "learnable in 15-20 trials", medium as "learnable in 40 trials", and hard as "learnable in 80 trials"[21]. Once learnable rules had been established, experiments could be conducted to test the effects of feedback, response type, instruction type, and type of instances presented.

[19] *Initially the experiments did not include the possibility of winning a prize. However, as it got progressively more difficult to recruit participants, a prize was introduced, which had the desired effect of raising participant numbers.*

[20] *The date was set to one month after the deadline of that particular experiment.*

[21] These "definitions" seemed reasonable considering how long Ss were willing to take part in the experiment

8.2.1 Pilot Studies

The pilot experiments consisted of 80 learning trials and 20 test trials. On each trial a stimulus consisting of eight pronounceable syllables, e.g. LaTaGaVa-RaPaWaBa, was presented. Ss were informed that there were rules governing the letter strings, and that they should try to figure out the rules. Ss were required to repeat the stimulus out loud. Next they had to indicate whether or not they thought the stimulus followed the rules; then they received feedback on their answers. After every 20 trials there was a break during which Ss were required to write down any rule(s) they thought they knew or any strategies they were using, or anything else they were thinking. After the learning phase came the testing phase, in which Ss were tested on new positive and negative strings (not in the training set). They again had to indicate whether or not they thought the string followed the rules, but they no longer received feedback on their answers.

The Ss' learning curves indicated how many trials on average were needed to learn the rule(s). The results of several pilot studies showed that only very simple rules were learnable to 100% accuracy in 80 trials.

8.2.2 Pilot Study Results

8.2.2.1 Easy rule

The easy rule (i.e. "learnable in 15-20 learning trials") was established as: One of the syllables has to be repeated in both halves of the string for the string to be ruleful. If no syllable is repeated in the string, the string is unruleful.

Examples of ruleful strings are:

Ta**Ka**LaGa-VaRaSa**Ka**
Ba**Va**TaYa-Wa**Va**PaLa
DaKaZaHa-WaJa**Da**Ma (the repeated syllables are in bold print)

Examples of unruleful strings are:

KaGaLaPa-VaBaDaRa
DaTaZaLa-NaBaYaWa
HaPaKaJa-RaDaSaGa

The learning curve in Figure 8.4 shows that this rule was learnable to approximately 100% in 15-20 trials. Ss were performing at an overall average of 96% in the test phase (i.e. trials 81-100). Ss were performing at an average of 88% in trials 15-20.

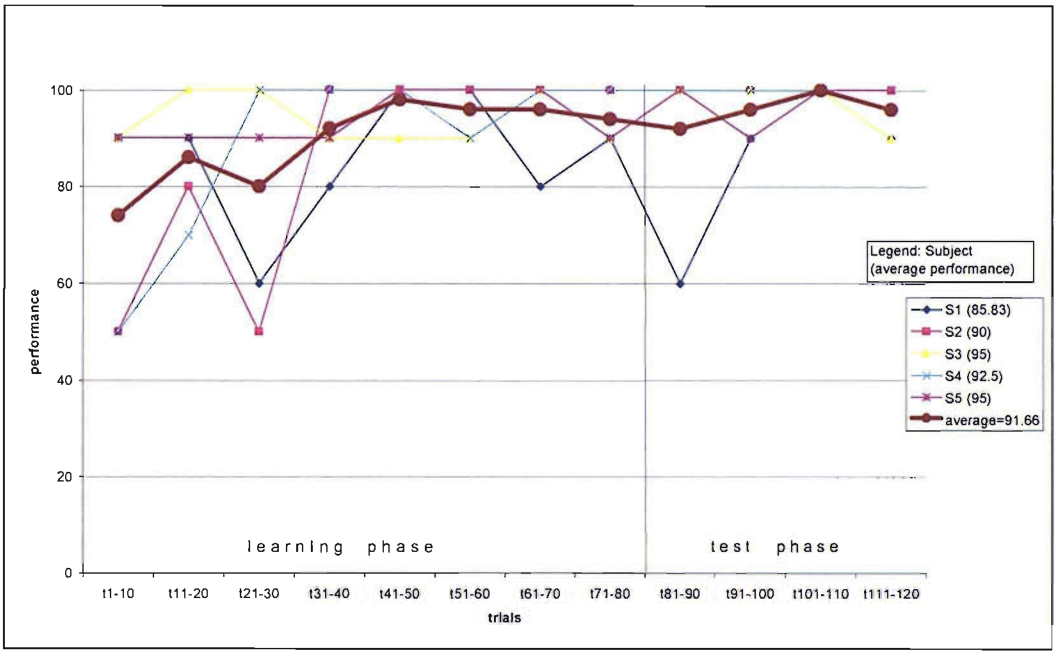


Figure 8.4 Learning curves of easy rule

8.2.2.2 Medium rule

The medium rule was established as: One of the syllables in the first half of the string is repeated in the second half of the string. If this repeated syllable is in the mirror image position in the second half of the string, then the string is ruleful. If it is in any other position, the string is unruleful.

Examples of ruleful strings are:

GaTaZa**La**-**La**DaBaVa
Wa**Da**PaKa-BaLa**Da**Sa
VaYaZaKa-RaJaMa**Va** (the ruleful part of the strings are shown in bold print)

Examples of unruleful strings are

ZaMaNaVa-Da**Za**PaRa
HaVaJa**Sa**-Ba**Sa**PaWa
YaLaKaGa-HaNa**SaSa**

Figure 8.5 shows the learning curves of Ss who learnt the medium rule. Ss were performing at an overall average of 74.6% correct in the test phase. In trials 31-40, Ss were performing at 67.14% correct.

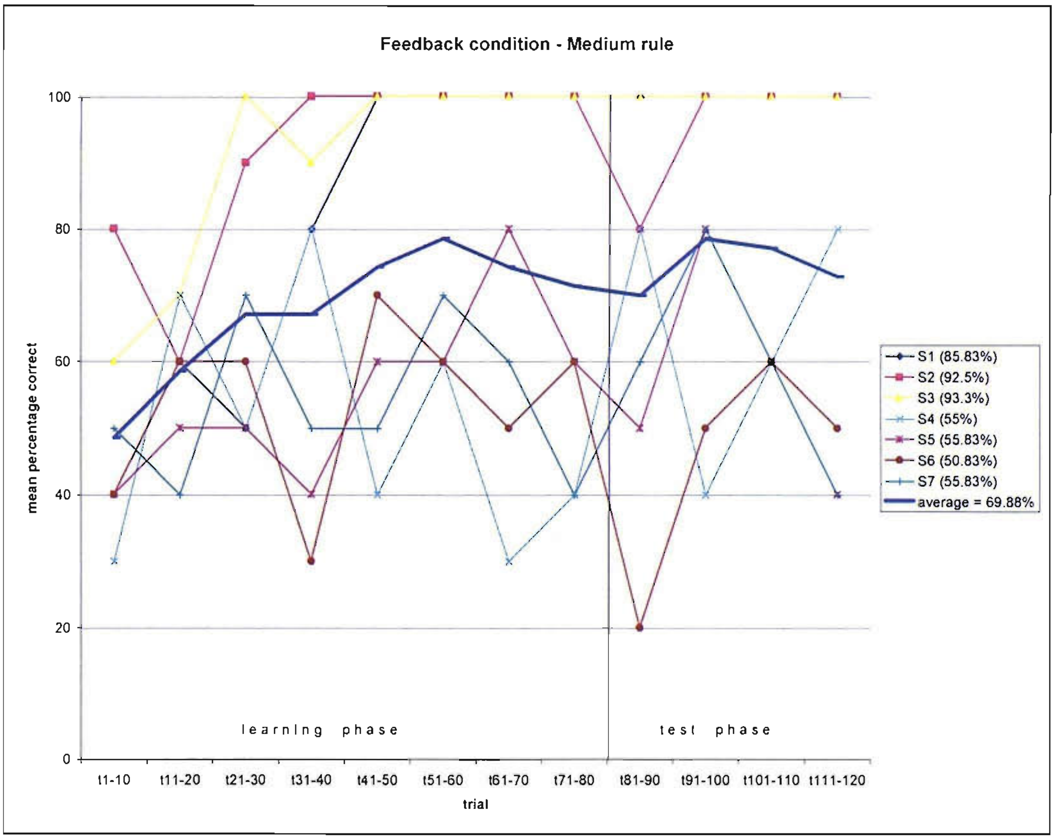


Figure 8.5 Learning curves of medium rule

8.2.2.3 Hard rule

The hard rule was established as: Ruleful strings have two identical syllables, which are separated by one other syllable. Unruleful strings have two identical syllables, which are either adjacent or separated by more than one other syllable.

Examples of ruleful strings are:

HaRa**Ha**Ta-BaKaDaYa
PaGa**La**Na-**La**VaDaSa
MaPaWaYa-Ta**Za**Ga**Za**

Examples of unruleful strings are:

Ka**Ha**GaLa-VaBa**Ha**Pa
RaTaZaNa-**MaMa**YaWa
Ga**Wa**DaHa-**Wa**MaVaPa

Figure 8.6 shows that Ss were performing at an overall average of 64.25% correct in the test phase. Ss were performing at 69% correct in trials 71-80.

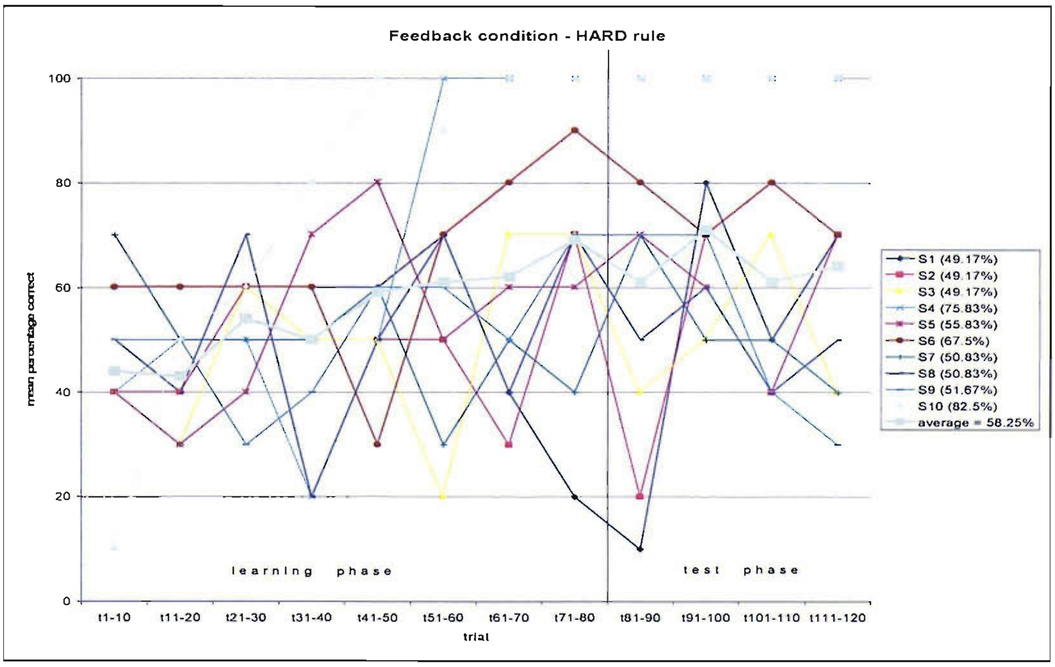


Figure 8.6 Learning curves of hard rule

9 Effect of Feedback on Learning

This experiment investigated the effect of feedback on learning. Reber defines feedback in learning as “any information about the correctness or appropriateness of a response” (Reber, Penguin Dictionary of Psychology, 1985). In this thesis, the term feedback is used in the more general sense of “any information about the correctness or appropriateness of a stimulus or response” and it can be *active*, *passive*, *explicit* or *implicit.* The various uses of the term feedback can be illustrated using the example of the rulefulness of a string. In this thesis, Reber’s definition of feedback corresponds to ‘*active feedback*’ and occurs when Ss make a response to indicate whether or not they think the string they have just seen follows a certain rule (“Yes, it is ruleful” or “No, it is not ruleful”) ; immediately after responding they are informed whether or not their response was correct. *Passive feedback*, on the other hand, occurs when Ss make no response about the rulefulness of the string, hence they cannot be given any subsequent information about the correctness of their response; they are merely informed whether or not the string is (or was) ruleful.

A further distinction is made between explicit and implicit feedback. *Explicit feedback* is feedback about a string’s rulefulness given when Ss are aware that it is feedback about a rule, i.e. when Ss are told explicitly what the feedback means. Active and passive feedback are thus both forms of explicit feedback. *Implicit feedback*, on the other hand, is information about a string’s rulefulness given when Ss are not explicitly informed of what it means. Presenting ruleful strings on a red background and unruleful strings on a green background (without explanation) is an example of implicit feedback in the form of colour, since Ss are not told what the colours mean, but it still differentiates the two kinds of string.

The working hypothesis was that *all* forms of feedback would have a beneficial effect on learning, i.e. that Ss who received feedback in the learning phase would perform better than those who did not. In this experiment, Ss in the feedback group will be given immediate, active feedback after each responses about whether they are correct or not. Ss in the no feedback group will not be given immediate feedback on their responses.

9.1 Method

9.1.1 Design

The design of this experiment is a 3 x 2 design with difficulty (easy, medium, and hard) and feedback (with and without) as independent factors and task performance and verbalisation ability as dependent variables. Since these were web-based experiments, Ss were asked to write down whether they knew any rule(s), what rule(s) they thought they knew, any strategies they were using, and any other comments they had after every 20 trials and at the end of the experiment. These written comments were then analysed as the verbalisation data (see also Section 8.1.4 and Figure 8.3 for a screenshot).

9.1.2 Subjects

56 voluntary Ss took part in this online experiment. There were 32 Ss in the feedback group, and 24 Ss in the no feedback group. In the feedback group, 10 Ss were assigned to the easy condition, 9 Ss to the medium condition, and 13 Ss to the hard condition. In the no feedback group, 7 Ss were assigned to the easy condition, 9 Ss to the medium condition and 8 Ss to the hard condition.

9.1.3 Procedure

Ss were told ("forewarned") that their task was to try to learn the rule. The learning phase consisted of 80 trials for all difficulty conditions (i.e. easy, medium and hard rule). In the learning phase, Ss were shown positive and negative strings. On each trial they were presented with a string, which they had to repeat out loud while clicking the right mouse button. Once they had repeated the string, Ss were asked whether or not they thought the string was ruleful, by clicking on Yes or No. In the *feedback condition*, Ss were informed (on-screen) after every trial whether or not their response was correct, while in the *no-feedback condition*, Ss were not informed. In the test phase, Ss had to indicate whether or not each presented (new) string was ruleful, but they were not given any feedback on their response.

After every 20 trials in both the learning and test phases, Ss in both conditions were told how many trials they had classified correctly in the immediately preceding block of 20 trials. Thus, Ss in the no-feedback condition were receiving some delayed feedback on their overall success-rate across the 20 preceding trials, but no corrective feedback on each actual response. Although this delayed success-rate feedback may influence performance somewhat, in an experiment of this length it is necessary to provide Ss with at least this minimal reinforcement to keep them motivated and to ensure that they complete the experiment (Reips, 2000). Without this intermediate feedback, the chances would be high that Ss in the no-feedback condition would drop out of the experiment without finishing (see also Section 8.1.2).

In these breaks Ss were also asked whether they knew the rules and what they thought they were and what strategies they were using to decide on the rulefulness. They were asked to key in everything they were thinking, no matter how irrelevant they thought it was.

9.2 Easy rule

9.2.1 With Feedback

10 Ss took part in this condition, 8 male, 2 female. The average performance in the test phase was 89.25%. The easy rule had been pre-tested and calibrated to be learnable within about 20 trials (see pilot studies, Chapter 9). Ss were performing at an average of 75% in trials 11-20.

In Figure 9.1. the learning curves of all Ss in the feedback group in the easy condition are shown with the average learning curve in boldface. Most Ss could correctly distinguish ruleful from unruleful strings after about 30 trials, although most were also still making some mistakes in later trials, probably due to fatigue or lack of concentration.

Seven of the ten Ss could describe the rule after the maximum 40 trials. As soon as these Ss learnt the rule, performance improved from chance to near perfect, and they verbalised the rule at first opportunity (i.e. in the next break).

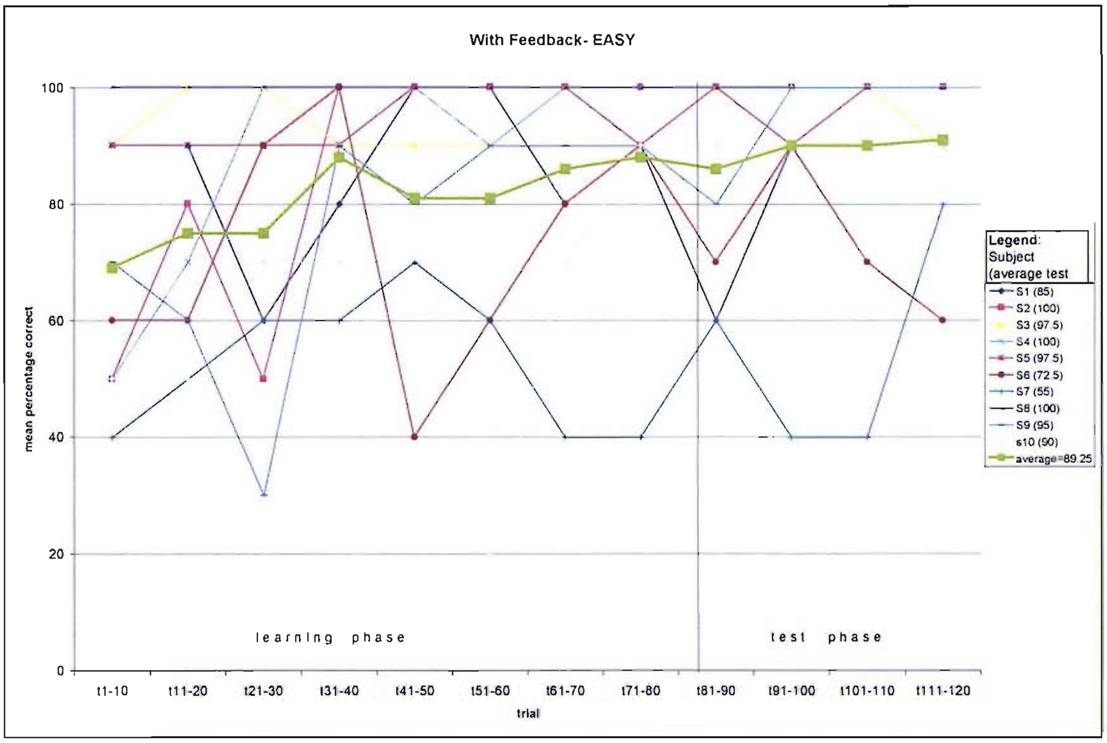


Figure 9.1: Learning curves for Ss in the feedback group in the easy condition

Of the remaining three Ss, one did not write anything down, but was performing at a high level of correctness; a further S only verbalised incorrect rules, and his overall performance in the test phase was only 55%. A further S stated the correct rule after 20 test trials (so after a total of 100 trials – 80 learning trials and 20 of the 40 test trials), but is performing at a very high level (90% correct) from learning trial 60 onwards. This S may have learnt to respond correctly (perhaps also implicitly) at an earlier stage than his ability to verbalise the correct rule. However, it is also possible that this S did not write down the rule until completely certain of it. Ss were asked to write down any rules they were using during the experiment, but were not explicitly asked from which trial onwards they were using these rules from.

9.2.2 With No Feedback

Seven Ss took part in this condition: 4 male, 3 female. Overall performance in the test phase was 53.93% correct.

Figure 9.2 displays the learning curves of Ss in the no-feedback group in the easy condition. The average learning curve is in boldface. One S had an average of 85% in the learning phase, and this S also stated the rule explicitly ("matching instances on each side"), although he did not seem to be entirely confident about the rule, both in performance and verbalisation. All other Ss failed to learn the rule, and were performing at chance level.

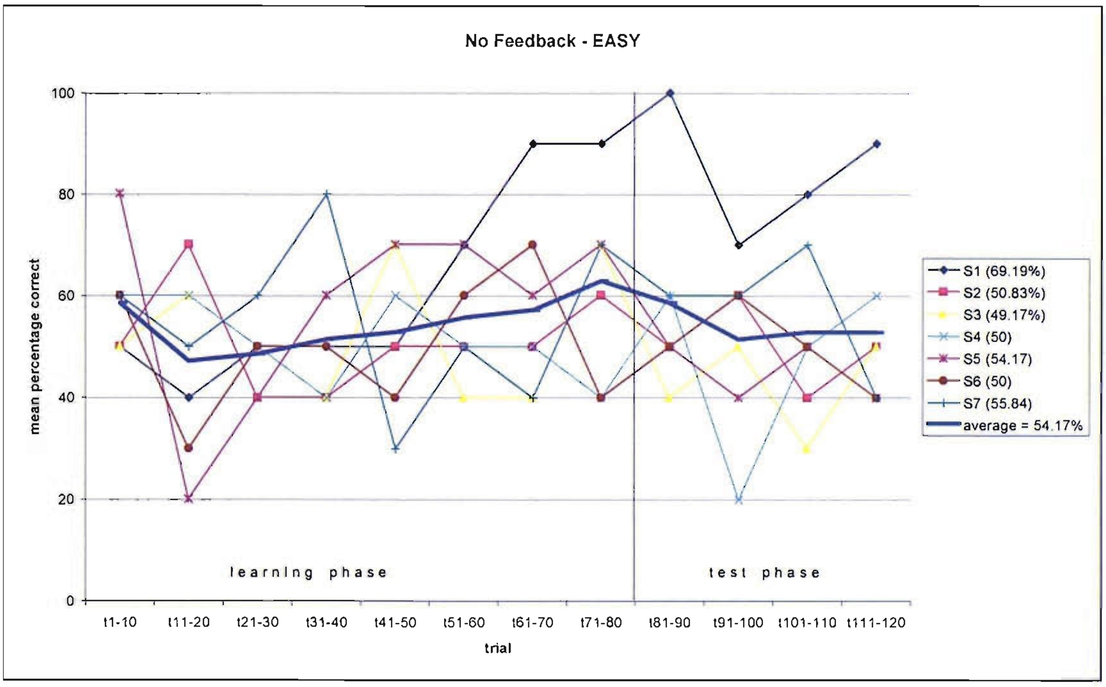


Figure 9.2 Learning curves of Ss with no feedback in easy condition

9.2.3 Verbalisation

Figure 9.3 shows all verbalisers' learning curves and the average learning curves in the feedback and no feedback group in the easy condition. Individual verbalisers from the feedback group are displayed with continuous lines, verbalisers from the no-feedback condition with dashed lines; averages are in boldface. The yellow highlighted arrows show the verbalisation points (i.e. times/points in the experiment when Ss verbalised the rule for the first time), and the orange highlighted arrows show the verbalisation points of the no-feedback group.

Seven of the ten Ss in the feedback group verbalised the rule, while only one of the seven Ss in the no-feedback group could verbalise the rule.

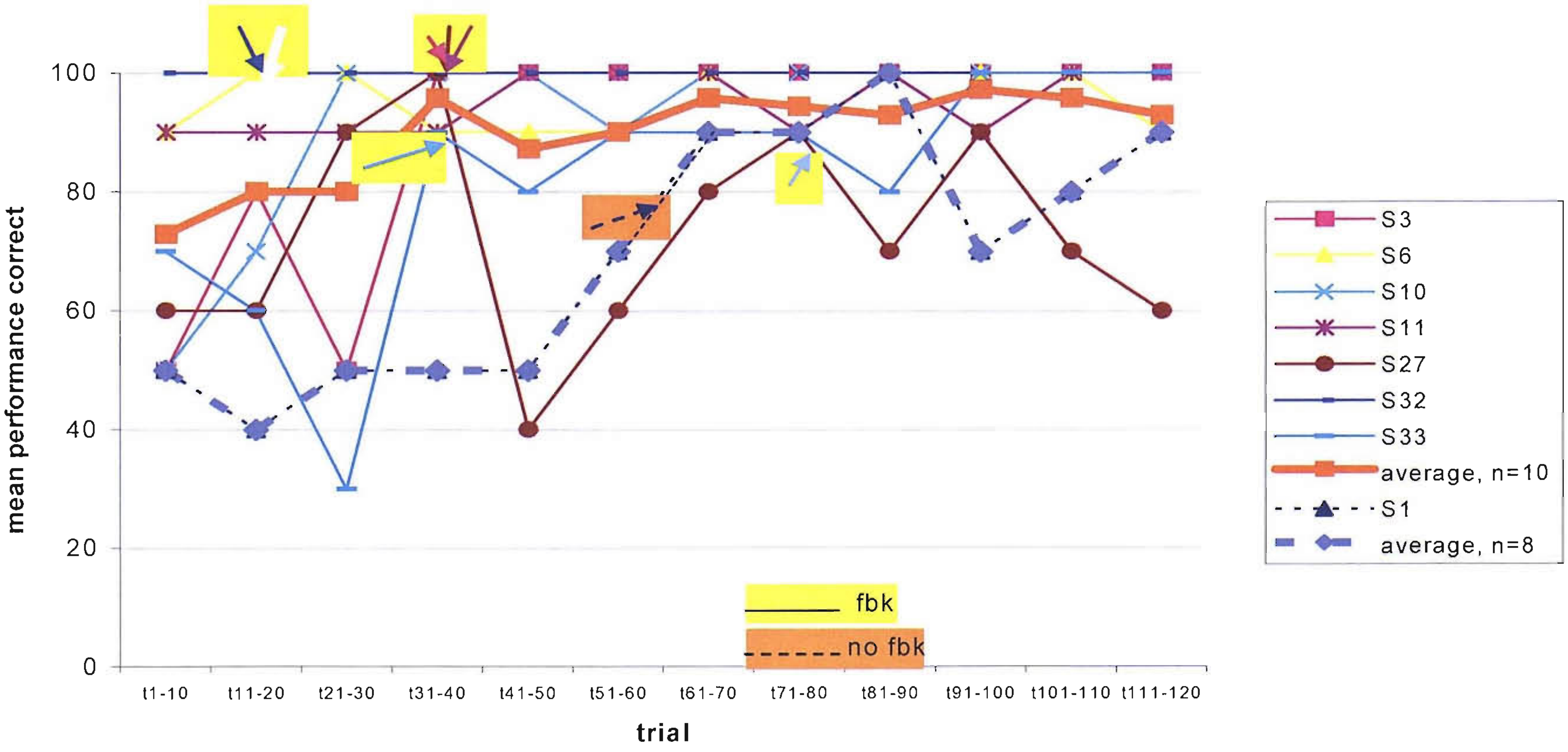


Figure 9.3: Learning curves of verbalisers in the feedback (continuous line) and no-feedback (dashed line) groups with verbalisation points; easy condition

9.3 Medium rule

9.3.1 With Feedback

Nine Ss took part in this condition, 4 male, and 5 female. Ss were performing at an overall average of 80.28% correct in the test phase. The medium rule was precalibrated in the pilot phase to be learnable in about 40 trials. In trials 31-40, Ss were performing at 73.33% correct.

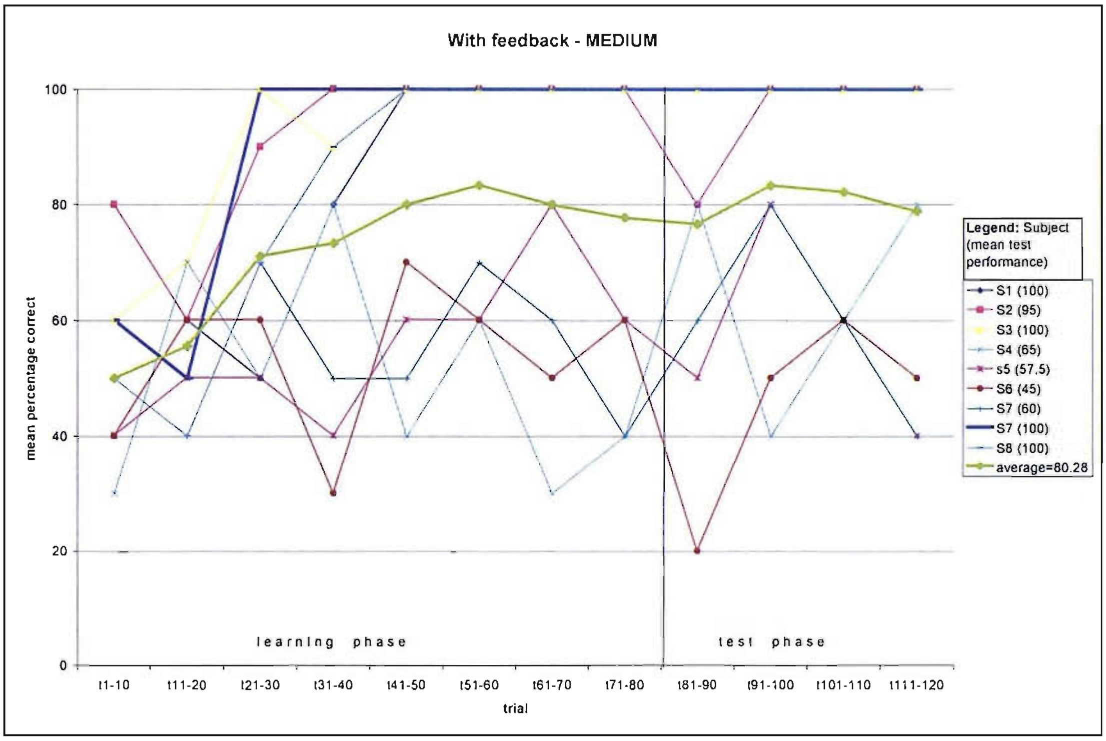


Figure 9.4 Learning curves of Ss in the feedback group in medium difficulty condition

Figure 9.4 shows the learning curves of the Ss with feedback in the medium difficulty condition, with their average in boldface. Most Ss seemed to learn this rule more gradually than the easy rule, first noticing the repeated syllables, and then later noticing the relevance of the position in the string. These Ss were performing at levels above chance when they had learnt the partial rule (repeated syllable), and their performance improved to near 100% correct when they had learnt the additional rule (position relevance). Some Ss only learned partial rules; for example, they noticed that it had something to do with the repeated syllable, but could not figure out the relevance of the position. These partial learners were performing at levels better than chance, but did not reach perfect performance.

9.3.2 With No Feedback

Nine Ss took part in this condition, 6 male and 3 female. Overall test performance was 73.61% correct.

Figure 9.5 shows the learning curves of Ss in the medium condition, who did not receive feedback on their answers. The average learning curve is shown in boldface. Several Ss learned partial rules, but also mentioned that they were very unsure about it. Four Ss (out of a total of nine Ss) learned the rule to approximately 100% correct performance and could explicitly state the rule; one S stated it after 40 learning trials, one after 80 learning trials, the other after all 120 trials (learning and test phase).
One S stated a correct partial rule after 40 learning trials, but was not performing according to this rule. As S received no feedback on whether his responses were correct or not, S could not and did not attribute his correct responses to this partial rule and dropped the rule in favour of other (incorrect) rules.

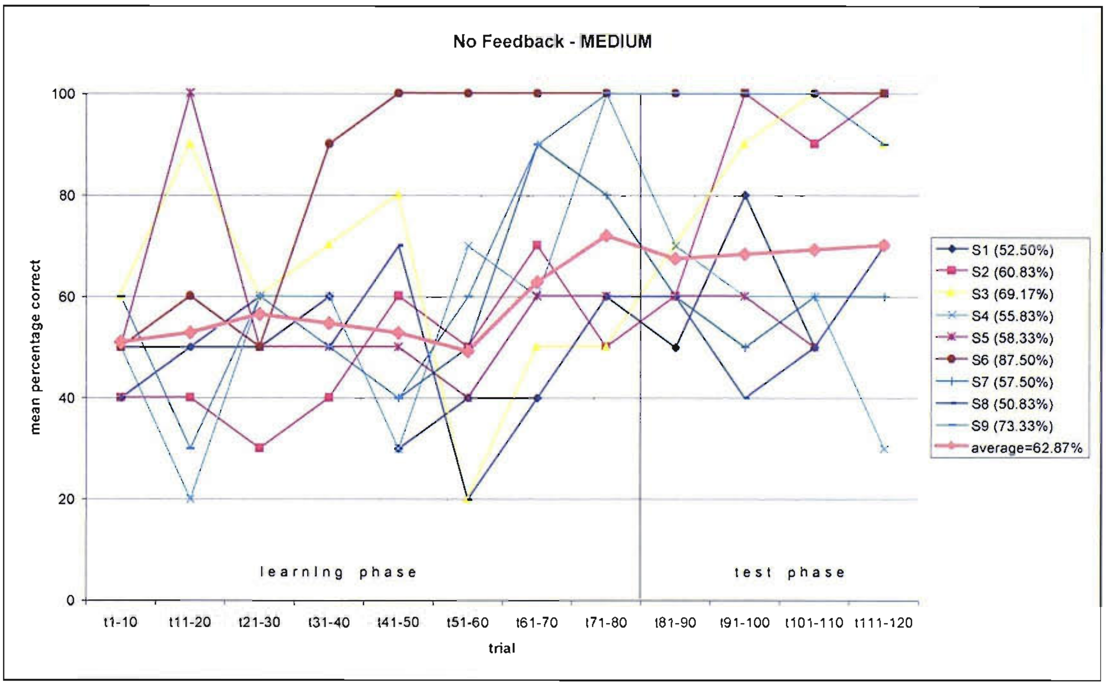


Figure 9.5 Learning curves of Ss with no feedback in medium condition

9.3.3 Verbalisation

Figure 9.6 shows the learning curves of all verbalisers in the feedback and the no-feedback group with the first verbalisation points of each verbaliser. The feedback group is shown with continuous lines, the no-feedback group with dashed lines; the respective average learning curves are in bolder print. The verbalisers in the feedback group are shown with yellow highlighted continuous arrows, and the verbalisers in the no-feedback group are shown with orange highlighted dashed arrows.

The verbalisers in the feedback group were able to verbalise the rule much earlier than the no-feedback verbalisers. It can be seen from this figure that Ss who could verbalise the rule were also performing at high levels (between 80 and 100% correct).

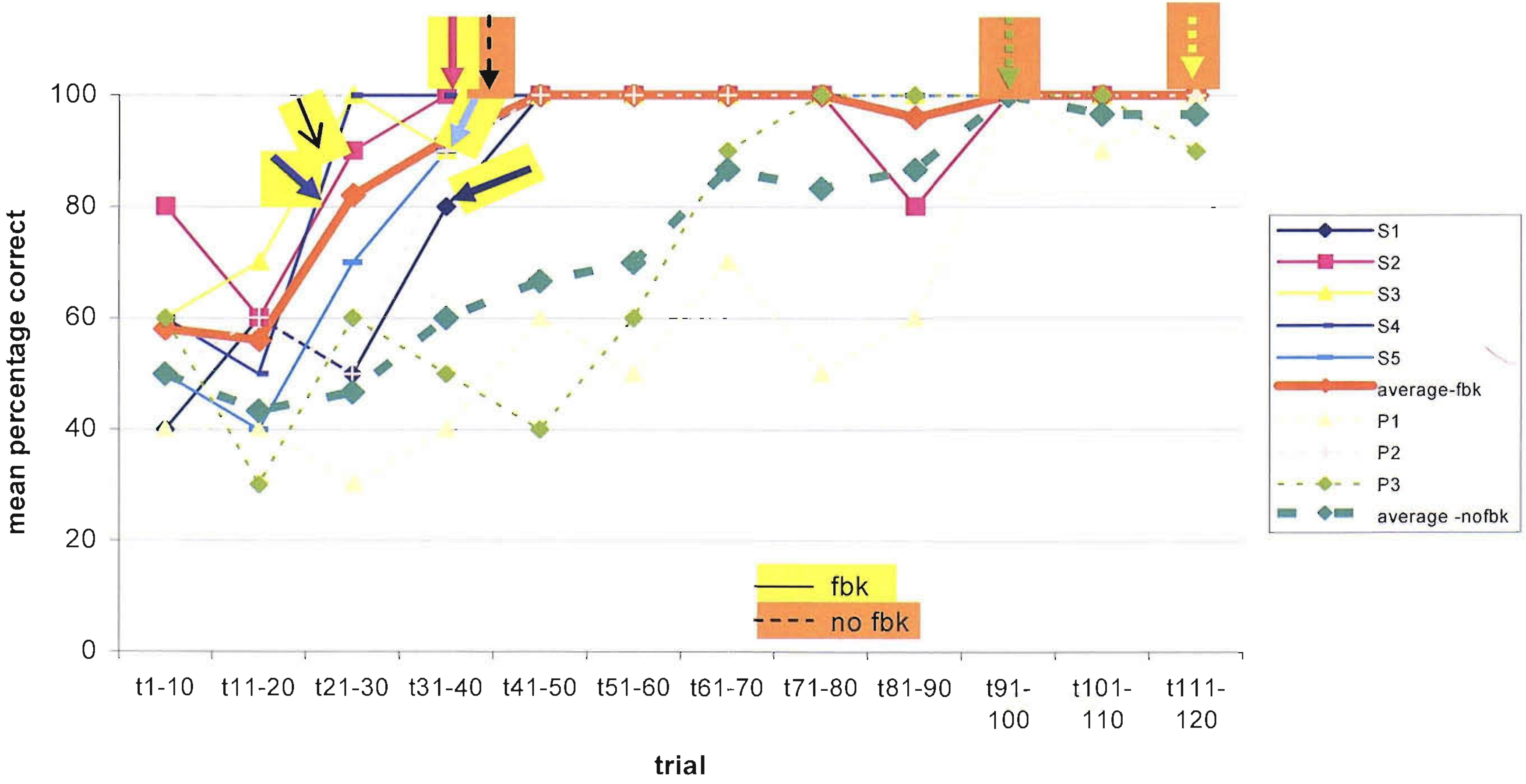


Figure 9.6: Learning curves of verbalisers in the feedback (continuous line) and no-feedback (dashed line) group with initial verbalisation points; medium condition

9.4 Hard rule

9.4.1 With Feedback

Thirteen Ss took part in this condition, 9 male, 4 female. Overall test performance was 63.46% correct. In the pilot studies the hard rule was predetermined to be learnable in about 80 trials. In trials 71-80, Ss were performing at a level of 68.46% correct.

Figure 9.7 shows the learning curves of the Ss with feedback in the hard condition. The boldface line is the average learning curve. Two Ss learned the rule explicitly and could describe it in words. Most Ss learned partial rules, for example “the same syllable may not repeat directly (like Sa Sa)”. Half of the Ss could verbalise the correct rule, or partial rules, while half could not explicitly state the rule.

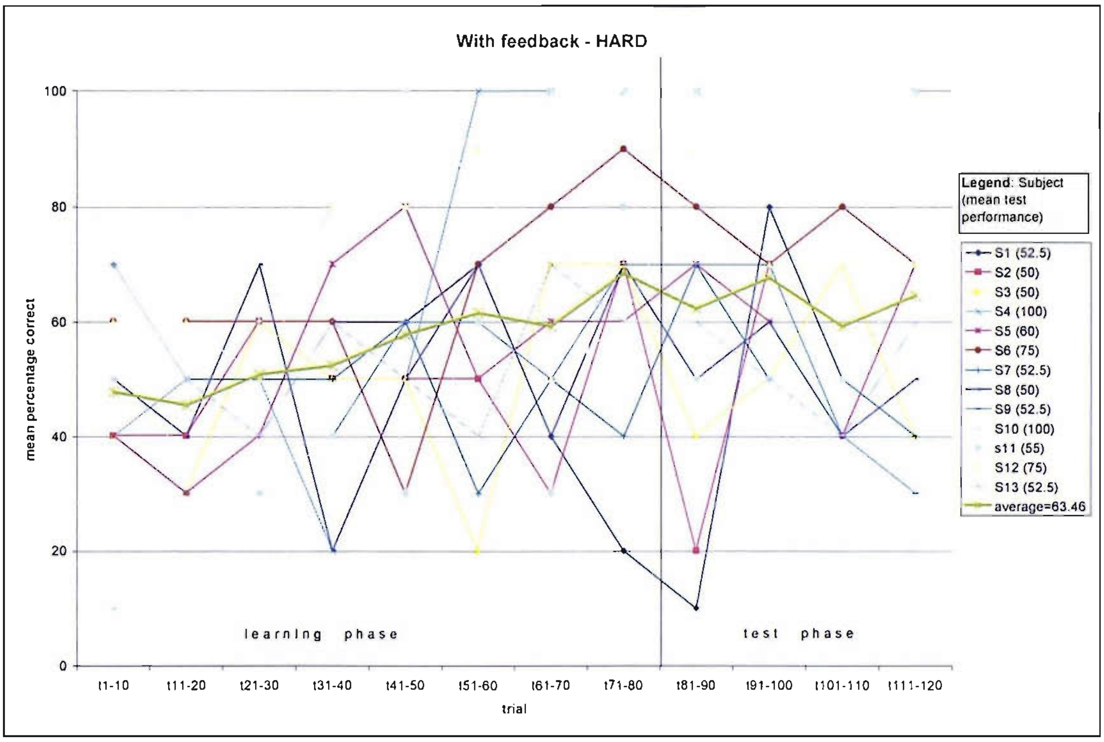


Figure 9.7 Learning curves of Ss in feedback group in the hard difficulty condition

9.4.2 With No Feedback

There were eight Ss in this condition, 5 male, 3 female. Overall test performance of Ss was 56.43%.

Figure 9.8 shows the learning curves of Ss in the hard condition who did not receive feedback on their answers. Two Ss learned the rule and could explicitly state it. Most Ss did not even learn partial rules and were performing at chance level. One S stated a partial rule after 60 learning trials, but the performance does not reflect this. This S also stated many other incorrect rules, but without feedback had no way of knowing whether the rules were correct or not.

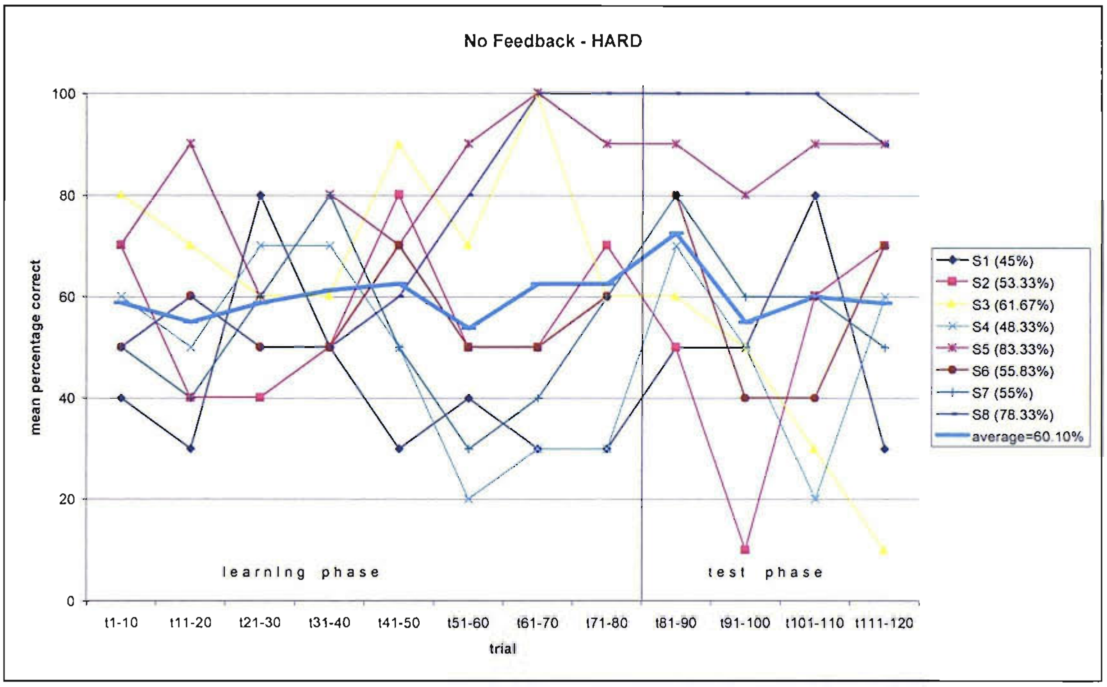


Figure 9.8 Learning curves of Ss with no feedback in the hard condition

9.4.3 Verbalisation

Figure 9.9 displays the learning curves of verbalisers in the feedback group with continuous lines, and of no-feedback verbalisers with dashed lines; the average learning curves are in boldface. The verbalisation points are shown with continuous arrows highlighted in yellow for the feedback group, and dashed arrows highlighted in orange for the no-feedback group.

In the hard condition, Ss in the feedback group verbalised the rule earlier in the experiment than Ss in the no-feedback group.

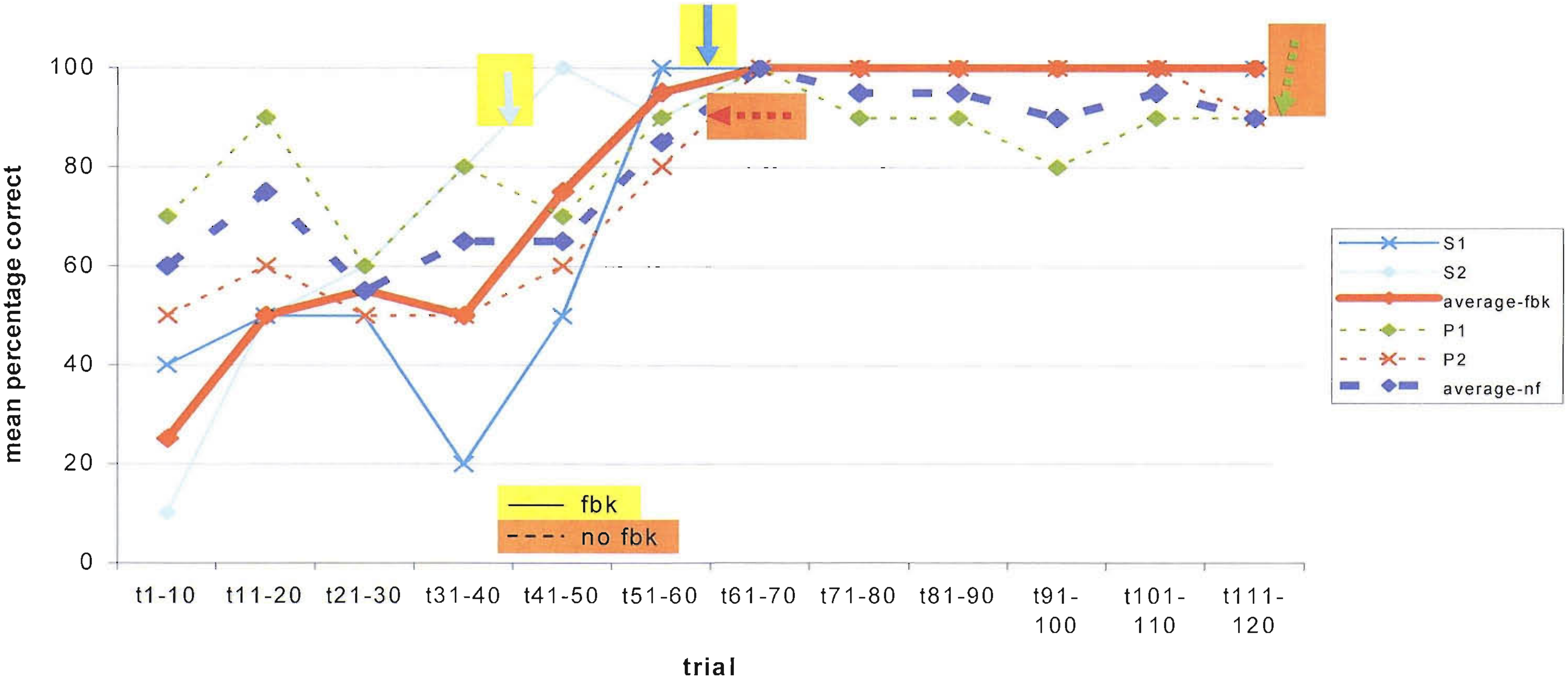


Figure 9.9: Learning curves of verbalisers in the feedback (continuous line) and no-feedback (dashed line) group with initial verbalisation points: the hard condition

9.5 Results

9.5.1 Performance

A two-way factorial ANOVA with test performance as dependent variable and feedback and difficulty as independent variables showed a significant main effect of feedback ($F_{1,50}$ = 8.316, p<0.01), and a significant interaction between feedback and difficulty ($F_{2,50}$ = 4.085, p<0.05).

Figure 9.10 shows the mean percentage of correct answers in the test phase for the three difficulty conditions with and without feedback. Ss in the easy condition were performing significantly better when they received feedback on their answers than when they received no feedback (t=4.870, df=15, p<0.01). T-Tests revealed that there were no significant differences in performance for the medium and hard condition, although it can be seen from the figures that the feedback group were performing at a higher level in the medium and hard condition as well.

It seems that in the feedback condition, the easier the task, the better performance is at test (i.e. Ss in the easy condition were performing better than Ss in the medium condition who were performing better than Ss in the hard condition). However, in the no feedback condition, this did not seem to be the case. In the more difficult conditions (medium and hard rule), receiving feedback did not improve performance. Indeed, collapsing the medium and hard condition into one shows that test performance is virtually the same for the feedback and no feedback conditions. This can be seen in Figure 9.11.

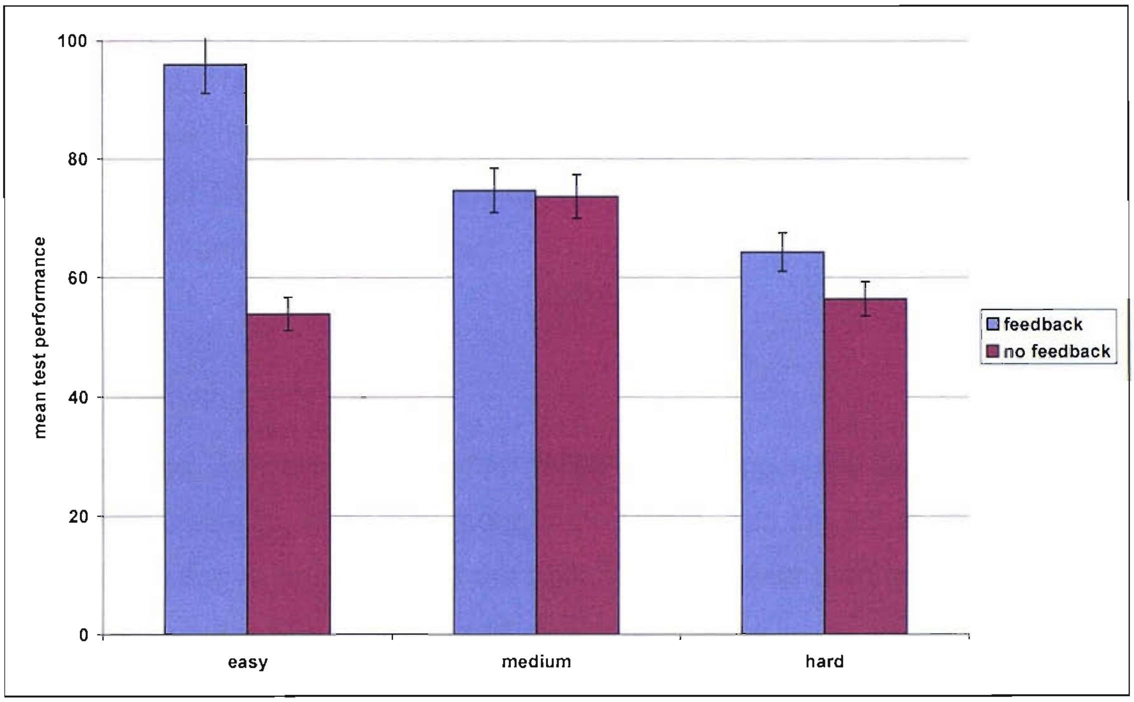


Figure 9.10 Mean percentage of correct answers in the test phase for the three difficulty conditions with and without feedback

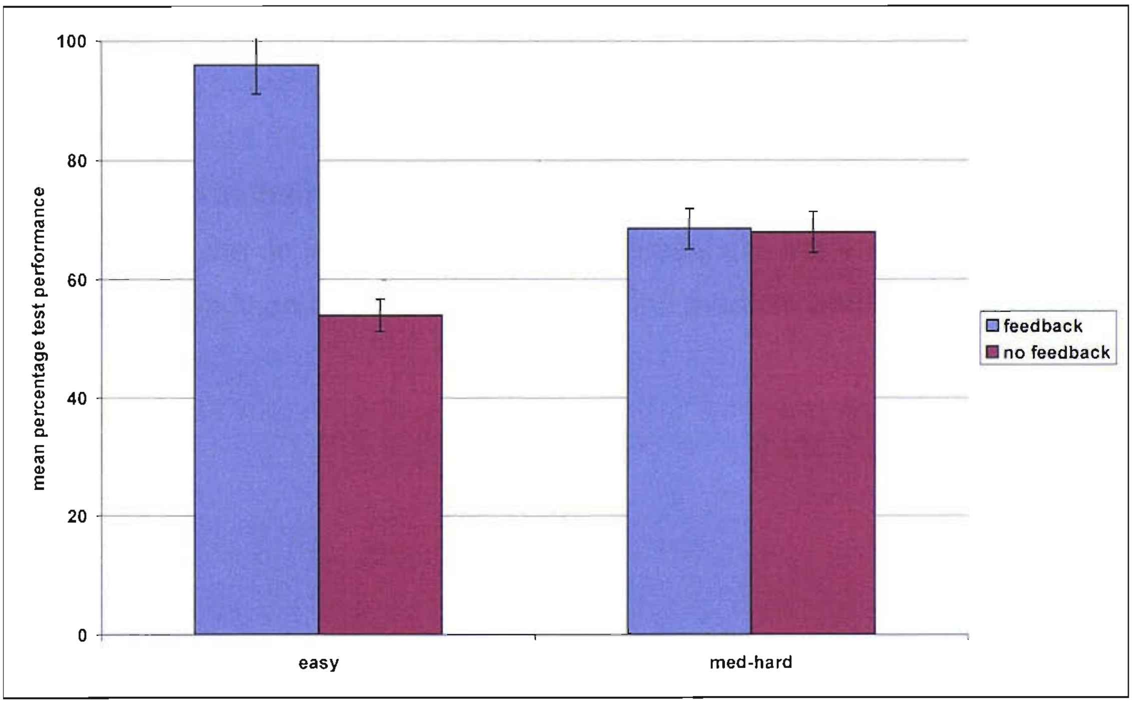


Figure 9.11 Mean percentage of correct answers in the test phase for the easy condition and the more difficult conditions (medium and hard condition collapsed into one) with and without feedback

An analysis of covariance (ANCOVA) was conducted with test performance as dependent variable, feedback and difficulty as independent variables, and time in the experiment that the rule was verbalised (verbalisation time) as covariate. After adjusting for verbalisation time, the interaction between feedback and difficulty is still significant, F=4.484, p<0.05. This indicates that feedback had an effect in the easy condition, while in the medium and hard conditions, feedback had no effect.

9.5.2 Verbalisation

In the breaks, Ss were asked to "verbalise" (by writing in the text box) whether they knew the rules, what they thought they were, and any other strategies they were using to help them do the task. This data was then looked at and it was noted whether Ss could verbalise, partially verbalise or not verbalise the rule(s). It was also noted when in the experiment Ss verbalised (or partially verbalised) the rule(s).

Figure 9.12 shows the number of Ss who could verbalise the rules ("verbalisers") in the feedback and no feedback condition for the easy, medium and hard difficulty conditions. There are more verbalisers in the feedback than in the no feedback condition across the three difficulty conditions. In the no feedback condition, interestingly, there are no verbalisers in the easy condition, but few verbalisers in the medium and hard condition.

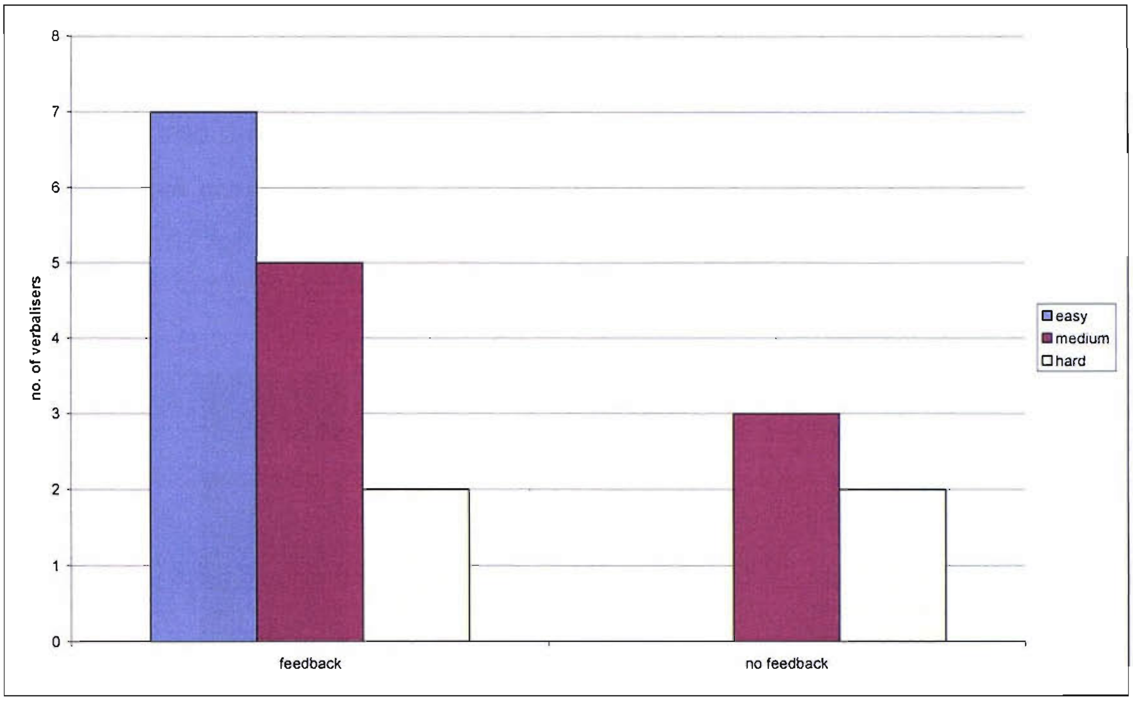


Figure 9.12 No. of verbalisers in the feedback and no feedback condition for each difficulty condition

There was a significant correlation between test performance and being able to verbalise the rule in both the feedback condition (r=0.838, df=30, p<0.01) and the no feedback condition (r=0.814, df=22, p<0.01). Ss who were performing at a high level were also able to verbalise the rule.

A two-way factorial ANOVA with verbalisation as dependent variable and feedback and difficulty as independent variables showed a significant main effect of difficulty ($F_{2,50}$=5.160, p<0.01), and a significant interaction between feedback and difficulty ($F_{2,64}$ = 4.613, p<0.01). This shows that the easier the rule, the more likely Ss were able to verbalise it. However, the interaction between feedback and difficulty indicates that in the no feedback group, the differences between the difficulty conditions did not seem to be as salient. In Figures 9.13., 9.14, and 9.15 the percentages of subjects who could verbalise, partly verbalise and not verbalise the rules are shown for each difficulty condition.

Since the easy rule is a single rule (one of the syllables from the first half of the string has to be repeated in the second half of the string) and cannot be separated into partial rules, it makes sense that no S learnt any partial rules in the feedback condition.

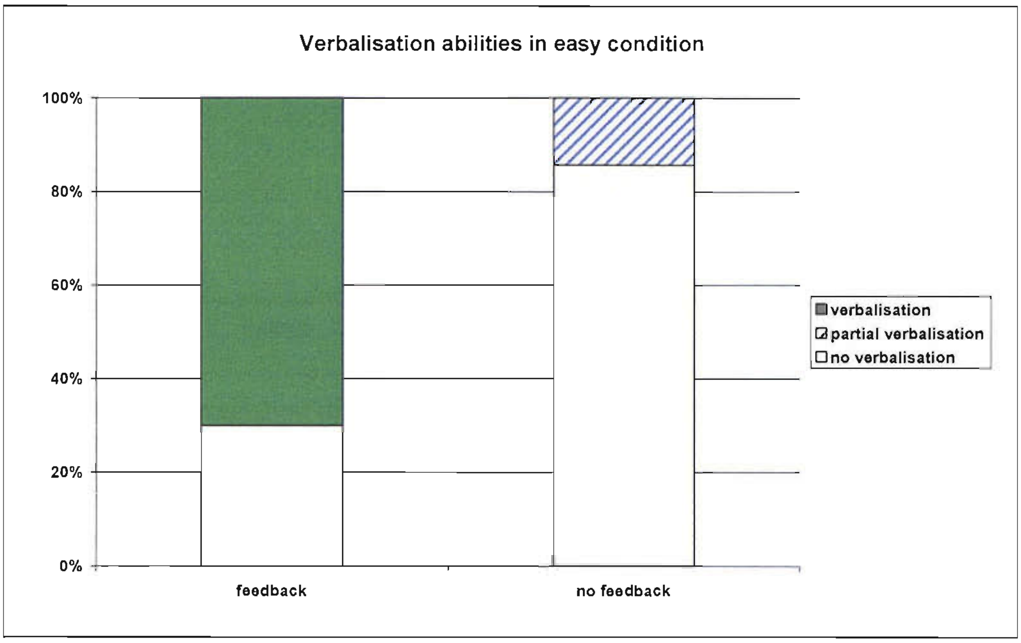


Figure 9.13: Percentage of Ss who could verbalise, partly verbalise and not verbalise the rules in the easy condition

Either you learn the easy rule and can verbalise it or you do not learn it and cannot verbalise it. In the no feedback condition there were several "partial verbalisers". These Ss noticed that some strings started and ended with the same syllable and thought that might have something to do with the rule. This is more like memorisation of certain fragments than learning a partial rule.

The medium and hard rules are both dual rules and can more obviously be separated into partial rules. The most commonly learnt partial rule for both the medium and hard rule was the recurrence of syllables, which however, would not have aided them in their responses, since both ruleful and unruleful strings contained a repeated syllable. Another commonly learnt partial rule was that the repeated syllable could not occur next to each other (e.g. SaHaGaKa-**LaLa**BaRa would be unruleful).

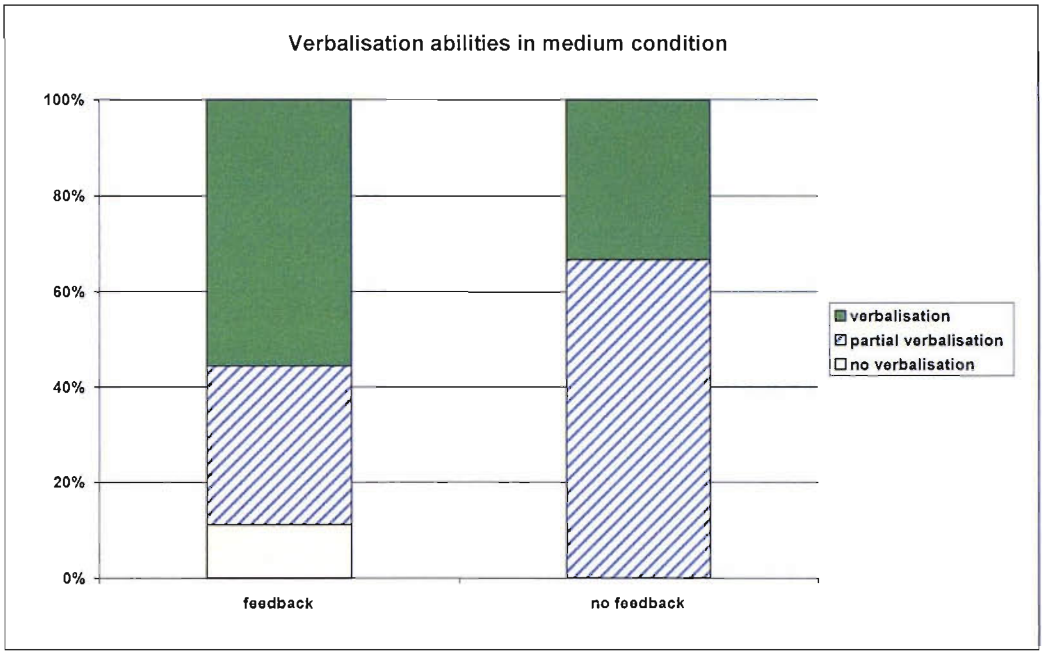


Figure 9.14: Percentage of Ss who could verbalise, partly verbalise and not verbalise the rules in the medium condition

No S in the easy condition with no feedback could verbalise the rule, whereas a few Ss in the medium and hard condition could. Collapsing medium and hard condition into one, as in Figure 9.16, shows that for Ss who received feedback there were equal numbers of verbalisers, partial verbalisers and non-verbalisers in the more difficult condition. For Ss in the no feedback condition there were more partial verbalisers than non-verbalisers. There was an equal number of verbalisers in the feedback and no feedback condition, although from Figure 9.14 and 9.15 it can be seen that most of the verbalisers in the feedback condition were in the medium condition.

Overall performance level was also significantly correlated with how long it took until the rule could be verbalised ($r=-0.632$, $df=12$, $p<0.05$) by Ss who received feedback. Those Ss who were performing well overall were also verbalising the rule earlier.

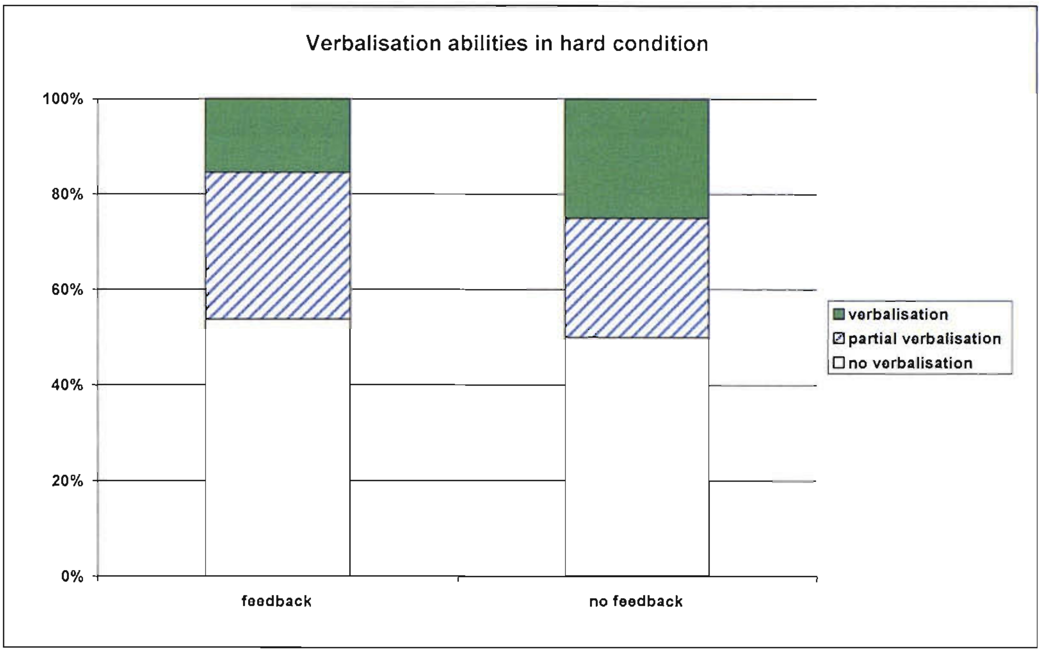


Figure 9.15: Percentage of Ss who could verbalise, partly verbalise and not verbalise the rules in the hard condition

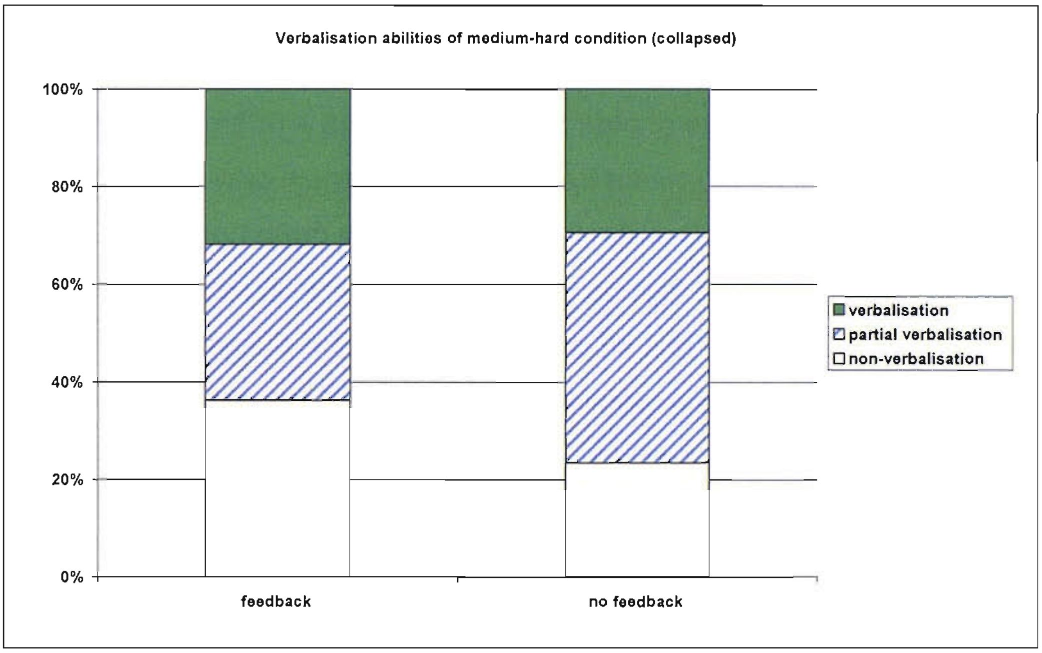


Figure 9.16: Percentage of Ss who could verbalise, partially verbalise and not verbalise the rules in the feedback and no feedback conditions when the medium and hard condition are collapsed into one.

9.6 Conclusions

The results of this experiment show that Ss who received immediate feedback on their responses performed better than Ss who did not receive feedback. This effect was particularly strong in the easy condition. Ss in the no feedback condition did learn some rules, mostly partial rules. However, they found it hard to ascertain whether the rules they were using were yielding any correct results. The no feedback Ss received delayed feedback after every 20 trials informing them of how many trials (out of 20) they got right in the previous block, but they did not know which feature to attribute their correct results to; they were struggling with the credit/blame assignment problem (see Section 2.2). This became especially clear in those no-feedback Ss who stated the correct rule or a partial rule, but then abandoned it later on, attributing their correct results to other, incorrect rules they were entertaining.

A further finding of this experiment was that those Ss who were performing at or very close to 100% correct were also able to explicitly verbalise the rule(s) accurately. Ss who were performing at an intermediate level (i.e. between 60 and 75% correct) often verbalised partial rules. Ss who were performing at (or close to) chance could not verbalise any correct rules at the end of the experiment. There is a chance that some Ss first learnt the rule implicitly, and only later became explicitly aware of how they were getting it correct. Since Ss in this experiment were only asked to verbalise their strategies and any rules they thought they knew in the breaks after every 20 trials, it is difficult to judge whether they learnt the rule implicitly first. Nevertheless, all Ss who were performing well at the end of the experiment (and who was actually writing down their rules and strategies as instructed) could also verbalise the rules.

10 Effect of Response on Learning

This experiment analysed the effect of response type (active versus passive) on learning. In both the active and the passive group, Ss were shown a string and asked to click on the right mouse button and say the syllables aloud to themselves. This was to ensure that Ss were paying attention serially to each syllable and not just processing the stimulus as a whole. In the active group, Ss were then required to make an active response concerning the rulefulness of the string. They had to signal whether or not they thought it followed the rule(s) by clicking on either of two buttons 'Yes' or 'No'. Ss were then told whether or not their response was correct. The passive group made no active response and was merely told after each presentation whether or not the string had followed the rule. It was expected that active trial-and-error responding followed by feedback (compared to mere passive exposure followed by feedback) would have a positive effect on learning (success and/or speed), i.e. that active responders would perform better.

10.1 Method

10.1.1 Design

The design of this experiment is a 3 x 2 design with difficulty (easy, medium, and hard) and response type (active responding and passive exposure) as independent variables and task performance and verbalisation ability as dependent variables. Since these were web-based experiments, Ss were asked to write down whether they knew any rule(s), what rule(s) they thought they knew, any strategies they were using, and any other comments they had after every 20 trials and at the end of the experiment. These written comments were then analysed as the verbalisation data (see also Section 8.1.4 and Figure 8.3 for a screenshot).

10.1.2 Subjects

50 Ss took part in this experiment. 32 in the active group, 17 in the passive group. In the active group, 10 took part in the easy experiment, 9 in the medium experiment, and 13 in the hard experiment. In the passive group, 6 Ss took part in the easy experiment, 6 took part in the medium experiment and 5 in the hard experiment.

10.1.3 Procedure

Ss in all three difficulty conditions (i.e. easy, medium and hard) in both groups were presented with 80 positive and negative learning strings and then 40 new positive and negative test strings. Active Ss were told that they should try to learn the rule(s) governing the strings. They were shown a string of syllables and asked to click on the right mouse button while saying each successive syllable aloud. Each click on the right mouse button made the corresponding syllable appear in a box beneath it, to be said aloud. After completing the entire string, they were asked whether or not they thought it had been ruleful. They responded by clicking the 'Yes' or 'No' button. After responding they received feedback on whether or not their answer had been correct. Then the next trial appeared. This active condition is the same as the feedback condition in Chapter 9.

Passive Ss were told that their task was to try to learn the rule(s). On each trial they were presented with a string, which they had to repeat aloud while clicking the right mouse button, like the active group. Once they had repeated the string, Ss were told whether or not the string was ruleful without having to make a response.

The test phase was the same for both active and passive Ss. They were presented with novel positive and negative strings and again had to repeat the string aloud while clicking on the right mouse button. Once they had repeated the string, they were asked whether the string had been ruleful and had to click Yes or No. They were not given any feedback on their responses in the test phase.

Ss were asked after every 20 trials in the learning and test phase whether they knew the rules and what they thought they were. They were asked to write down everything they were thinking. Ss in the active group were also told on how many trials they had responded correctly in the immediately preceding block of 20 trials.

At the end of the experiment all Ss were told on how many trials they had responded correctly in the test phase and what the rule(s) were.

10.2 Easy rule

10.2.1 Active Response

Since the active response condition is the same as the condition with feedback in Chapter 9 (see Section 9.2.1), the same results were used. These have been reproduced here for convenience and comparison purposes.

10 Ss took part in this condition, 8 male, 2 female. The average performance in the test phase was 89.25% and Ss were performing at an average of 75% in trials 11-20.

In Figure 10.1 (= Figure 9.1) the learning curves of all Ss in the feedback group in the easy condition are shown with the average learning curve in boldface. Most Ss could correctly distinguish ruleful from unruleful strings after about 30 trials, although most were also still making some mistakes in later trials, probably due to fatigue or lack of concentration.

Seven of the ten Ss could describe the rule after the maximum 40 trials. As soon as these Ss learnt the rule, performance improved from chance to near perfect, and they verbalised the rule at first opportunity (i.e. in the next break). Of the remaining three Ss, one did not write anything down, but was performing at a high level of correctness; a further S only verbalised incorrect rules, and his overall performance in the test phase was only 55%. A further S stated the correct rule after 20 test trials (so after completing the 80 learning trials and 20 of the 40 test trials), but is performing at a very high level (90% correct) from learning trial 60 onwards. This S may have learnt to respond correctly (perhaps also implicitly) at an earlier stage than his ability to verbalise the correct rule. However, it is also possible that this S did not write down the rule until completely certain of it. Ss were asked to write down any rules they were using during the experiment, but were not explicitly asked from which trial onwards they were using these rules from.

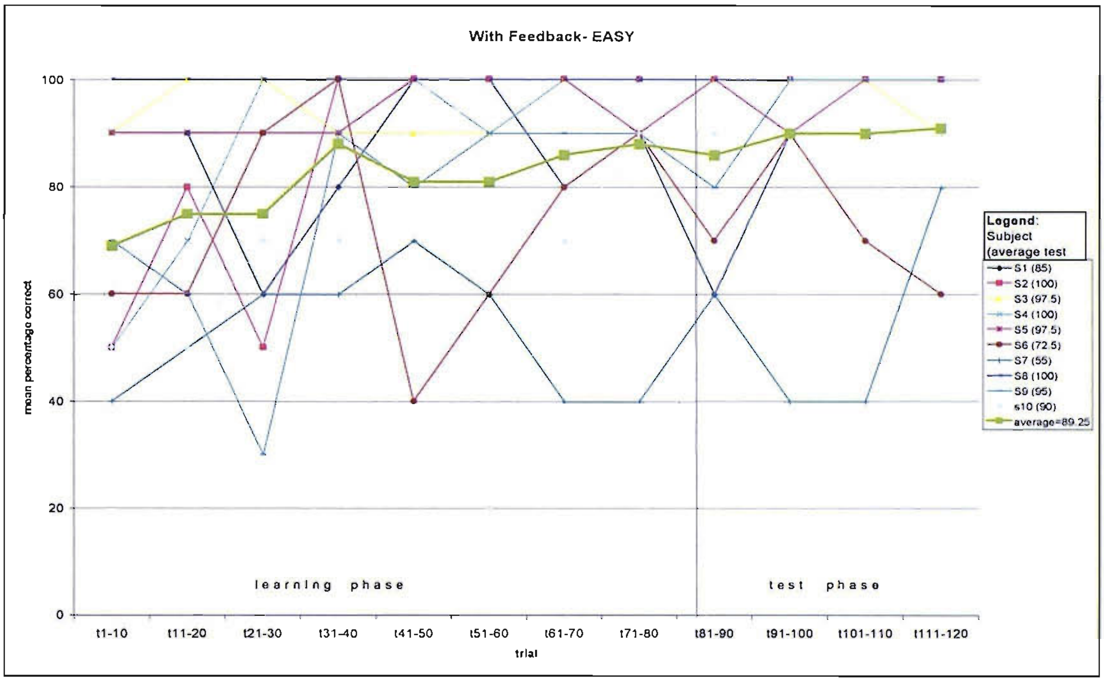


Figure 10.1: Learning curves for Ss in the feedback group in the easy condition (= Figure 9.1)

10.2.2 Passive Exposure

There were 6 Ss in the passive group: 3 male, 3 female. Average performance in the test phase was 82.08% correct.

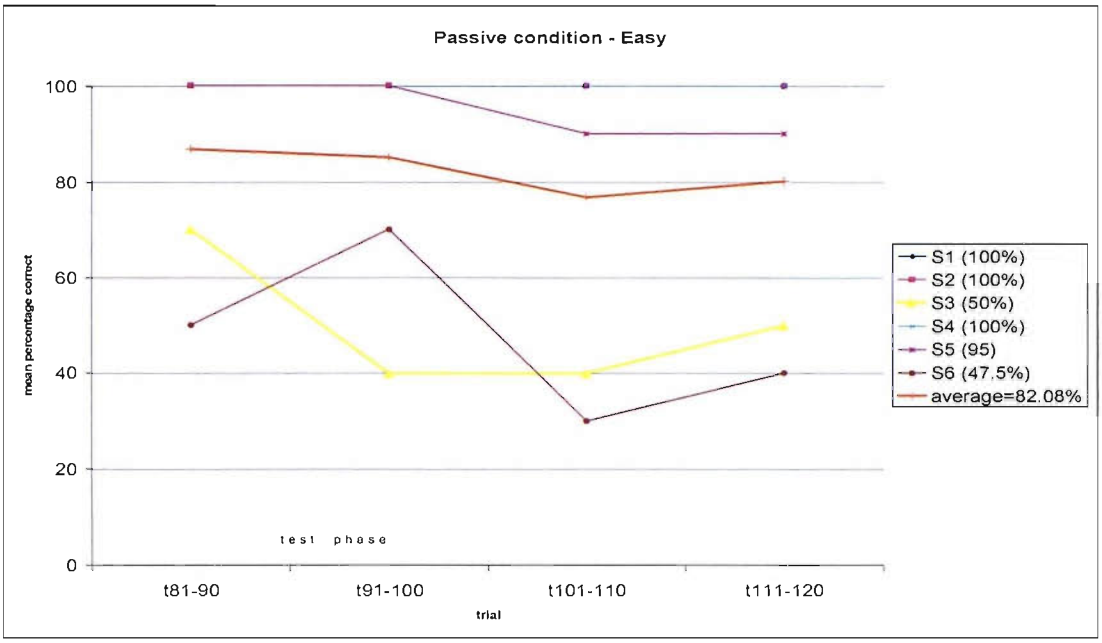


Figure 10.2: Learning curve of Ss in the passive group in the easy condition (since Ss are not making any response in the training phase there is no data from the training phase and only the test phase is shown)

In Figure 10.2 the performance curves of all Ss in the test phase (since they were not making any response in the training phase) are shown. The average curve is depicted in boldface. All Ss who were performing at 100% (or close to it) were able to state the rule explicitly. A further S explicitly stated a partial rule, but was performing at chance in the test phase. Another S was performing at chance and only wrote down incorrect rules. Ss seem to be getting worse over the 40 trials of the test phase. This can be attributable to fatigue and lapses in concentration, as evidenced by Ss comments ("lost concentration there for a while", "I'm getting bored now").

10.2.3 Verbalisation

Figure 10.3 shows the learning curves of all verbalisers with their average learning curve shown in boldface. Ss in the active group are shown with continuous lines, the passive group with dashed lines. The verbalisation points of verbalisers for the active group are indicated by yellow highlighted continuous arrows, for the passive group by orange highlighted dashed arrows.

Most verbalisers in both groups verbalised the rule early in the training phase, after 20 learning trials. One S in the passive group verbalised the rule after 80 learning trials. All verbalisers are performing at high levels.

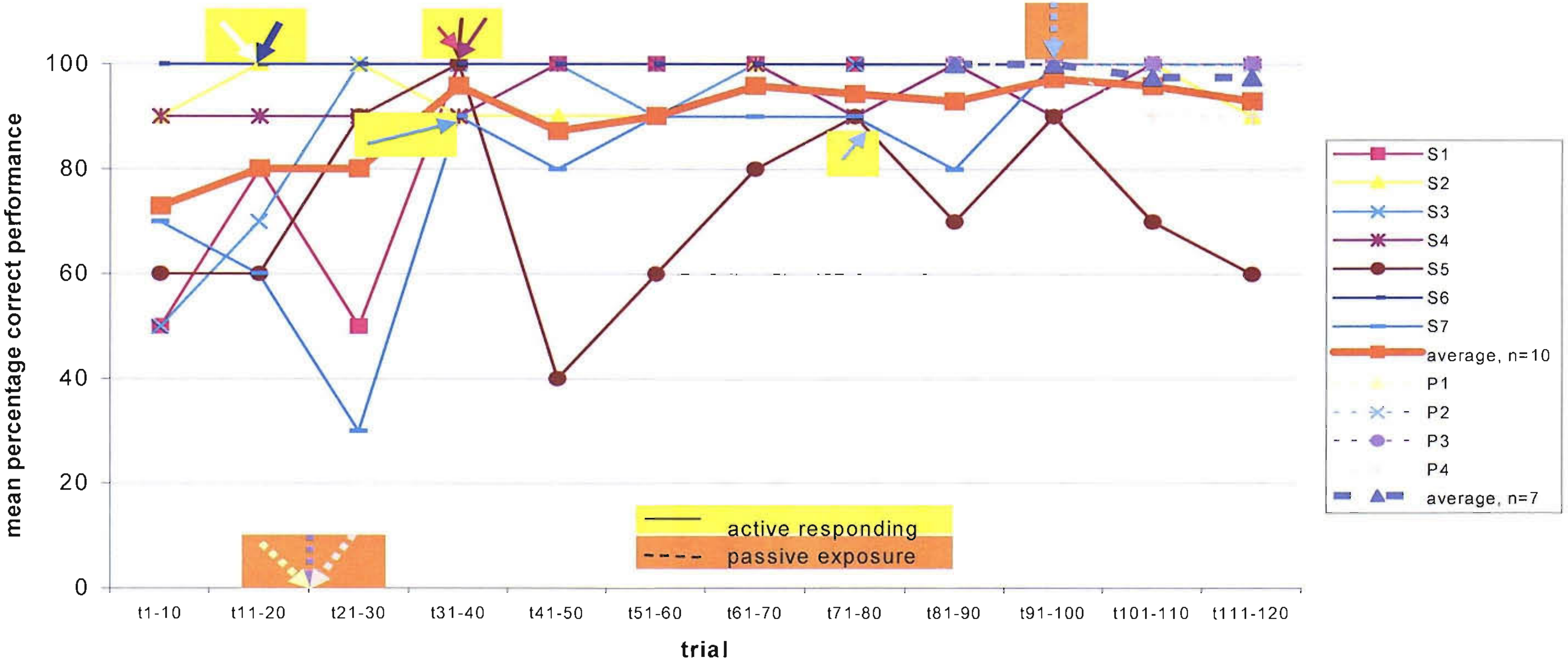


Figure 10.3: Easy condition: Learning curves for all verbalisers in the active (continuous line) and passive group (dashed line), with points at which rule was successfully verbalised

10.3 Medium rule

10.3.1 Active Response

The results of Ss in the active group were the same as those in the feedback group in section 9.3.1: Nine Ss took part in this condition, 4 male, and 5 female. Ss were performing at an overall average of 80.28% correct in the test phase. The medium rule was precalibrated in the pilot phase to be learnable in about 40 trials. In trials 31-40, Ss were performing at 73.33% correct.

Figure 10.4 (=Figure 9.4) shows the learning curves of the Ss with feedback in the medium difficulty condition, with their average in boldface. Most Ss seemed to learn this rule more gradually than the easy rule, first noticing the repeated syllables, and then later noticing the relevance of the position in the string. These Ss were performing at levels above chance when they had learnt the partial rule (repeated syllable), and their performance improved to near 100% correct when they had learnt the additional rule (position relevance). Some Ss only learnt partial rules; for example, they noticed that it had something to do with the repeated syllable, but could not figure out the relevance of the position. These partial learners were performing at levels better than chance, but did not reach perfect performance.

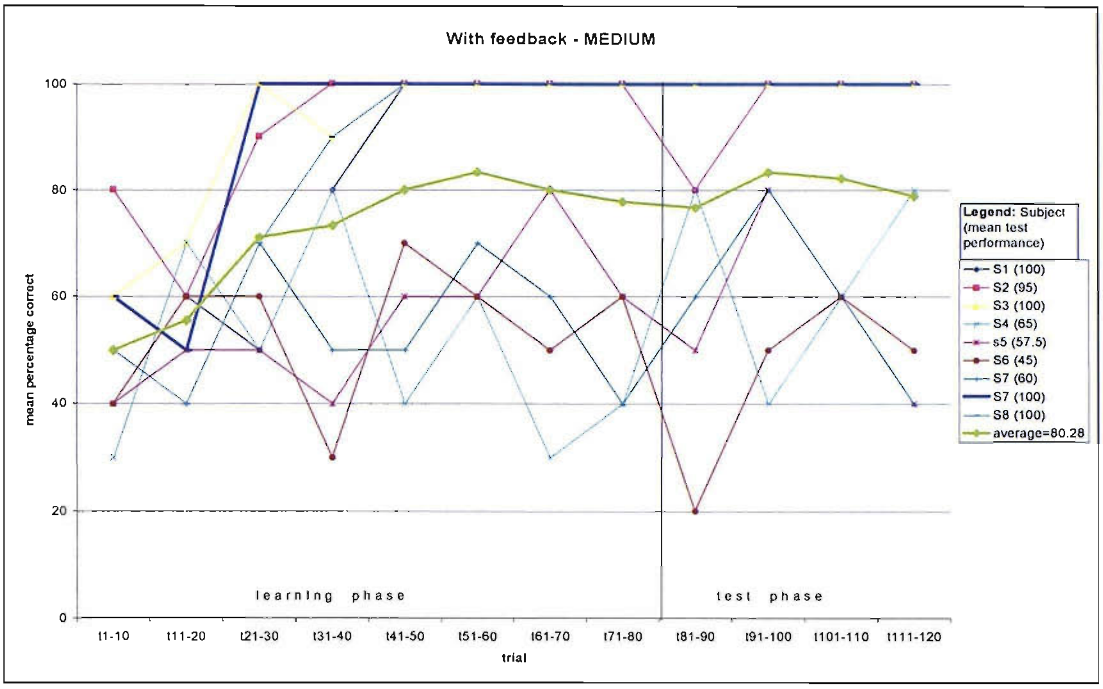


Figure 10.4 Learning curves of Ss in the feedback group in medium difficulty condition (=Figure 9.4)

10.3.2 Passive exposure

Six Ss took part in this experiment: 4 male, 2 female. Average performance in the test phase was 90.32%.

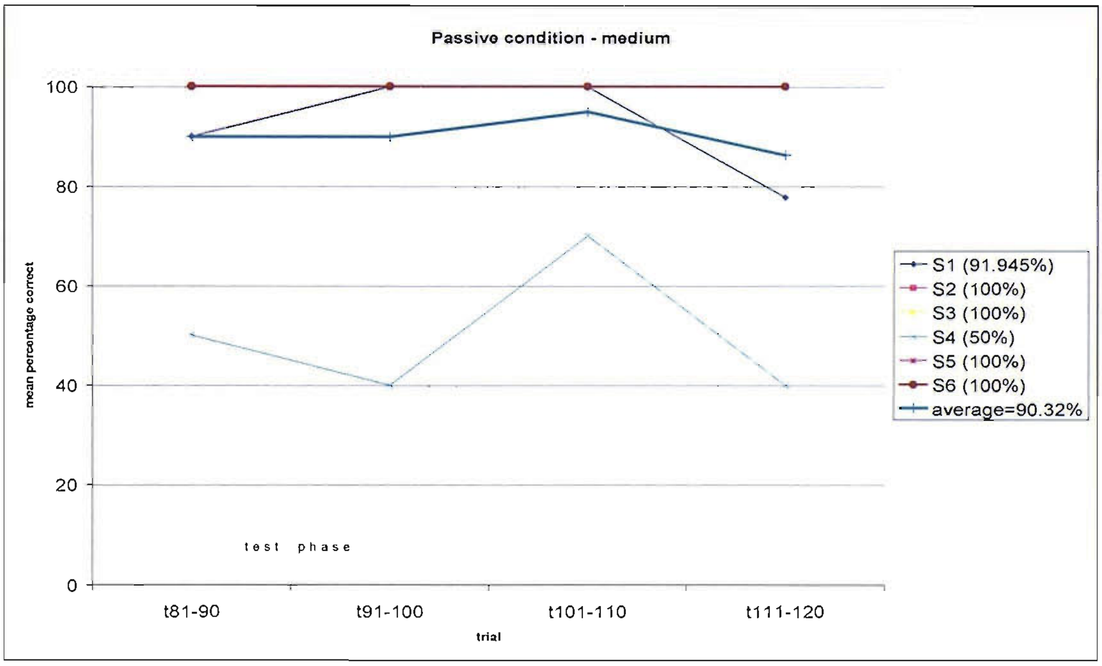


Figure 10.5: Learning curve of Ss in passive group in medium condition (since Ss are not making any response in the training phase there is no data from the training phase and only the test phase is shown)

Figure 10.5 shows the test performance curves of all Ss in the medium condition. The average performance curve is shown in boldface. All Ss who were performing above chance could also state the rule explicitly. The sole S who was performing at chance level, did not write anything down.

One subject hinted in the learning phase that he may be experiencing implicit learning (even though he was not actually making an active response: "now I'm confused (though I'm guessing right most of the time think the rule might be in my subconscious somewhere?!?)". This S was performing very well in the test phase and explicitly stated the rule after 20 test trials.

10.3.3 Verbalisation

Figure 10.6 shows the learning curves of all Ss in the active group with continuous lines, and all Ss in the passive group with dashed lines with their averages in boldface. The verbalisation points are indicated for each of the verbalisers by highlighted arrows. Verbalisers in the active group are depicted with yellow highlighted continuous arrows, while verbalisers in the passive group are depicted with orange highlighted dashed arrows.

Verbalisers in both groups were verbalising in the first half of the training phase at about the same time, after either 20 or 40 learning trials.

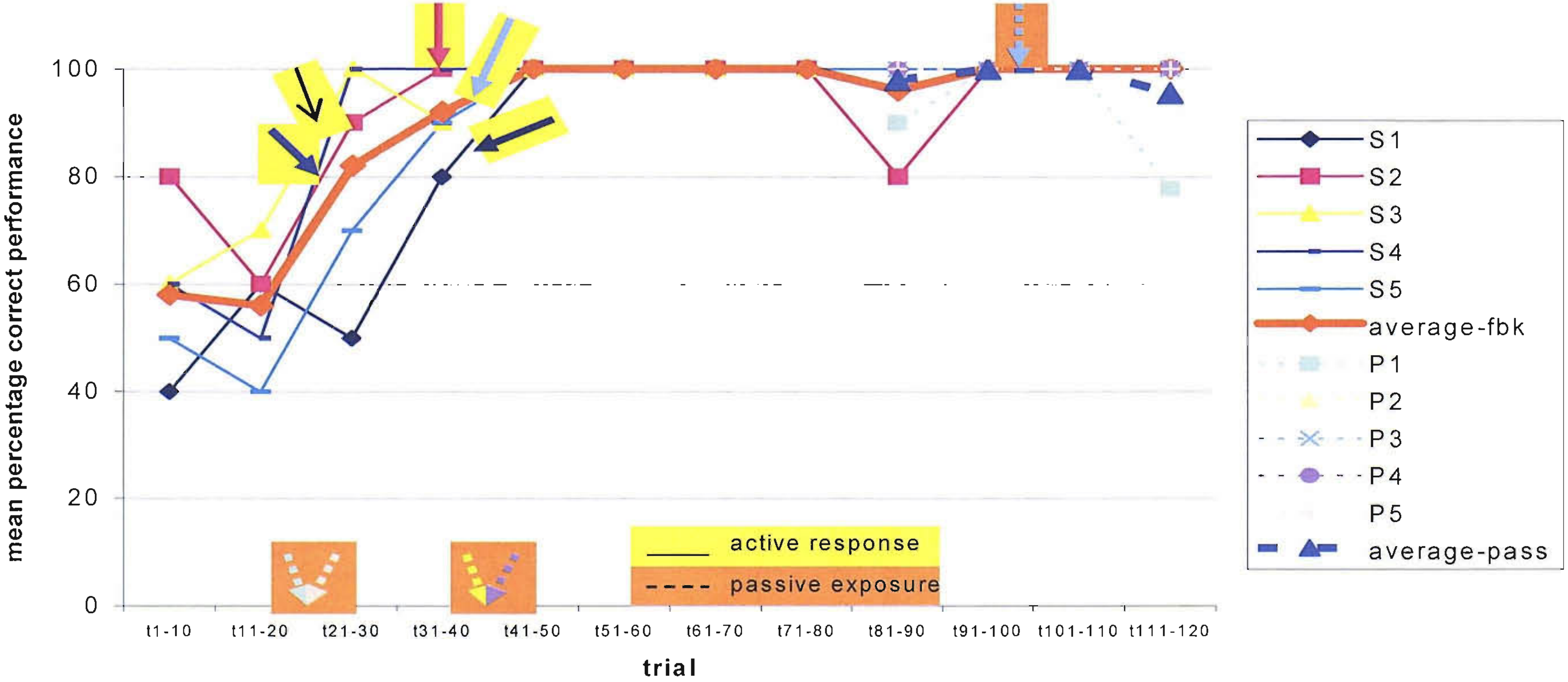


Figure 10.6: Medium condition: Learning curves for all verbalisers in the active (continuous line) and passive group (dashed line), with points at which rule was successfully verbalised.

10.4 Hard rule

10.4.1 Active Response

Learning curves and percentages are the same for Ss in the hard active response group as in the hard feedback group, see Section 9.4.1:

Thirteen Ss took part in this condition, 9 male, 4 female. Overall test performance was 63.46% correct. In the pilot studies the hard rule was predetermined to be learnable in about 80 trials. In trials 71-80, Ss were performing at a level of 68.46% correct.

Figure 10.7 (=Figure 9.7) shows the learning curves of the Ss with feedback in the hard condition. The boldface line is the average learning curve. Two Ss learned the rule explicitly and could describe it in words. Most Ss learned partial rules, for example "the same syllable may not repeat directly (like Sa Sa)". Half of the Ss could verbalise the correct rule, or partial rules, while half could not explicitly state the rule.

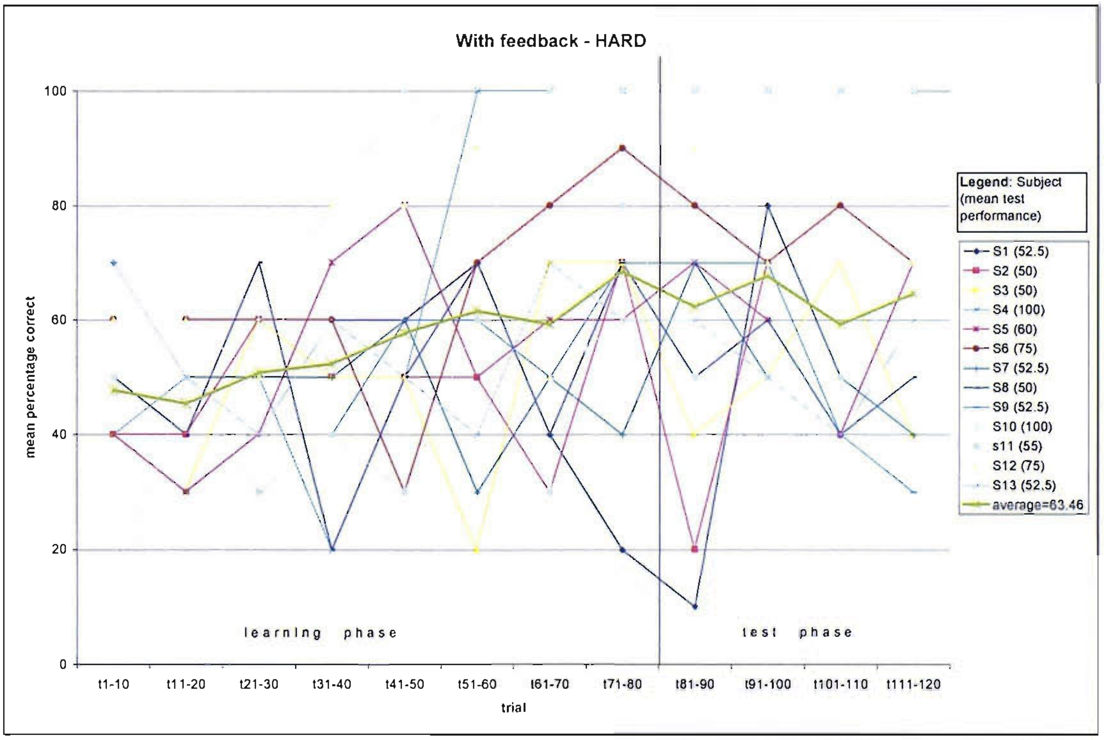


Figure 10.7 Learning curves of Ss in feedback group in the hard difficulty condition (= Figure 9.7)

10.4.2 Passive exposure

5 Ss took part in this experiment: 2 male, 3 female. Average performance in the test phase was 72% correct.

Figure 10.8 shows the test performance curves of all Ss in the hard condition. The mean performance curve is shown in boldface. All Ss who were performing at 100% could explicitly state the rule. The S who was performing above chance (67.5%) did not write anything down. Some other Ss learnt partial rules, but their performance did not reflect this as they were performing at chance.

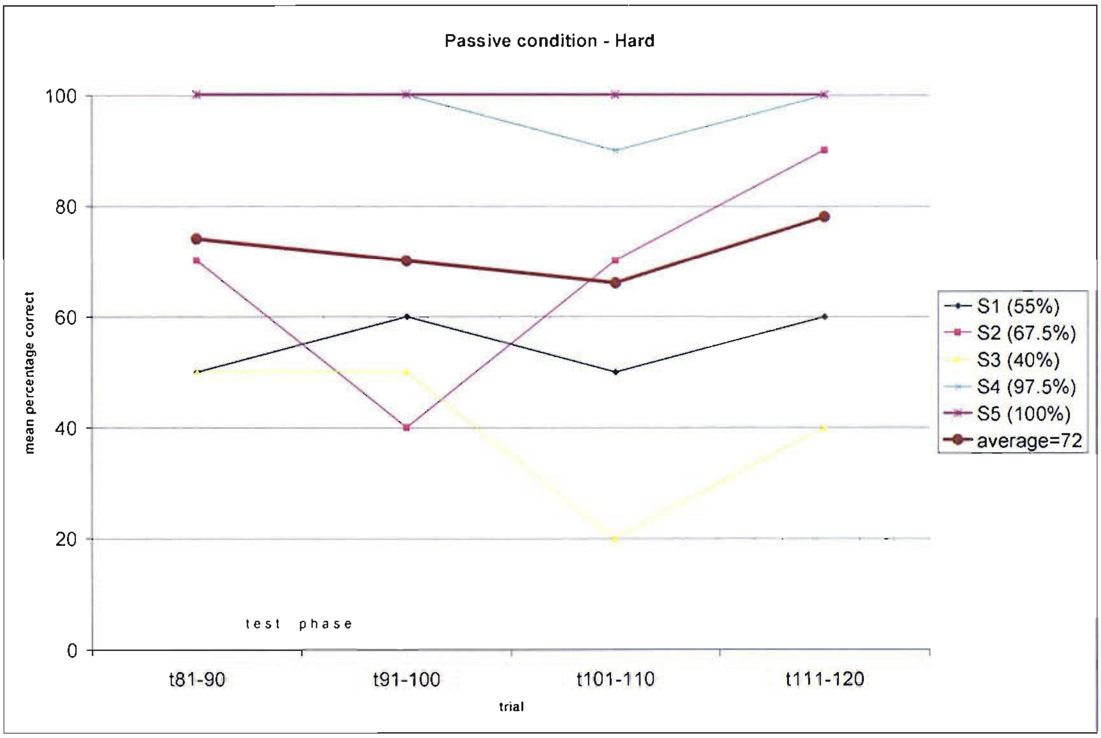


Figure 10.8: Learning curve of Ss in passive group in hard condition(since Ss are not making any response in the training phase there is no data from the training phase and only the test phase is shown)

10.4.3 Verbalisation

In Figure 10.9 the learning curves of all verbalisers in the active group are shown by a continuous line, verbalisers in the passive group by dashed lines. The verbalisation points are indicated by highlighted arrows, yellow continuous arrows for the active group, and orange dashed arrows for the passive group.

In the hard condition, Ss in the passive group were verbalising the rule earlier than Ss in the active group. Passive verbalisers stated the rule after 20 learning trials, active verbalisers stated the rule after 40 or 60 learning trials.

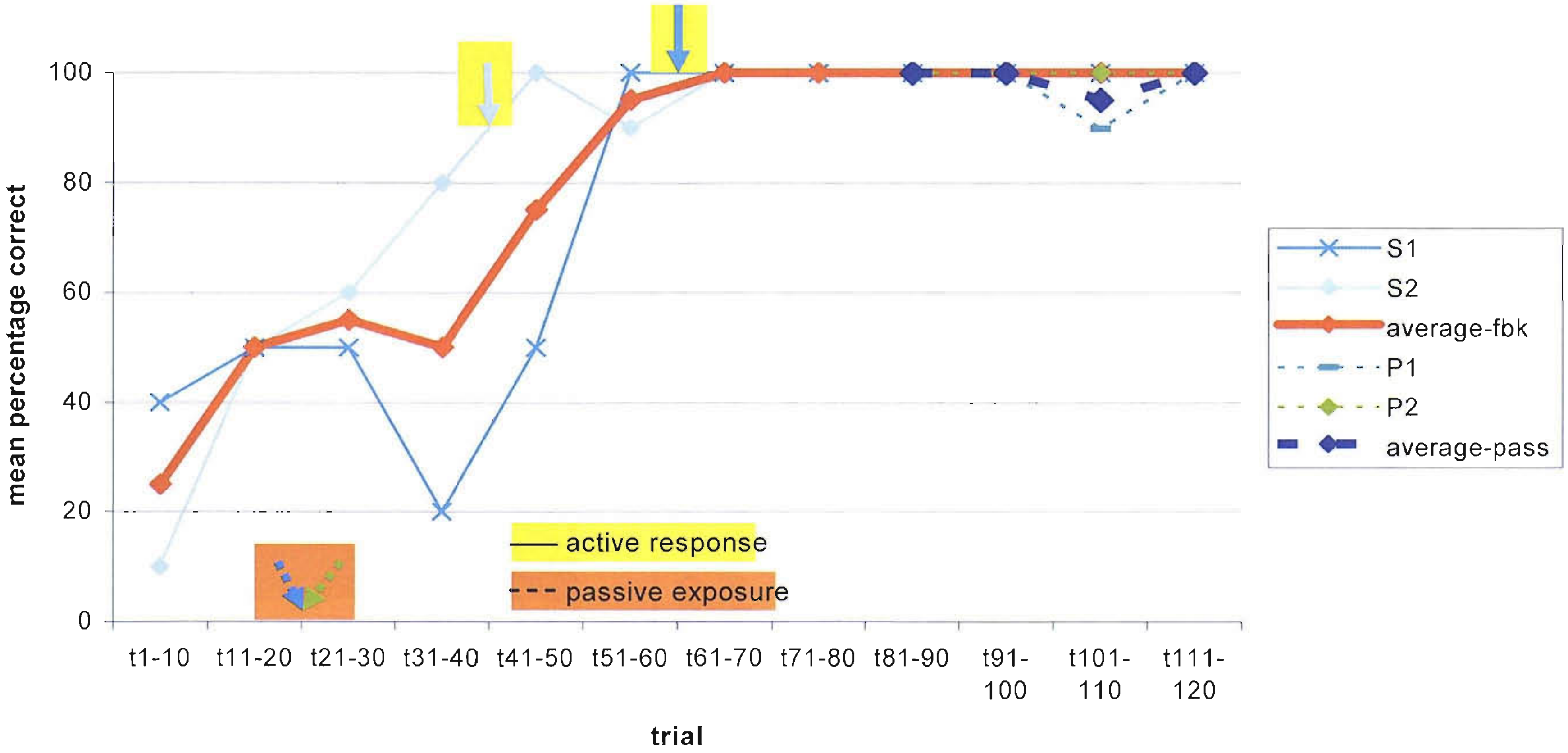


Figure 10.9: Hard condition: Learning curves for all verbalisers in the active (continuous line) and passive group (dashed line), with points at which rule was successfully verbalised.

10.5 Results

10.5.1 Performance

In a two-way ANOVA with test performance as dependent variable and response and difficulty as independent variables, there was a significant main effect of difficulty ($F_{2,43} = 3.582$, $p<0.05$). This indicates that Ss in the three difficulty conditions were performing differently. However, in Figure 10.5.1.1 this main effect of difficulty is not immediately apparent.

Several t-tests were conducted to get a more meaningful analysis of the data. The t-tests revealed that the only significant difference in means was between the hard and the easy condition in the active group (t=-3.617, df=21, p<0.01). Active Ss in the easy condition were performing significantly better than Ss in the hard condition. All other Ss were performing at about the same level.

There is no significant effect of response, suggesting that test-phase performance did not differ depending on whether Ss had to indicate actively that the string was ruleful or were merely presented the string and then told whether it was ruleful. This can also be seen in Figure 10.10.

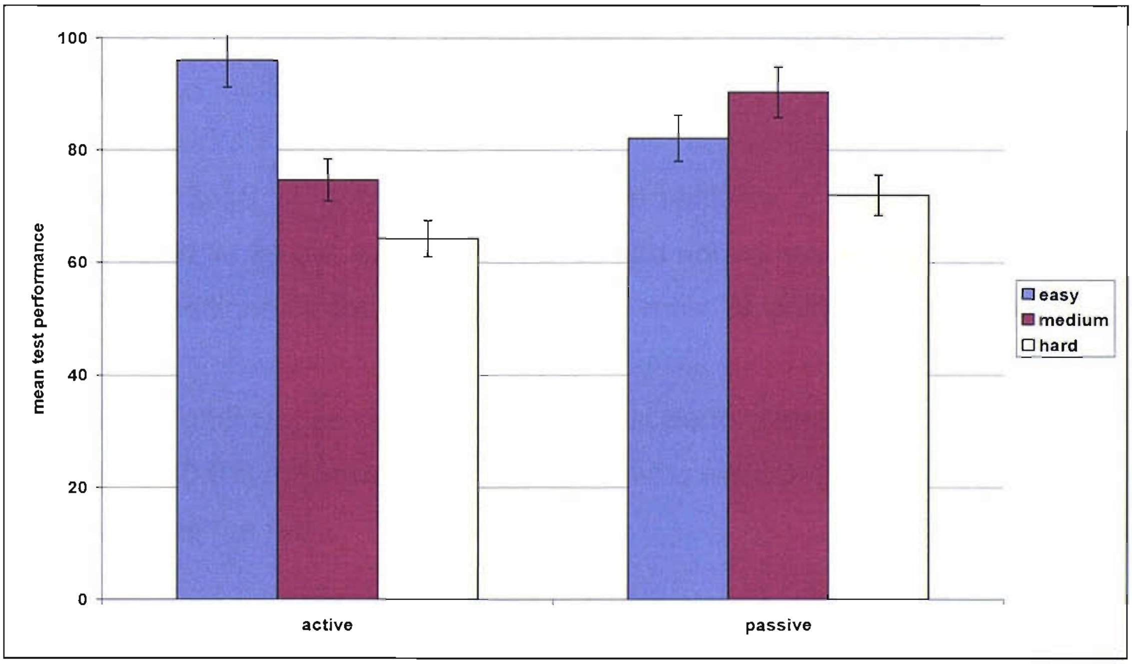


Figure 10.10: Mean Percentage of correct answers in test phase for the three difficulty conditions for the active and the passive group

10.5.2 Verbalisation

Figure 10.11 shows the number of verbalisers in the active and passive group for the easy, medium and hard difficulty condition.

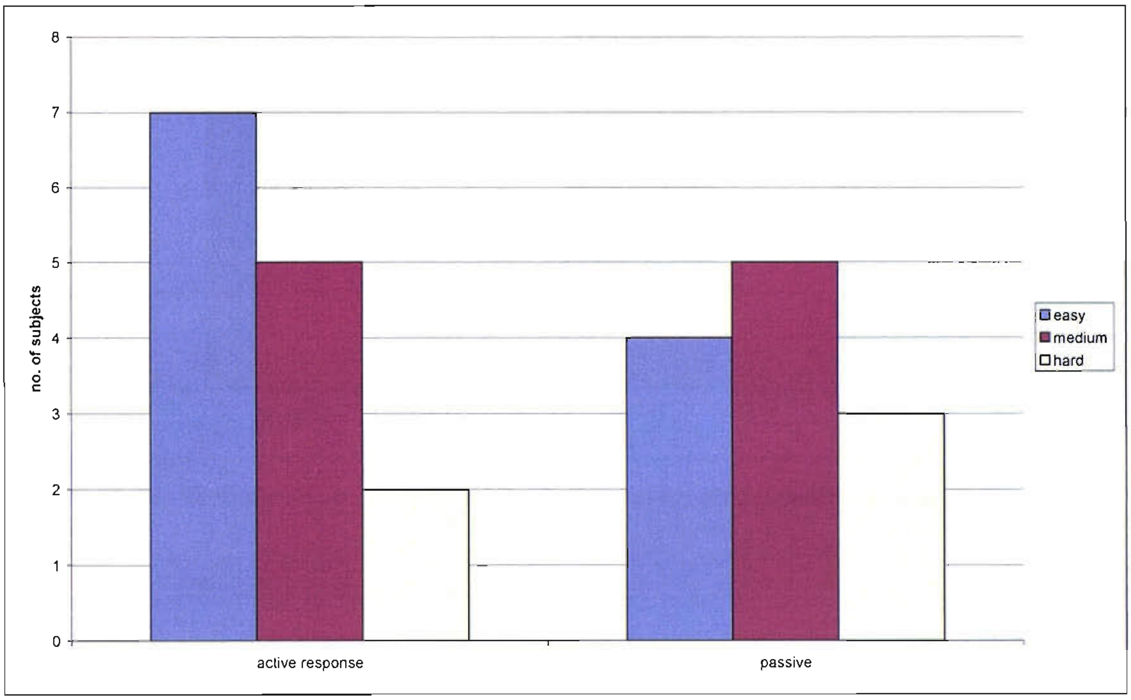


Figure 10.11 Number of Ss in the active and passive group who could verbalise the rules for the three difficulty conditions

A two-way ANOVA with verbalisation as the dependent variable and response and difficulty as independent variables indicated a significant main effect of difficulty ($F_{2,43} = 3.698$, $p<0.05$).

Figures 10.12, 10.13, and 10.14 show the percentages of active and passive Ss who could verbalise, partially verbalise and not verbalise, for the three difficulty conditions. In the passive condition most Ss could verbalise the rules.

There is a significant correlation between test performance and verbalisation ($r=0.922$, $p<0.01$) for passive Ss. Those Ss who were performing well could also verbalise the rule.

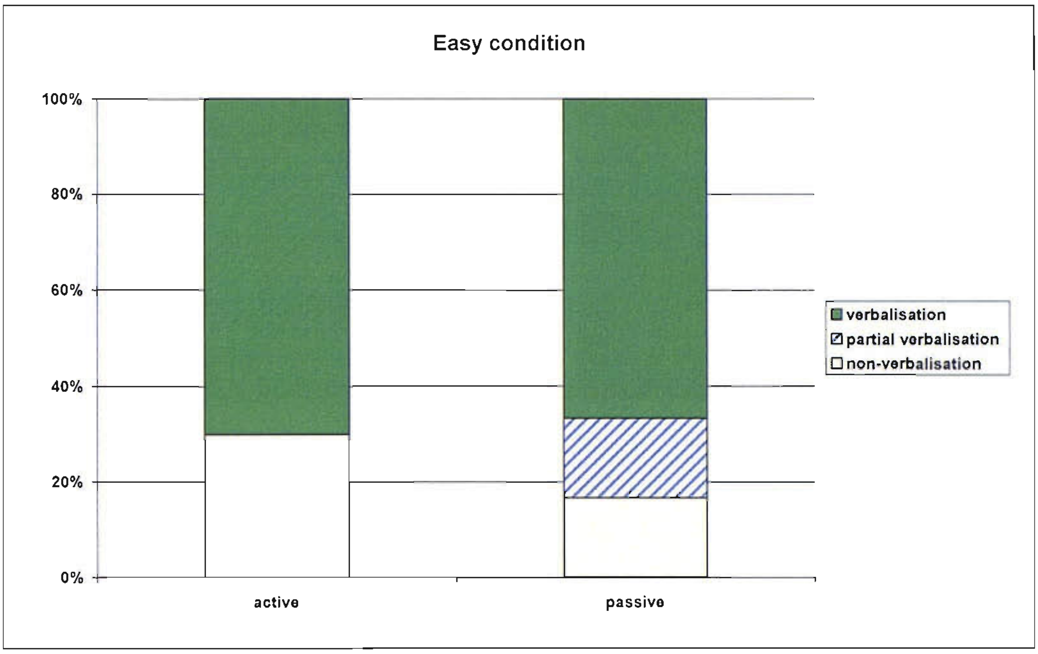


Figure 10.12: Verbalisation abilities of Ss in the active and passive group in the easy condition

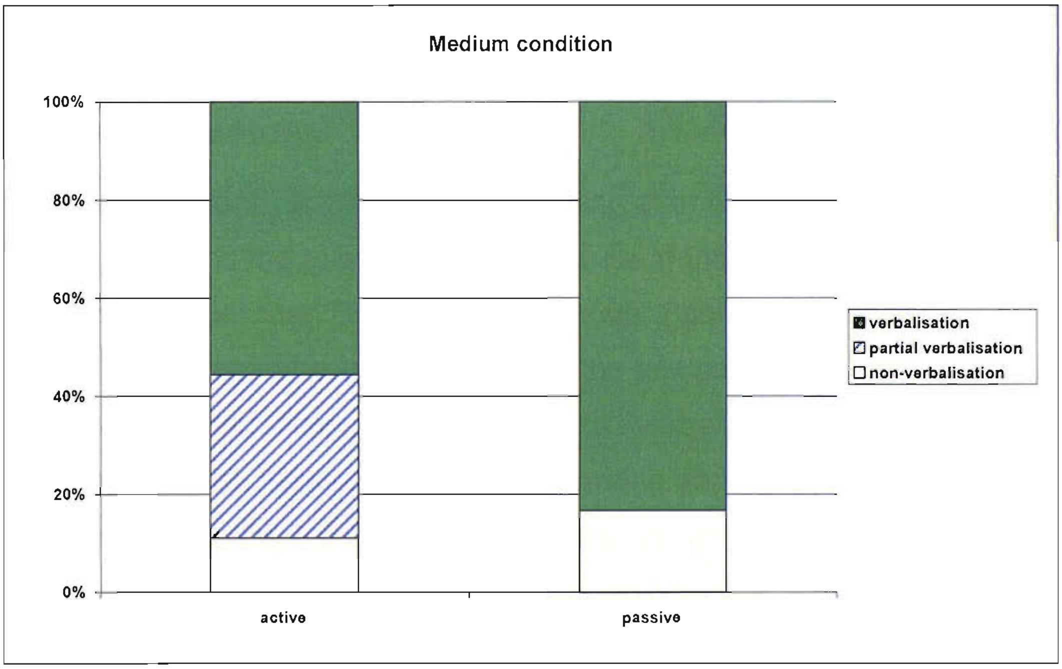


Figure10.13.: Verbalisation abilities of Ss in the active and passive group in the medium condition

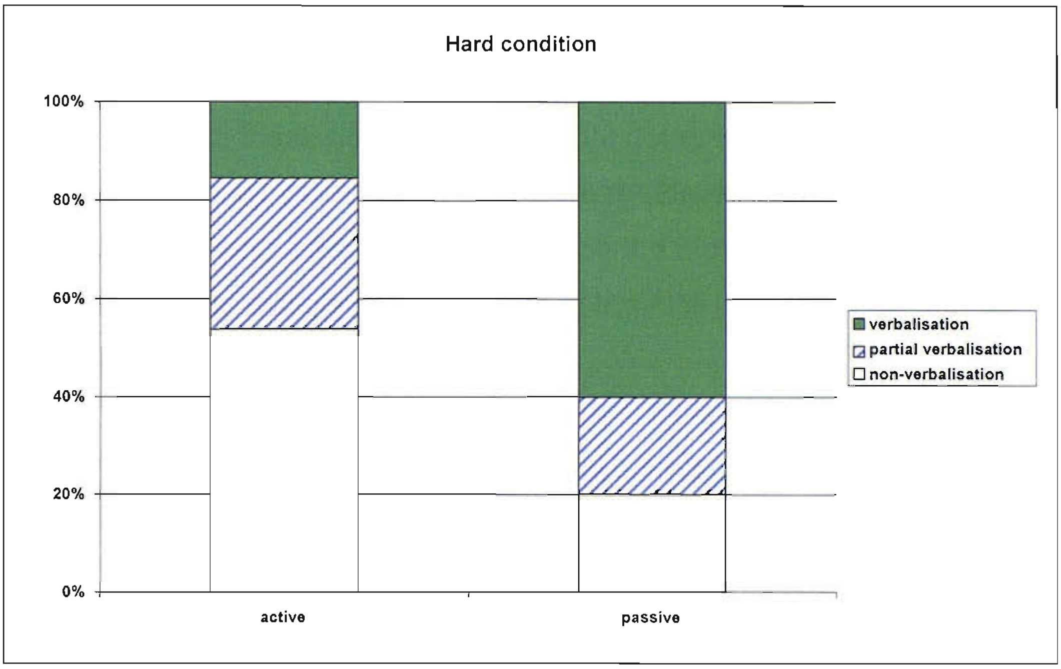


Figure 10.14: Verbalisation abilities of Ss in the active and passive group in the hard condition

10.6 Conclusions

In this experiment, Ss who were performing at or near 100% correct were also able to verbalise the rule(s). Interestingly, Ss in the passive group were performing better than Ss in the active group. Passive Ss not only had a higher percentage of correct answers in the test phase, but there were also more passive verbalisers (i.e. Ss who could verbalise the rules). Given the close association between success at learning and success at verbalisation (as also seen in the feedback experiment in Chapter 9), it is not surprising that both performance and verbalisation rates were higher in the passive group.

The hypothesis that those Ss who were responding actively to the stimulus would perform better than Ss who were not responding actively must therefore be rejected. One possible explanation is that the repetition of the stimulus (saying the syllables out loud while clicking a mouse button), which was designed to ensure that Ss paid attention to the stimulus, acted as a kind of "active response" to the stimulus. Ss in the passive group were in this sense not really passive, since they were saying the syllables aloud and clicking on the mouse button, unlike Ss in a classical “mere exposure experiment” (where Ss are merely exposed to the stimulus) (e.g. Zajonc, 1968; Newell & Bright, 2001). Although these passive Ss were not doing active trial-and-error hypothesis-testing, the fact that they were required to pay attention to the stimulus (by saying it aloud and clicking on the mouse button) may have been active enough to enable good learning of the rule (perhaps with mental hypothesis-testing).

Since Ss in the passive group did not have to make any decision about the rulefulness of the stimulus before they were told whether the stimulus was ruleful, the memory load on them may have been smaller, helping them perform better. It was noted in Section 5.4 that Ss in Dienes et al's (1991) study with an increased memory load found it difficult to remember which training stimuli had been presented in green and which in red, and this caused their performance to drop. In the experiment documented in Chapter 6 the increased memory load also showed a worse performance level. Similarly, Ss in the passive condition may have had smaller memory loads as they did not need to do explicit hypothesis-testing. Active Ss had to remember both whether or not their response had been correct, and whether or not their hypothesis was still valid. This may explain why Ss in the passive group were performing better than Ss in the active group. Or the difference may have been because the overt repetition of the stimulus together with the explicit hypothesis-testing was a handicap rather than an asset, compared to overt repetition with only mental hypothesis-testing.

As in the feedback experiment in Chapter 9, Ss who were performing well could also verbalise the rule(s). One S in the passive group thought he might have initially learnt the rule implicitly (in the learning phase), although later this S could explicitly state the correct rule and had a high performance rate. Since Ss in the passive group were not making any responses in the learning phase we have no further information apart from what S wrote down. Like in the feedback experiments in Chapter 9, those Ss who were performing at or close to 100% in the test phase could also explicitly name the rule(s).

11 Effect of Instructions on Learning

This experiment investigated the effect of training instructions on performance in the test phase. It is expected that telling Ss that there is a rule to be learnt (*forewarned* condition) will be beneficial to their performance, and telling them the actual rules (*forearmed* condition) will benefit them even more, while Ss who are not told that there is a rule (*rule-blind* condition) will have the worst test performance.

11.1 Method

10.1.1 Design

The design of this experiment is a 3 x 3 design with difficulty (easy, medium, and hard) and instructions (forewarned, forearmed and rule-blind group) as independent variables and task performance and verbalisation ability as dependent variables. Since these were web-based experiments, Ss were asked to write down whether they knew any rule(s), what rule(s) they thought they knew, any strategies they were using, and any other comments they had after every 20 trials and at the end of the experiment. These written comments were then analysed as the verbalisation data (see also Section 8.1.4 and Figure 8.3 for a screenshot).

11.1.2 Subjects

67 Ss took part in this experiment. There were 32 Ss in the *forewarned* group (those told in advance that there were underlying rules to be learnt), 9 Ss in the *forearmed* group (the ones explicitly told in advance what the rules were), and 26 Ss in the *rule-blind* group (neither told in advance that there were rules nor what the rules were). In the forewarned group, 10 Ss were in the easy rule condition, 9 in the medium, and 13 in the hard condition. In the forearmed group, there were three Ss in each difficulty condition. Ss in the forearmed condition were expected to perform at 100%

correct, as they were told the rule explicitly, so three Ss for each difficulty condition was deemed sufficient. In the rule-blind group, there were 9 in the easy condition, 8 in the medium, and 9 in the hard.

11.1.3 Procedure

The forewarned condition is the same as the condition with feedback (see Section 9.1.2) and active response (see Section 10.1.2). Ss were told that they should try to learn the rule(s) governing the strings. They were 80 learning strings, 40 positive and 40 negative. On each trial they were shown a string of syllables and asked to click on the right mouse button while saying each successive syllable aloud. Each click on the right mouse button made the corresponding syllable appear in a box beneath it, to be said aloud. After completing the entire string, they were asked whether or not they thought it had been ruleful. They responded by clicking the 'Yes' or 'No' button. After responding they received feedback on whether or not their answer had been correct. Then the next trial appeared.

In the forearmed condition, Ss were told in advance what the rules were. For 80 learning trials, Ss were shown positive and negative strings. On each trial they were presented with a string, which they had to repeat aloud while clicking the right mouse button. Once they had repeated the string, Ss were asked whether they thought the string was ruleful by responding with Yes or No. They were given feedback on whether their answer was right or wrong. Ss in this condition were asked to respond as quickly and as accurately as possible.

In the rule-blind condition, Ss were not informed about the existence of rules. They were told only that they would see strings of syllables from a foreign language and that they were being tested on how fast they could reproduce them under different conditions. They were presented with positive and negative strings shown on a different colour background (counterbalanced across Ss) to differentiate the two types of string. Ss were required to perform different tasks with the two colours. One task was to click the right mouse button and say the syllables out loud. The other task was to say the syllables aloud as soon as they appeared (automatically) in the boxes underneath the stimulus. Once they had responded, the next trial appeared. After the training phase Ss in the rule-blind condition were told that the strings on one colour background all followed a particular rule, while the other strings (on the other colour background) did not.

The test phase was the same for Ss in forewarned, forearmed and rule-blind conditions. 40 novel strings were presented on a white background and Ss were asked to indicate by clicking on Yes or No whether or not they thought the string followed the rule underlying the coloured presentations in the training phase. They were not given feedback on their responses in the test phase.

All Ss in the forewarned condition were asked to write down whether they knew any rule(s), what rule(s) they thought they knew, any strategies they were using, and any other comments they had after every 20 trials and at the end of the experiment. Since Ss in the rule-blind condition were not told about the existence of rules, they were simply asked after every 20 trials to write down any strategies they were using to help them with their task and any other comments they had. Subjects in the forearmed condition were asked to write down any comments they had.

11.2 Easy rule

11.2.1 Forewarned condition

The forewarned condition is the same as the feedback condition (see Section 9.2.1) and the active response condition (see Section 10.2.1). The results are shown here again.

10 Ss took part in this condition, 8 male, 2 female. The average performance in the test phase was 89.25%.

In Figure 11.1 (=Figure 9.1) the learning curves of all Ss in the feedback group in the easy condition are shown with the average learning curve in boldface. Most Ss could correctly distinguish ruleful from unruleful strings after about 30 trials, although most were also still making some mistakes in later trials, probably due to fatigue or lack of concentration.

Seven of the ten Ss could describe the rule after the maximum 40 trials. As soon as these Ss learnt the rule, performance improved from chance to near perfect, and they verbalised the rule at first opportunity (i.e. in the next break). Of the remaining three Ss, one did not write anything down, but was performing at a high level of correctness; a further S only verbalised incorrect rules, and his overall performance in the test phase was only 55%.

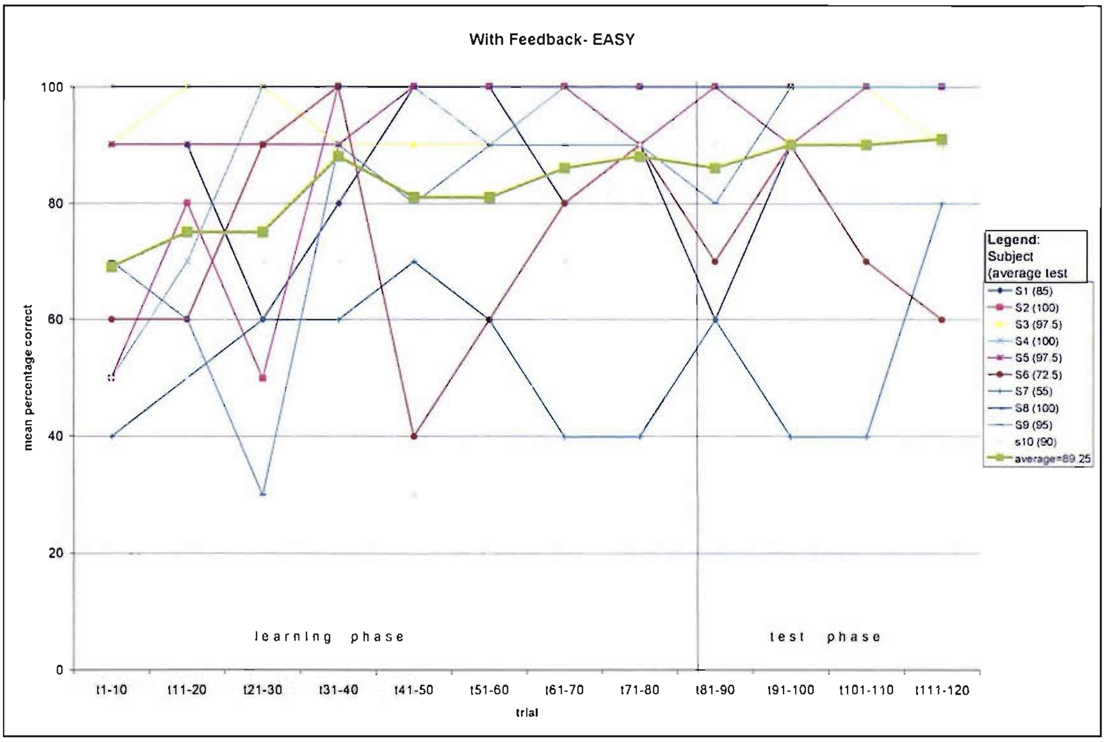


Figure 11.1: Learning curves for Ss in the feedback group in the easy condition (=Figure 9.1)

A further S stated the correct rule in combination with an incorrect rule after 20 test trials (so after completing the 80 learning trials and 20 of the 40 test trials), but is performing at a very high level (90% correct) from learning trial 60 onwards. This S may have learnt to respond correctly (perhaps also implicitly) at an earlier stage than his ability to verbalise the correct rule. However, it is also possible that this S did not write down the rule until completely certain of it. Ss were asked to write down any rules they were using during the experiment, but were not explicitly asked from which trial onwards they were using these rules from.

11.2.2 Forearmed condition

3 Ss took part in this experiment: 2 male and 1 female. The average test performance of Ss in the test phase was 99.17% correct. All Ss were performing at top level, since they had been told the rule in advance.

This condition also serves as a control condition and shows that Ss can learn the rules and perform at 100% when explicitly told the rules in advance. It also shows that although Ss know the rules and can perform at 100%, mistakes are still evident and can be attributed to lapses in concentration and fatigue. Ss stated that they made mistakes because they were not paying sufficient attention and trying to finish quicker. Lapses in performance by Ss who have stated the rules and have been performing at 100% in earlier trials, can thus be ascribed to lack of concentration, fatigue and a wish to finish the experiment once they had achieved the aim of the experiment (figuring out the rule(s)).

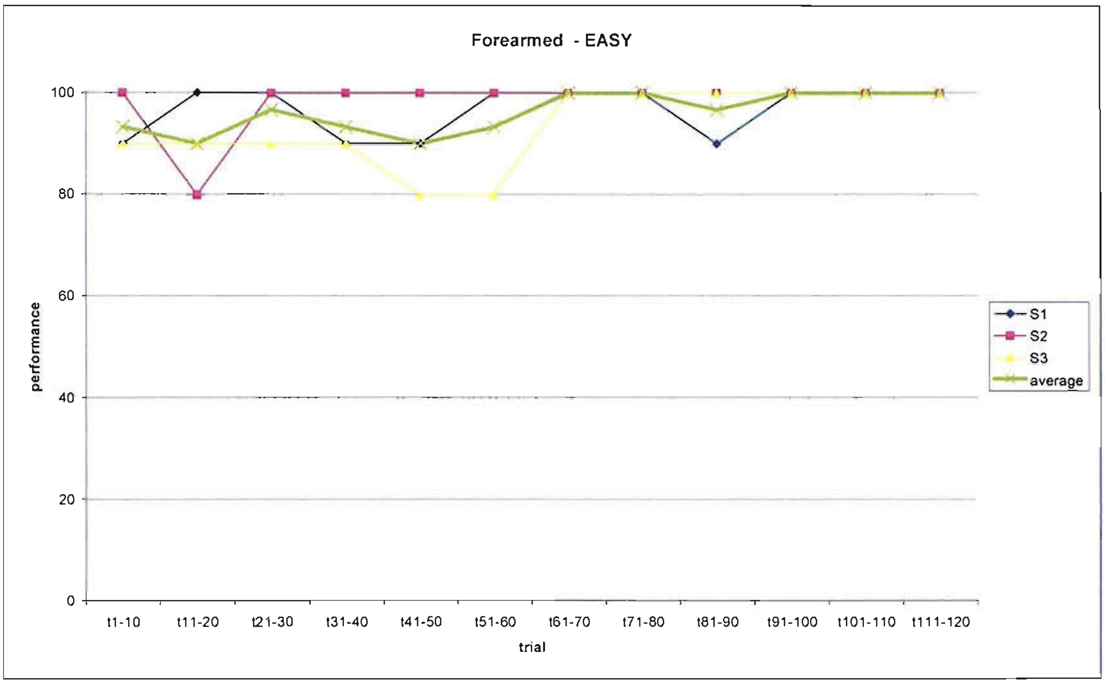


Figure 11.2 Learning curves of Ss in the forearmed group in the easy condition

11.2.3 Rule-blind condition

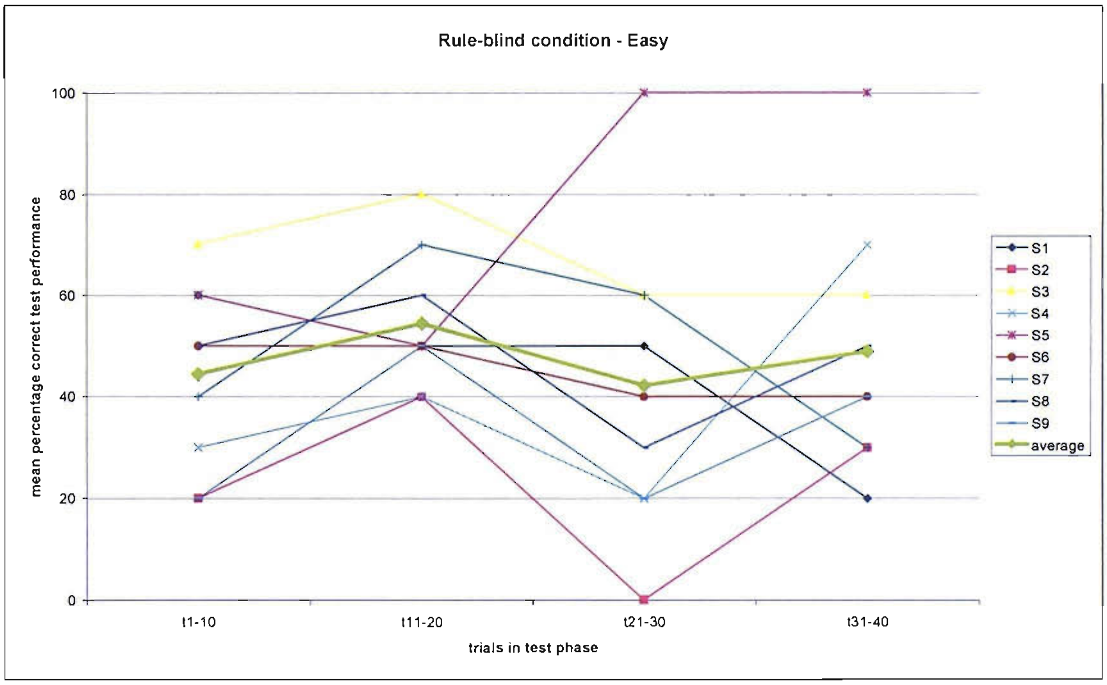


Figure 11.3 Learning curves of Ss in the rule-blind group in the easy condition

Nine 9 Ss took part in this experiment: 4 male, 5 female. Ss in the rule-blind condition were performing at an average level of 47.5% correct in the test phase.

Only one subject explicitly stated the correct rule and was performing at 100% correct: S12 explicitly stated the correct rule after 20 test trials, and was performing at 100% after that. S10 explicitly named the rule after 40 test trials (i.e. at the end of the experiment) and was performing at 70% in the last ten trials. S9 stated the correct rule after 60 learning trials. This S verbalised the rule without knowing that there were any rules to be learnt. However, this S's average score in the test phase was 67.5% correct, which is above chance but not at top level. This suggests that this subject did not realise that their hypothesis was correct, and was entertaining and using other hypotheses in the test phase, rather than the correct one.

11.2.4 Verbalisation

Figure 11.4 shows the learning curves of verbalisers in the forewarned group with continuous lines, the forearmed group with dotted lines, and the rule-blind group with dashed lines. The average learning curves are shown in boldface. The verbalisation times are shown with highlighted arrows for each S. Ss in the forewarned condition verbalised the rule early in the experiment and this is indicated by yellow highlighted arrows in the figure. Several Ss in the rule-blind condition could also verbalise the rules, as indicated by orange highlighted dashed arrows in the figure.

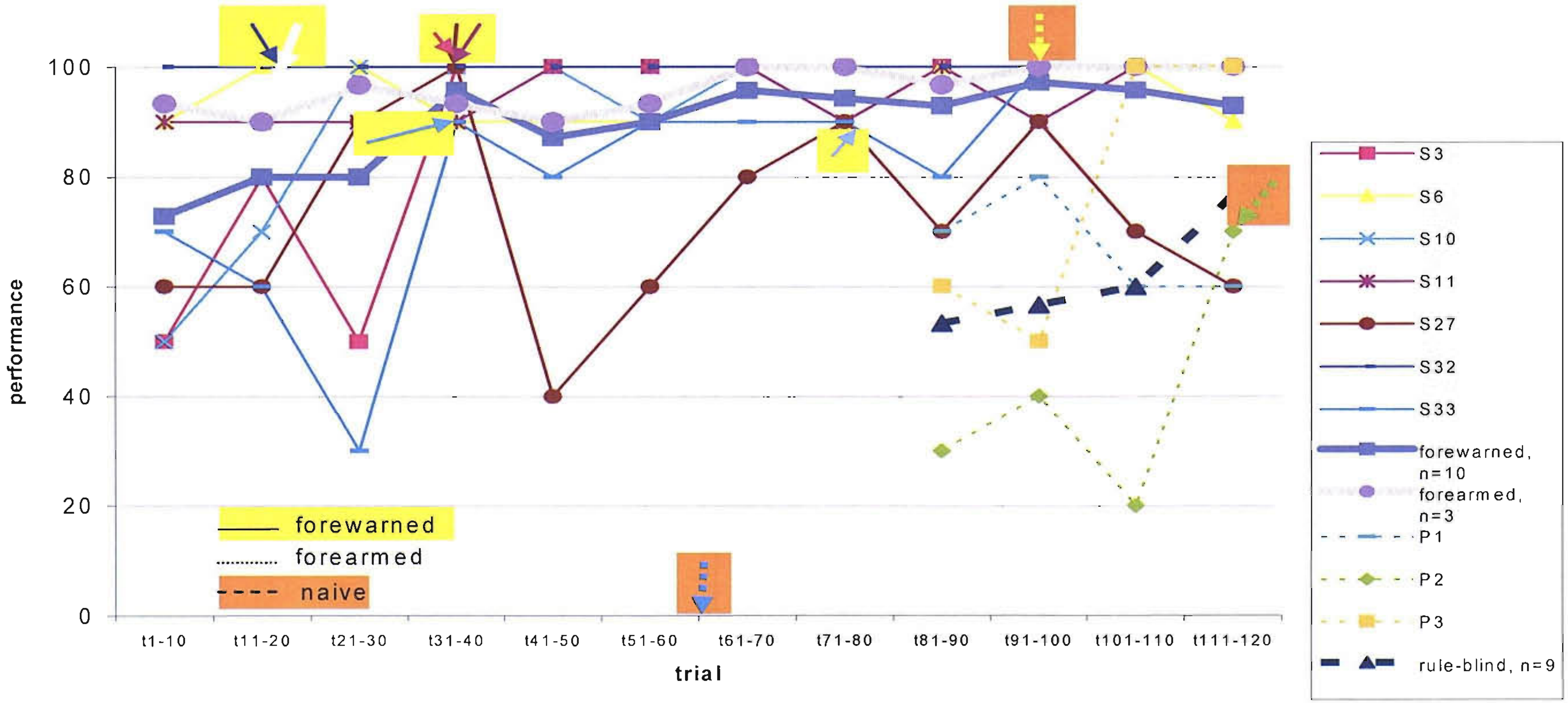


Figure 11.4: Easy condition: Learning curves of verbalisers in the forewarned, forearmed and rule-blind group with the point of rule-verbalisation indicated.

11.3 Medium rule

11.3.1 Forewarned condition

The forewarned condition is the same as the feedback condition in Chapter 9 and the active response condition in Chapter 10. Nine Ss took part in this condition, 4 male, and 5 female. Ss were performing at an overall average of 80.28% correct in the test phase.

Figure 11.5 (=Figure 9.4) shows the learning curves of the Ss with feedback in the medium difficulty condition, with their average in boldface. Most Ss seemed to learn this rule more gradually than the easy rule, first noticing the repeated syllables, and then later noticing the relevance of the position in the string. These Ss were performing at levels above chance when they had learnt the partial rule (repeated syllable), and their performance improved to near 100% correct when they had learnt the additional rule (position relevance). Some Ss only learned partial rules; for example, they noticed that it had something to do with the repeated syllable, but could not figure out the relevance of the position. These partial learners were performing at levels better than chance, but did not reach perfect performance.

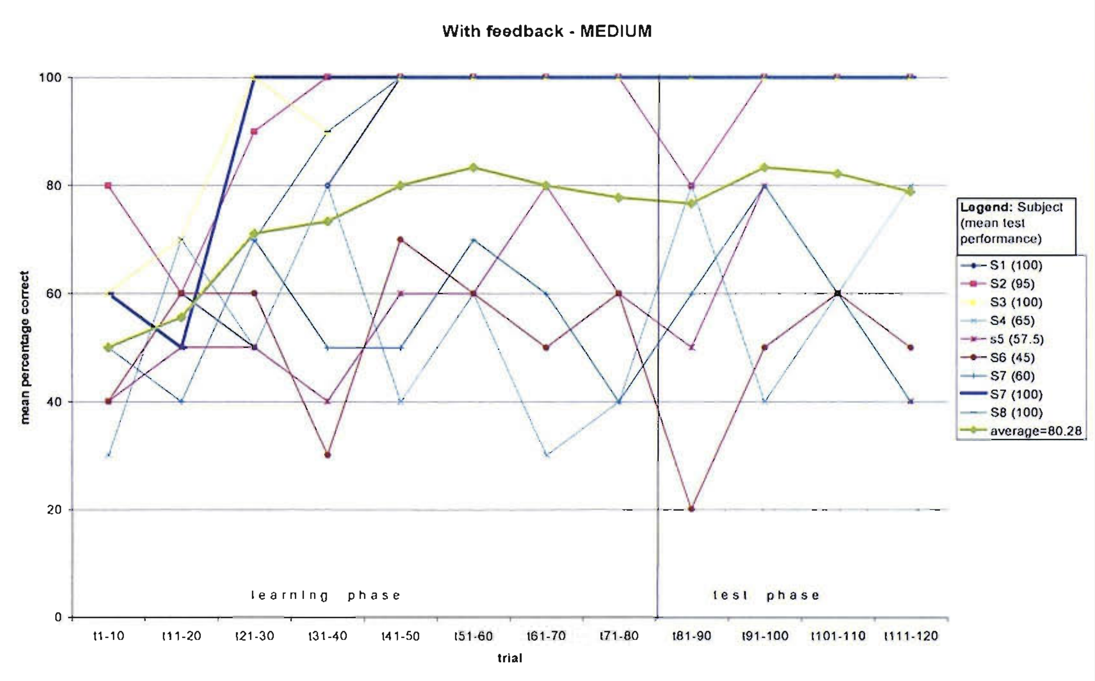


Figure 11.5 Learning curves of Ss in the feedback group in medium difficulty condition (=Figure 9.4)

11.3.2 Forearmed condition

3 Ss took part in this experiment: 2 male, 1 female. Average test performance of Ss in the test phase when they were told the rules in advance was 98.33% correct in the medium difficulty condition.

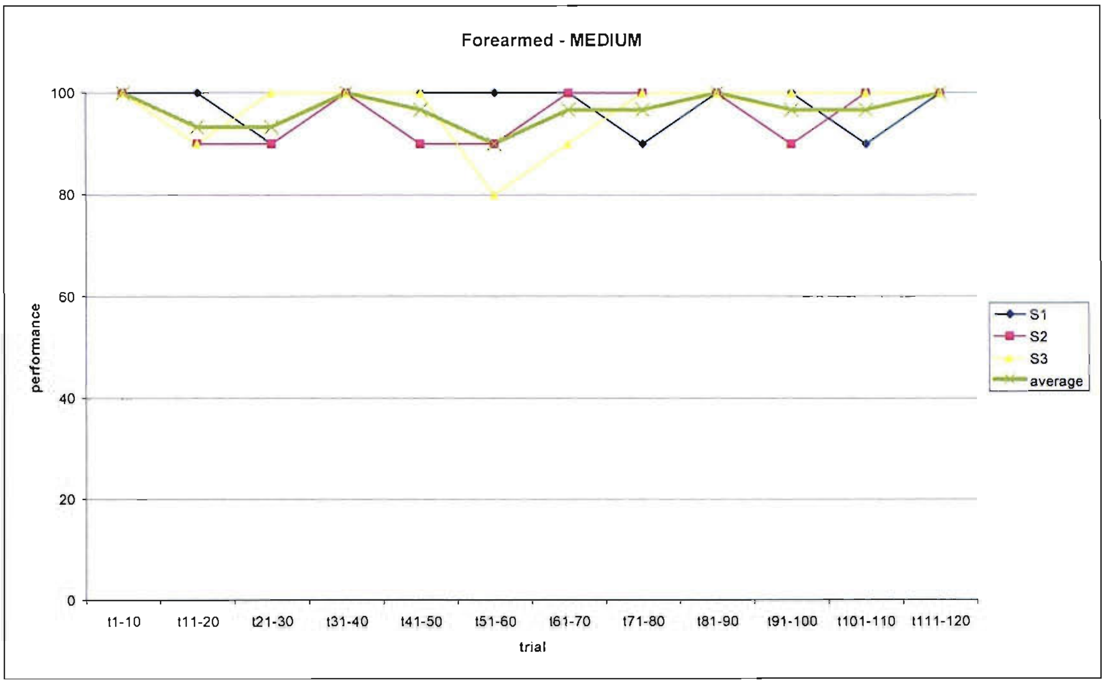


Figure 11.6 Learning curves of Ss in the forearmed group in the medium difficulty condition

Figure 11.6 shows that all Ss in the forearmed group were performing at or close to 100% correctly.

11.3.3 Rule-blind condition

8 Ss took part in this experiment: 6 male, 2 female. Ss in the medium rule-blind condition were performing at an average level of 53.44% correct in the test phase.

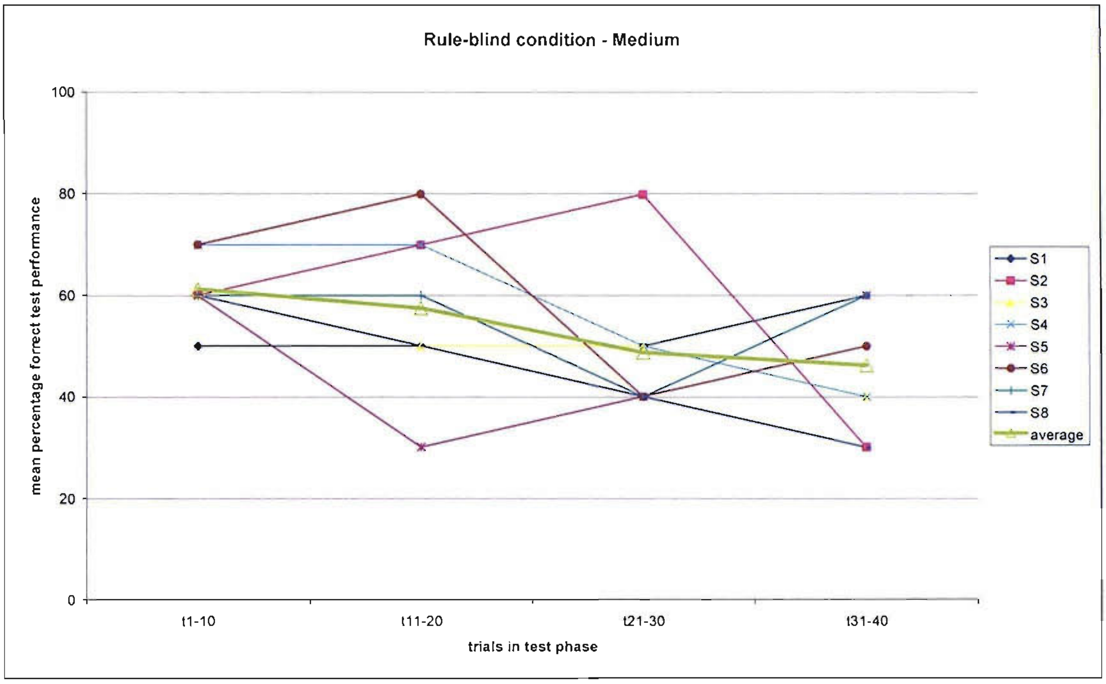


Figure 11.7 Learning curves of Ss in the rule-blind group in the medium difficulty condition

Figure 11.7 shows the learning curves of Ss in the rule-blind group. None could explicitly state the rule. One S stated a partial rule at the very end of the experiment ("When the last syllable of the first part of the string is the same as the first syllable of the second part of the string") and was performing at an average level of 60%. All others are performing at or close to chance.

11.3.4 Verbalisation

Figure 11.8 shows the learning curves of Ss in the forewarned group with continuous lines, Ss in the forearmed group with dotted lines, and Ss in the rule-blind group with dashed lines, with their averages shown in boldface. The rule verbalisation times are shown with highlighted arrows. Verbalisers in the forewarned group are shown with yellow highlighted arrows; all verbalised the rules within the first 40 learning trials. There were no verbalisers in the rule-blind group. (Ss in the forearmed group were told the rule in advance, so knew the rule from the beginning.)

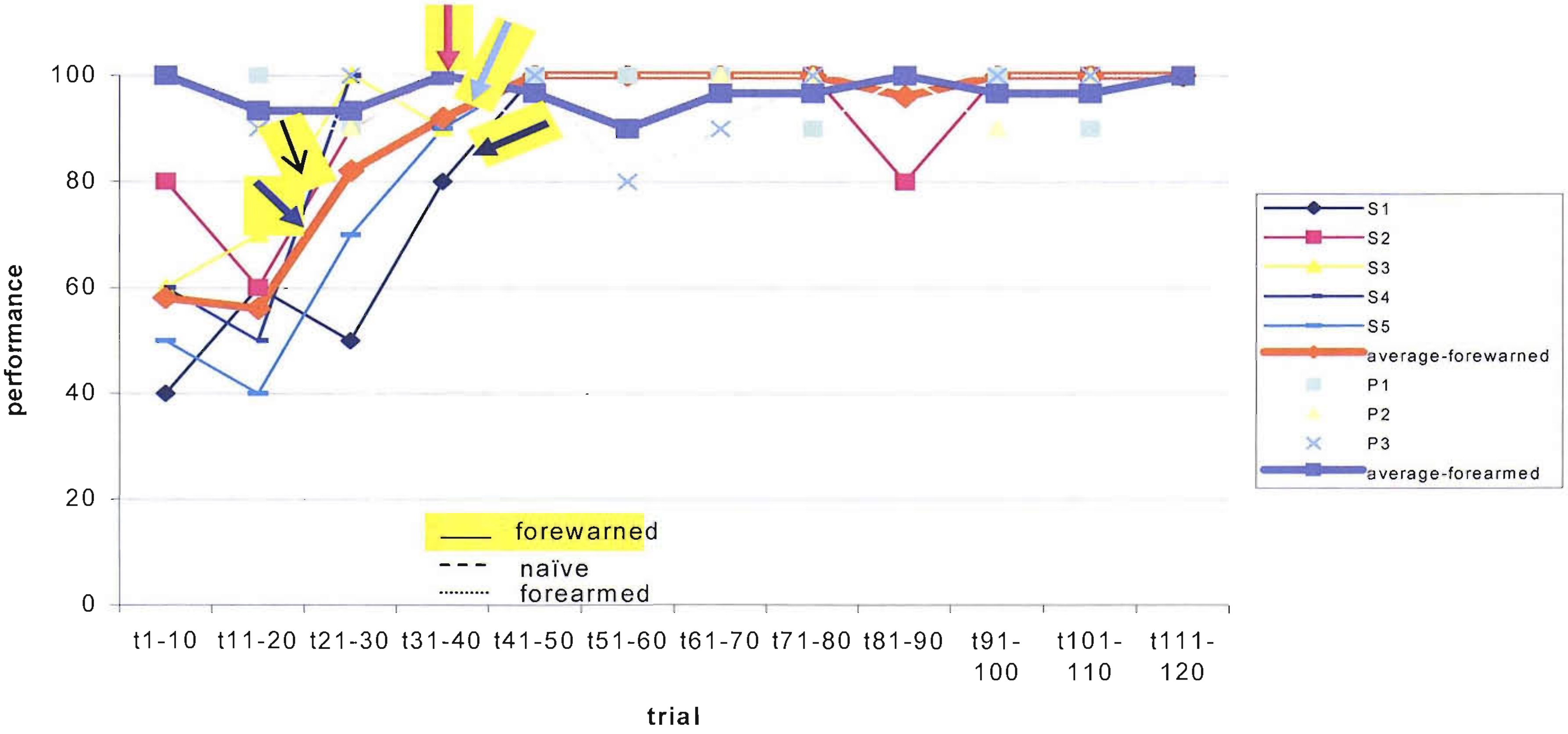


Figure 11.8 Medium difficulty condition: Learning curves of verbalisers in the forewarned (continuous lines), forearmed (dotted lines), and rule-blind (dashed lines) groups, with verbalisation.

11.4 Hard rule

11.4.1 Forewarned condition

The forewarned condition is the same as the feedback condition in Section 9.4.1 and the active response in Section 10.4.1, so the experiment was not repeated. The data from the experiment was the same.

Thirteen Ss took part in this condition, 9 male, 4 female. Overall test performance was 63.46% correct.

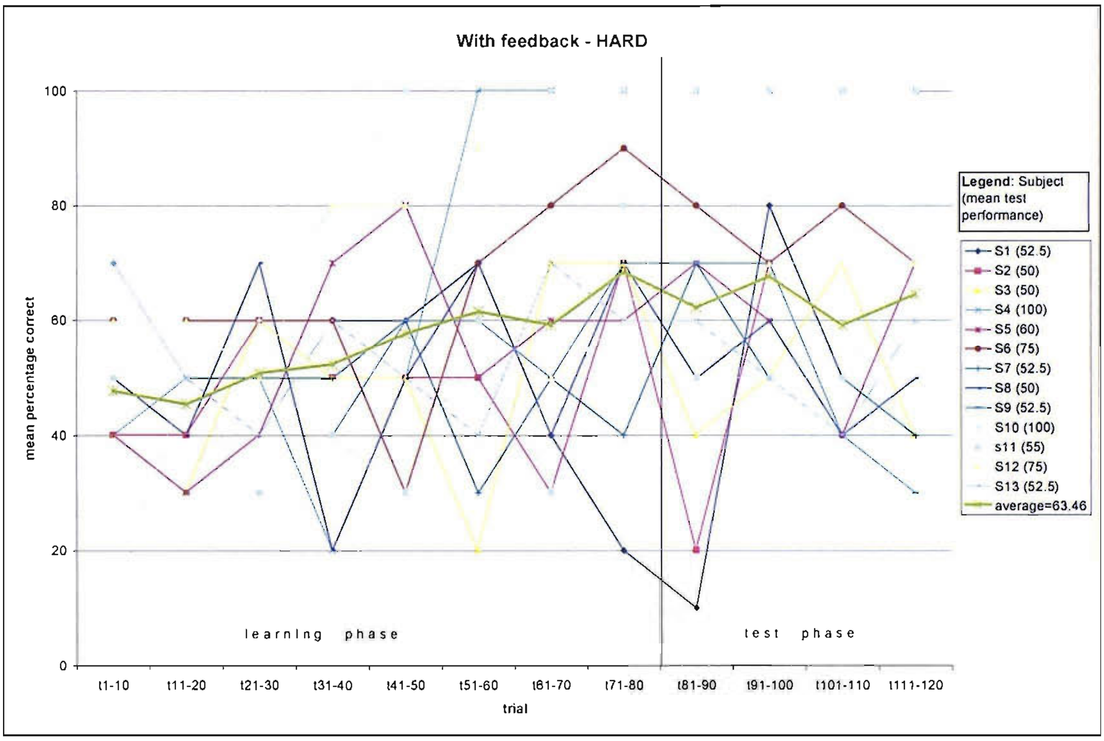


Figure 11.9 Learning curves of Ss in feedback group in the hard difficulty condition (=Figure 9.7)

Figure 11.9 shows the learning curves of the Ss with feedback in the hard condition. The boldface line is the average learning curve. Two Ss learned the rule explicitly and could describe it in words. Most Ss learned partial rules, for example "the same syllable may not repeat directly (like Sa Sa)". Half of the Ss could verbalise the correct rule, or partial rules, while half could not explicitly state the rule.

11.4.2 Forearmed condition

3 Ss took part in this experiment: 2 male, 1 female. Ss were performing at an average level of 98.33% in the test phase of the hard condition.

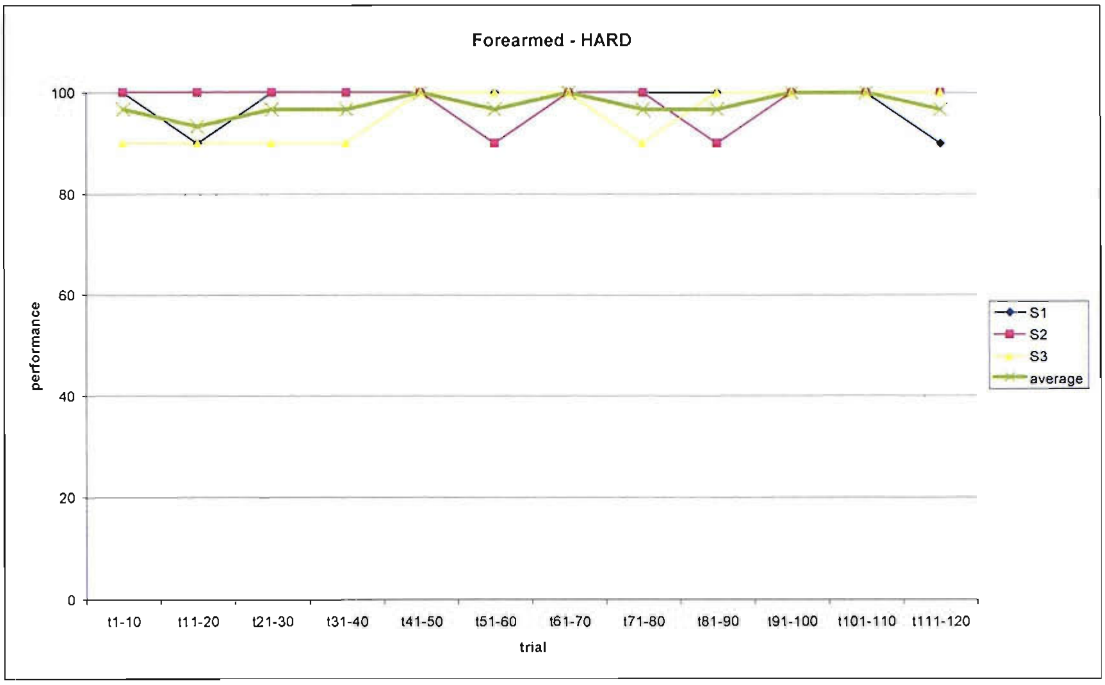


Figure 11.10 Learning curves of Ss in the forearmed group in the hard condition

Figure 11.10 shows the learning curves of Ss in the forearmed group. Ss were forearmed with the rule, so they are all performing at or very near 100% correct.

11.4.3 Rule-blind condition

9 Ss took part in this experiment: 7 male, 2 female. The average test performance of Ss was 58.89% correct.

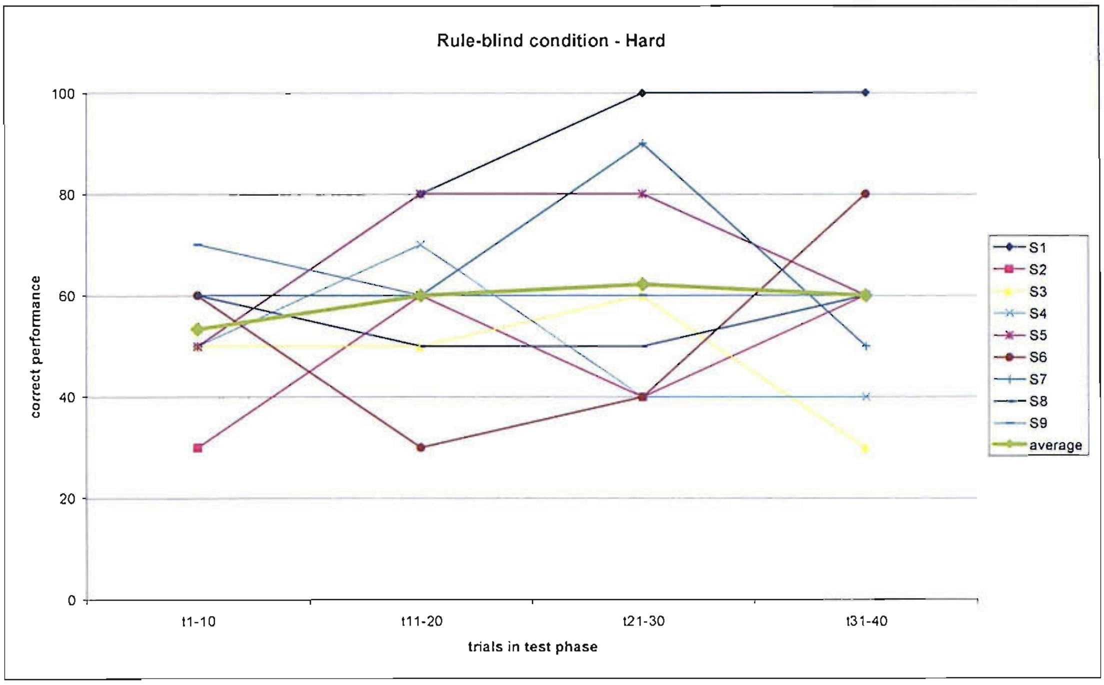


Figure 11.11 Learning curves of Ss in the rule-blind group in the hard condition

Two Ss could explicitly state the rule after 20 test trials, but only one S seemed to also be responding according to this rule, and performing at 100% in the last 20 test trials. The other S was performing at 90% correct in test trials 21-30, but performance dropped to 50% again in the last ten test trials. Three other Ss could name partial rules. Two of these did not have a test score better than chance (50% and 52.5% resp.), the third S was performing better than chance at 62.5%.

11.4.4 Verbalisation

In Figure 11.12 the learning curves of verbalisers in the forewarned group are shown with continuous lines, the learning curves of verbalisers in the forearmed group are shown with dotted lines, and the learning curves of verbalisers in the rule-blind group are shown with dashed lines. The rule verbalisation times are indicated for each verbaliser by highlighted arrows. Ss in the forewarned group are indicated by yellow highlighted arrows, and were verbalising the rule earlier in the experiment than Ss in the rule-blind group, which are indicated by orange highlighted arrows.

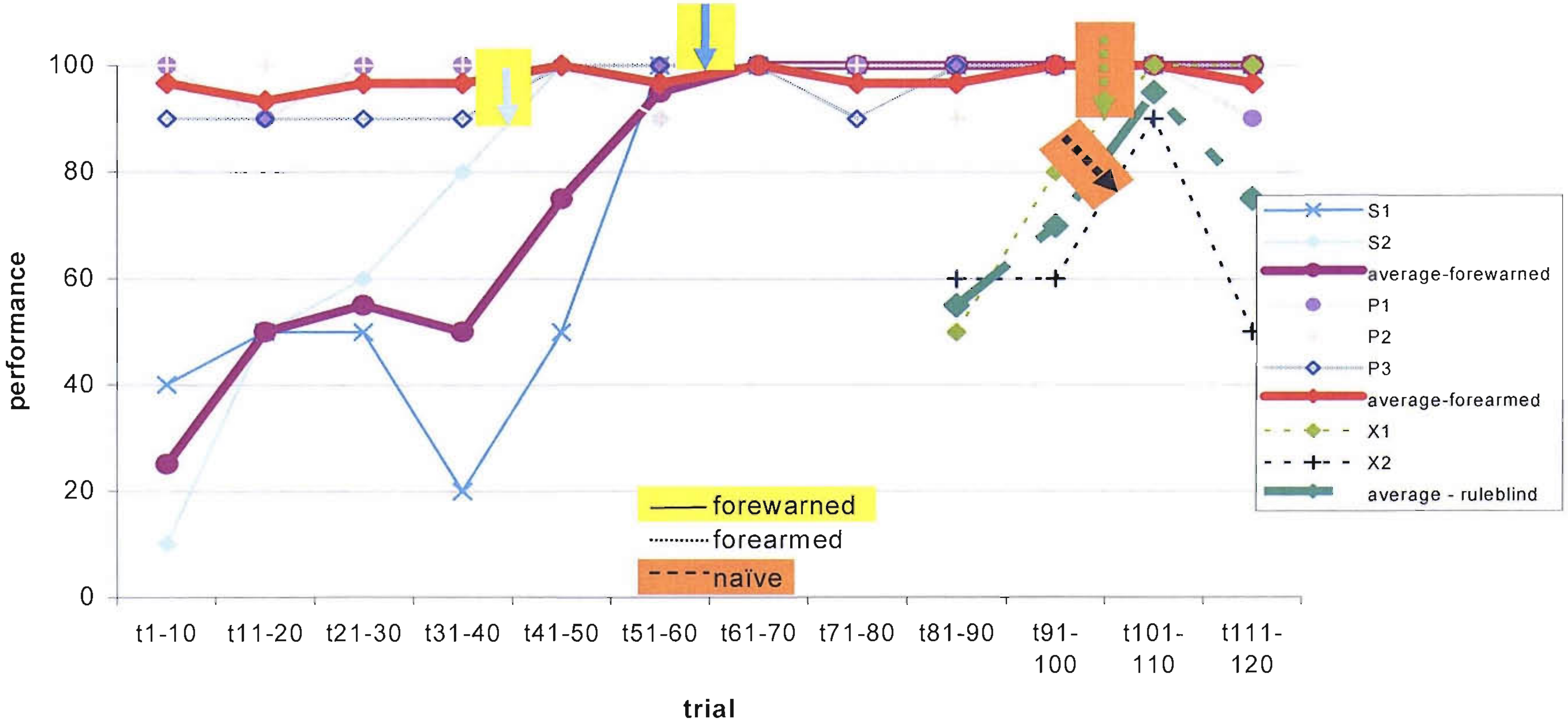


Figure 11.12 Learning curves of verbalisers in the forewarned (continuous lines), forearmed (dotted lines), and rule-blind (dashed lines) groups with verbalisation times in the hard condition

11.5 Results

11.5.1 Performance

A two-way ANOVA with test performance as dependent variable and instruction type and difficulty as independent variables revealed a significant main effect of instruction type ($F_{2,58}$=34.569, p<0.01) and a significant interaction between instruction type and difficulty ($F_{4,58}$=3.843, p<0.01). This indicates that Ss were performing differently depending on which type of instructions they received (i.e. forewarned about rules, forearmed with rules, or rule-blind condition).

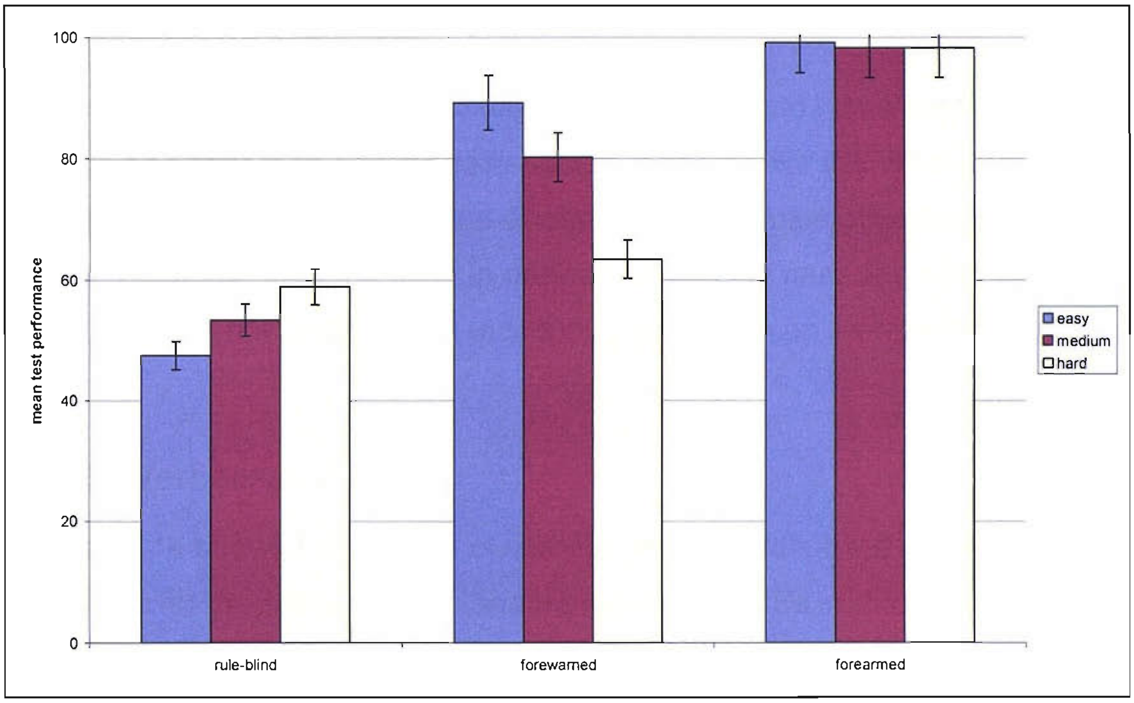


Figure 11.13 Mean test performance of Ss in all three difficulty conditions for each of the three instruction types

Figure 11.13 and Table 11.1 show that in the rule-blind condition Ss were performing at chance level. Ss who were forewarned about the existence of rules were performing better than chance and Ss who were forearmed with the actual rules were performing close to 100% correctly.

	Easy	medium	hard
rule-blind	47.5	53.44	58.89
forewarned	89.25	80.28	63.46
forearmed	99.17	98.33	98.33

Table 11.1 Mean percentages of correct test performance for each group and each difficulty condition

T-tests showed that there were no significant differences between the difficulty levels for the rule-blind condition, nor for the forearmed condition. There was a significant difference between the easy and hard difficulty level in the forewarned condition (t=-3.617, df=21, p<0.01). Ss in the easy condition were performing significantly better than Ss in the hard condition.

Interestingly, in the rule-blind condition, Ss with the hard stimuli seemed to be performing better than Ss with easy stimuli. It seems like the difficulty was reversed when Ss were not aware of any underlying rules governing the strings. However, the difference in performance in the hard and easy difficulty conditions was not significant, hence may just have been an effect of chance.

11.5.2 Verbalisation

Figure 11.14 shows the number of verbalisers in the rule-blind and forewarned conditions for the easy, medium and hard conditions. Ss in the forearmed condition were not asked to verbalise any rules, as they had been told the rules in advance. There are far fewer verbalisers in the rule-blind condition, although the few that there are, are surprising enough, since they had not been looking for rules.

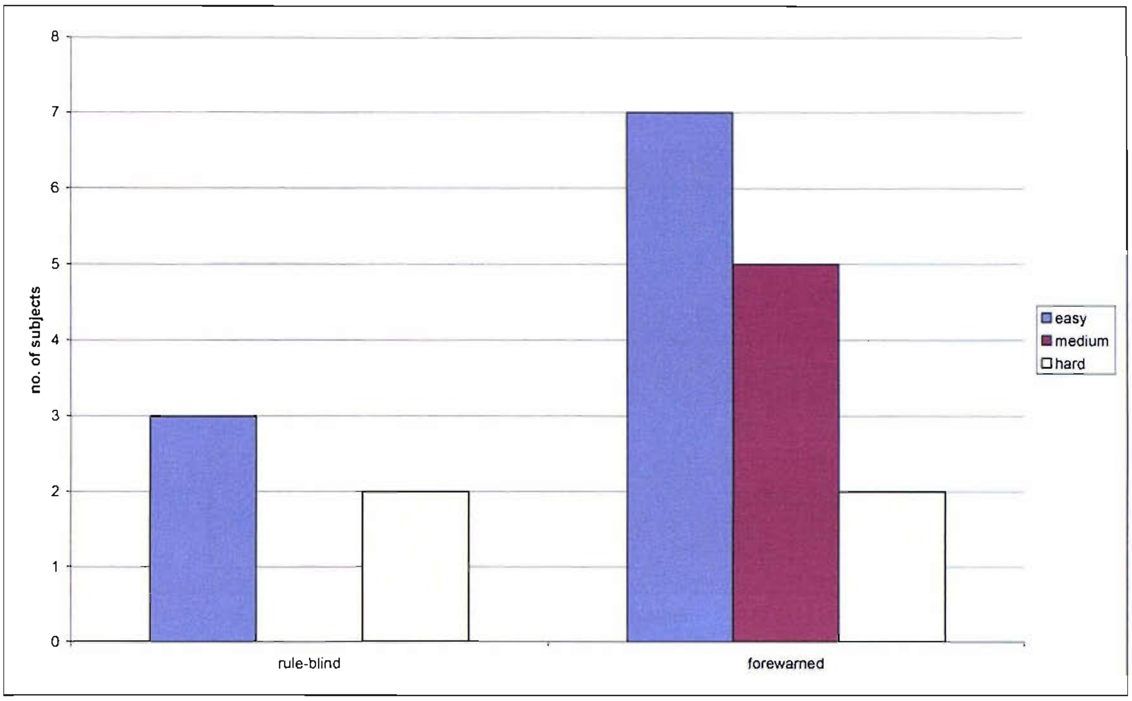


Figure 11.14 Number of Ss who could verbalise the rules in the rule-blind and forewarned condition for the three difficulty levels

A two-way ANOVA with verbalisation as dependent variable and instruction type and difficulty as independent variables showed a significant main effect of instruction type ($F_{1,52}$=7.852, p<0.01) and a significant interaction between instruction type and difficulty ($F_{2,52}$=4.535, p<0.05).

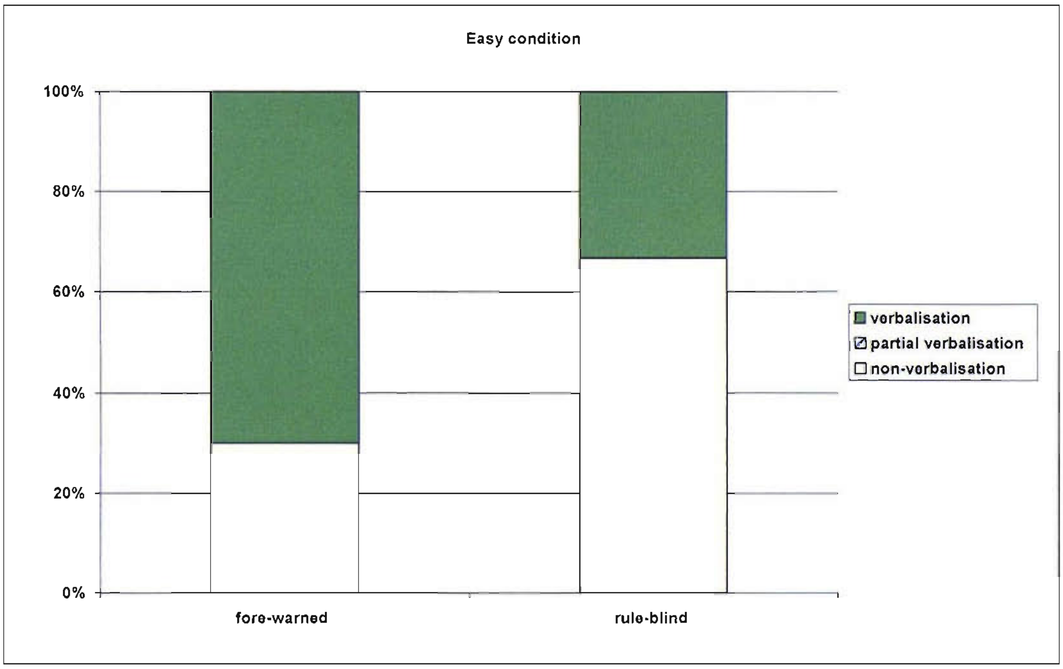


Figure 11.15: Verbalisation abilities of subjects in the fore-warned and rule-blind conditions in the easy condition

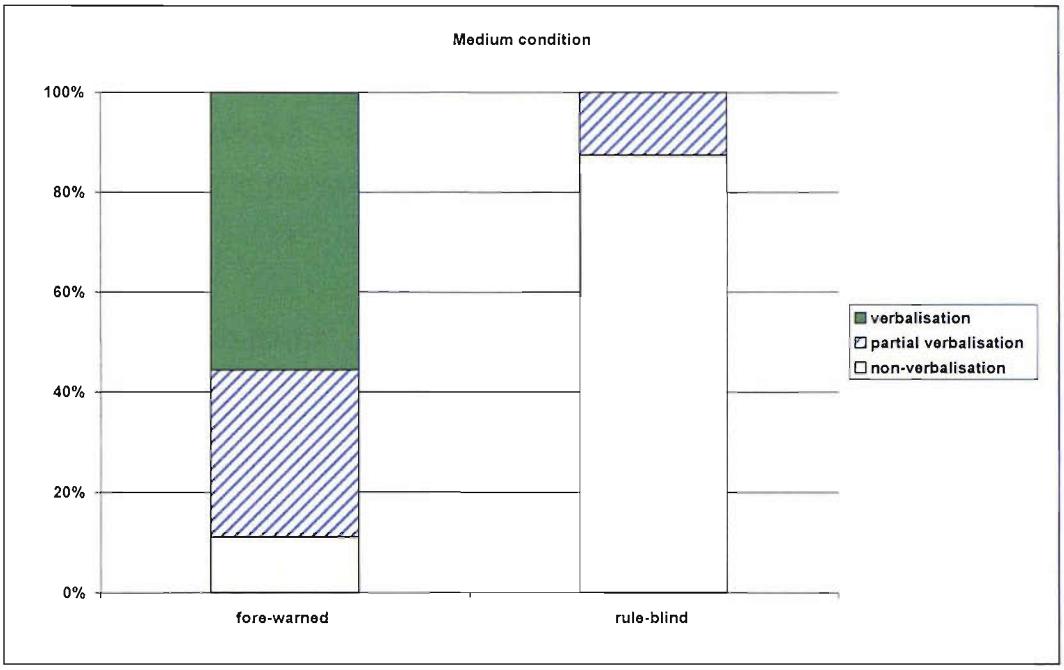


Figure 11.16: Verbalisation abilities of subjects in the fore-warned and rule-blind conditions in the medium condition

Figures 11.15, 11.16 and 11.17 show the percentages of fore-warned and rule-blind Ss who could verbalise, partially verbalise and not verbalise the rules. (Obviously, Ss who were told the rules in advance (forearmed condition) could verbalise the rules.)

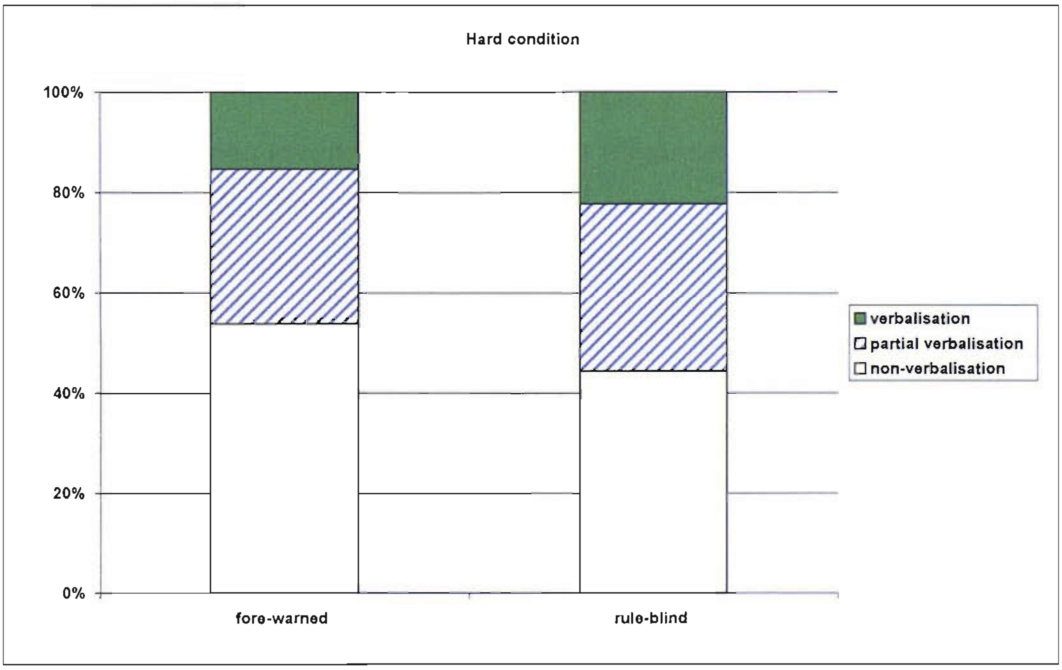


Figure 11.17: Verbalisation abilities of subjects in the fore-warned and rule-blind conditions in the hard condition

There is a significant positive correlation between test performance and verbalisation for both the rule-blind condition (r=0.555, df=24, p<0.01) and the forewarned condition (r=0.838, df=30, p<0.01). Those Ss who were performing well in the test phase could also verbalise the rules.

11.6 Conclusions

Several conclusions can be drawn from these data. This experiment confirms the previous finding that when Ss were performing at a high level, they could usually also verbalise the rules correctly. There was no evidence of implicit learning in this experiment.

Prior instruction and information have large effects on performance. Unsurprisingly, Ss forearmed with the rules in advance (forearmed condition) perform 100% correctly throughout the experiment. Ss forewarned that there are rules they should try to learn (forewarned condition) perform mostly correctly, at a level of around 70% correct in the test phase, while Ss who do not know about the existence of rules (rule-blind condition) perform at about chance level. There were no significant differences in performance for each difficulty level in the forearmed condition or in the rule-blind condition. This indicates, first, that there were no differences in difficulty in understanding and applying the rules when they are given in advance (forearmed condition). Second, when Ss did not know about the existence of rules, the easy rule (in this experiment) was not easier to learn than the hard rule.

12 Effect of Positive/Negative Instantiation on Learning

In this chapter the effect of positive and negative instantiation in training was investigated. Because they were originally inspired by natural language learning, most classical AGL experiments presented positive evidence only, i.e. only instances that followed the rule(s). In other areas of rule learning (e.g. concept learning, Bruner et al, 1956) as corroborated by work on formal learning theory (Gold, 1967), it has been found that to successfully learn an unknown rule, Ss need to sample both positive and negative instances of the rule, i.e. cases that obey and disobey the rule, because otherwise they have no way of knowing what rule(s) or feature(s) distinguish the positive from the negative ones. To illustrate, consider the following: Imagine Ss are presented with 10 instances of a red circle and asked to name the rule. The most likely rule would be "red circles". However, the rule could also be "circles", "red", or even "coloured shapes". If Ss are then presented with a blue circle, they have no way of knowing from their sample of red circles, whether the rule is indeed "red circles", in which case a blue circle would be a negative instance, or whether the rule is instead "circles", in which case a blue circle would follow the rule and be a positive instance.

In this experiment, one group of Ss received both positive and negative instances in the training phase. A second group received only positive instances, and a third group received only negative instances. It was predicted that the group which was presented with both positive and negative instances of the rule (henceforth positive-negative group) would be able to learn the rule, while the group with only positive instances (henceforth positive-only group) and the group with only negative instances (henceforth negative-only group) would not be able to learn the rule.

12.1 Method

12.1.1 Design

The design of this experiment is a 3 x 3 design with difficulty (easy, medium, and hard) and instantiation group (positive and negative instances, positive instances only, and negative instances only group) as independent variables and task performance and verbalisation ability as dependent variables. Since these were web-based experiments, Ss were asked to write down whether they knew any rule(s), what rule(s) they thought they knew, any strategies they were using, and any other comments they had after every 20 trials and at the end of the experiment. These written comments were then analysed as the verbalisation data (see also Section 8.1.4 and Figure 8.3 for a screenshot).

12.1.2 Subjects

There were 32 Ss in the positive-negative group. 10 Ss were assigned to the easy condition, 9 Ss to the medium condition, and 13 Ss to the hard condition.
19 Ss took part in the positive-only experiment, 7 Ss in the easy group, and 6 each in the medium and hard condition.
10 Ss took part in the negative-only experiment. 4 Ss were assigned to the easy condition, and 3 each to the medium and hard condition.

12.1.3 Stimuli

The stimuli for the positive-only group consisted of the same 40 positive strings as for the positive-negative group. The stimuli for the negative-only group consisted of the same 40 negative strings as the positive-negative group.

The rules for the negative-only group were reformulated to be the "negative rules". The easy rule was reformulated into the negative-easy rule: "Ruleful strings have no syllable repeated in the whole string. Unruleful strings have one syllable from the first half of the string repeated in the second half of the string."

The medium rule was reformulated into the negative-medium rule: "Ruleful strings have one syllable from the first half of the string repeated in any other position but the mirror image position in the second half of the string. Unruleful strings have one syllable from the first half of the string repeated in the mirror image position in the second half of the string."

The hard rule was reformulated into the negative-hard rule: "Ruleful strings have two repeated syllables either adjacent to each other or separated by more than one syllable. Unruleful strings have two repeated syllables separated by one other syllable."

12.1.4 Procedure

In the positive-negative group, Ss were presented with 40 positive instances and 40 negative instances shown one after the other in random order. Ss were told that there was a rule they should try to learn. In the test phase Ss were randomly presented with 40 new positive and negative instances. This is the same as the feedback, active, forewarned condition (see Chapter 11).

In the positive-only group, Ss were presented randomly, one after the other with only 40 positive instances in the training phase. They were told that all the stimuli followed the same rule(s) and that they should try to learn the rule. In the test phase Ss were presented with 40 new stimuli consisting of 20 new positive instances and 20 negative instances. Their task was to indicate whether or not they thought the stimulus followed the same rules as in the training phase.

Ss in the negative-only group received the same instructions as the positive-only group but they were presented with 40 negative instances in the training phase. The rules for the negative-only group were reformulated to be the opposite of the positive-only group (see Section 12.1.2).

12.2 Easy rule

12.2.1 Positive-negative group

The positive-negative group was the same as the feedback condition in Chapter 9. It is reproduced here for ease of reference.

10 Ss took part in this condition, 8 male, 2 female. The average performance in the test phase was 89.25%. The easy rule had been pre-tested and calibrated to be learnable within about 20 trials (see pilot studies, Chapter 9). Ss were performing at an average of 75% in trials 11-20.

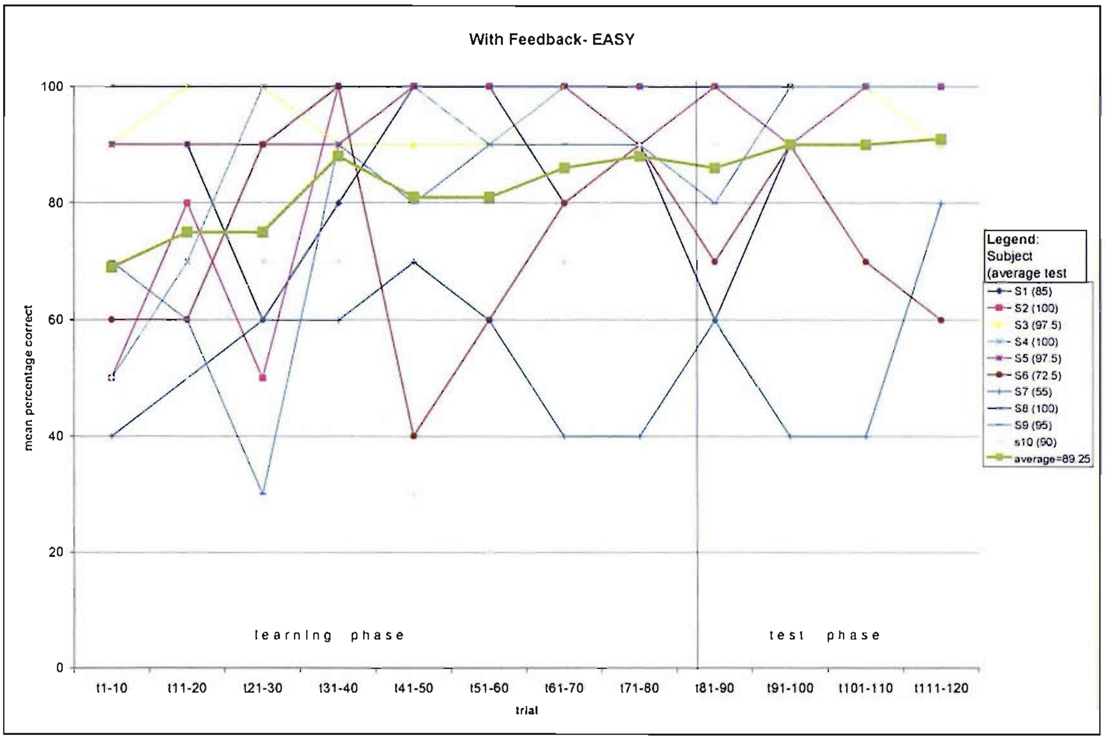


Figure 12.1: Learning curves for Ss in the feedback group in the easy condition (=Figure 9.1)

In Figure 12.1 the learning curves of all Ss in the feedback group in the easy condition are shown with the average learning curve in boldface. Most Ss could correctly distinguish ruleful from unruleful strings after about 30 trials, although most were also still making some mistakes in later trials, probably due to fatigue or lack of concentration.

Seven of the ten Ss could describe the rule after the maximum 40 trials. As soon as these Ss learnt the rule, performance improved from chance to near perfect, and they verbalised the rule at first opportunity (i.e. in the next break).

Of the remaining three Ss, one did not write anything down, but was performing at a high level of correctness; a further S only verbalised incorrect rules, and his overall performance in the test phase was only 55%. A further S stated the correct rule in combination with an incorrect rule after 20 test trials (so after completing the 80 learning trials and 20 of the 40 test trials), but is performing at a very high level (90% correct) from learning trial 60 onwards. This S may have learnt to respond correctly (perhaps also implicitly) at an earlier stage than his ability to verbalise the correct rule. However, it is also possible that this S did not write down the rule until completely certain of it. Ss were asked to write down any rules they were using during the experiment, but were not explicitly asked from which trial onwards they were using these rules from.

12.2.2 Positive Instances only

Seven Ss took part in this experiment: 3 male, 4 female. Average test performance was 71.1% correct.

Figure 12.2 shows the learning curves of all Ss in the easy condition, with their average performance shown in boldface. S2 was performing at chance level, and only named incorrect rules. S6 mentioned that he was looking for repetitions, but did not seem to learn the correct rule. However, his performance was above chance at 65%. S8 did not learn the rule and was performing at chance level. S9 seemed to have learnt the rule after 20 learning trials, but then did not seem to apply the rule until the second part of the test phase (i.e. trials 21 to 40).

This could to be due to the fact that S9 was not very confident that the rule was the correct one. S10 was following incorrect rules, but seemed to be performing at a level above chance (average 62.5%). This could be a chance fluctuation, but may also be an indication of implicit learning or memory of fragments. S11 verbalised the rule and was performing at top level throughout the test phase, but was not very sure or confident about her verbalisation ("...maybe not...", "...think it might be...", etc). S15 could verbalise the rule and was also performing well.

Overall, most Ss seemed to be performing at a level above chance. However, a lot of Ss were not very confident of their behaviour, and were thus not performing well.

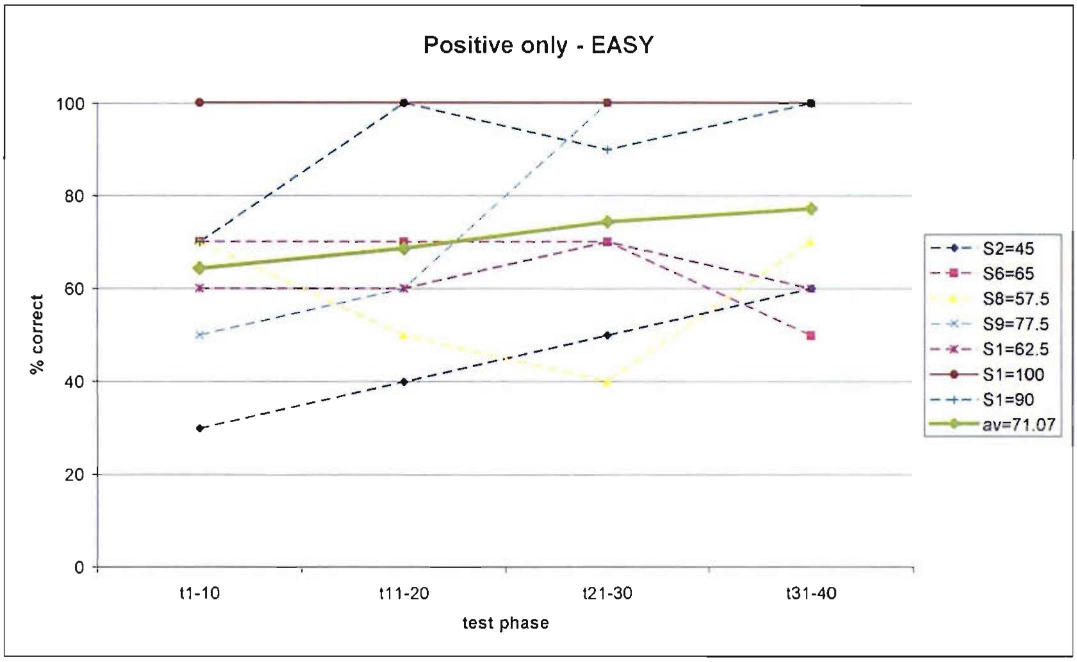


Figure 12.2: Learning curve of positive-only Ss in the easy condition

12.2.3 Negative Instances only

Four Ss took part in this experiment: 2 male, 2 female. The average test performance was 52.5% correct.

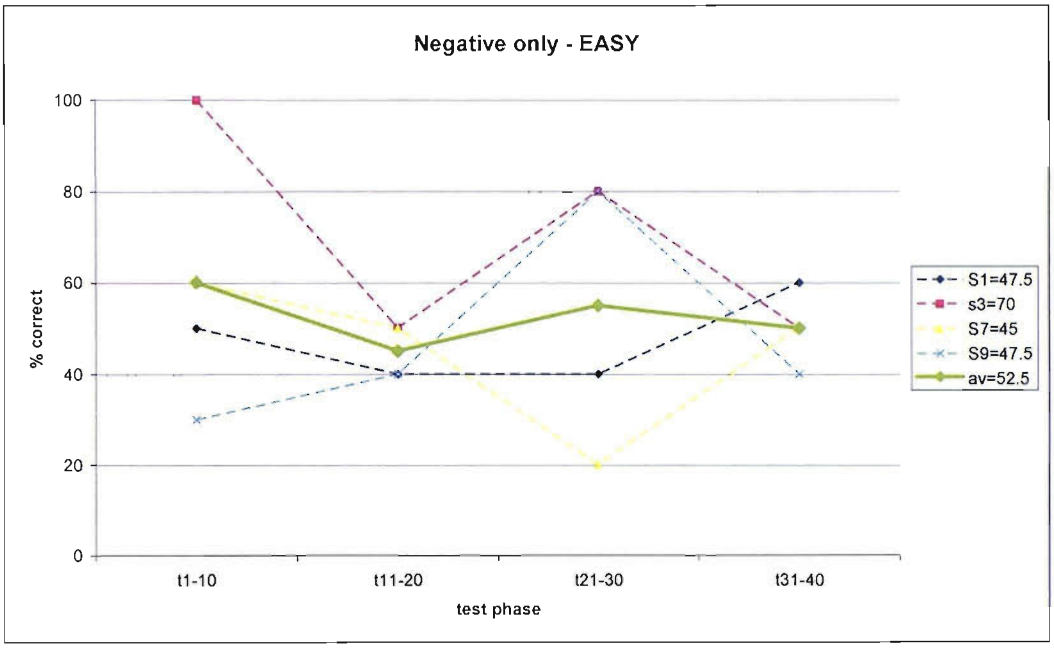


Figure 12.3: Learning curves of negative-only Ss in the easy condition

None of the Ss in the easy condition learnt the rule explicitly or implicitly. None of them could name even partial rules.

12.2.4 Verbalisation

Figure 12.4 shows the learning curves of verbalisers in the positive-negative group (continuous lines), positive-only verbalisers (dashed lines), and negative-only verbalisers (dotted lines), with their averages shown in boldface. The time of first verbalisation for verbalisers in each group is indicated with highlighted arrows. Ss in the positive-negative group who verbalised the rule did so in the training phase. Ss in the positive-only group who successfully verbalised the rule all did so after only 20 learning trials. There were no verbalisers in the negative-only group.

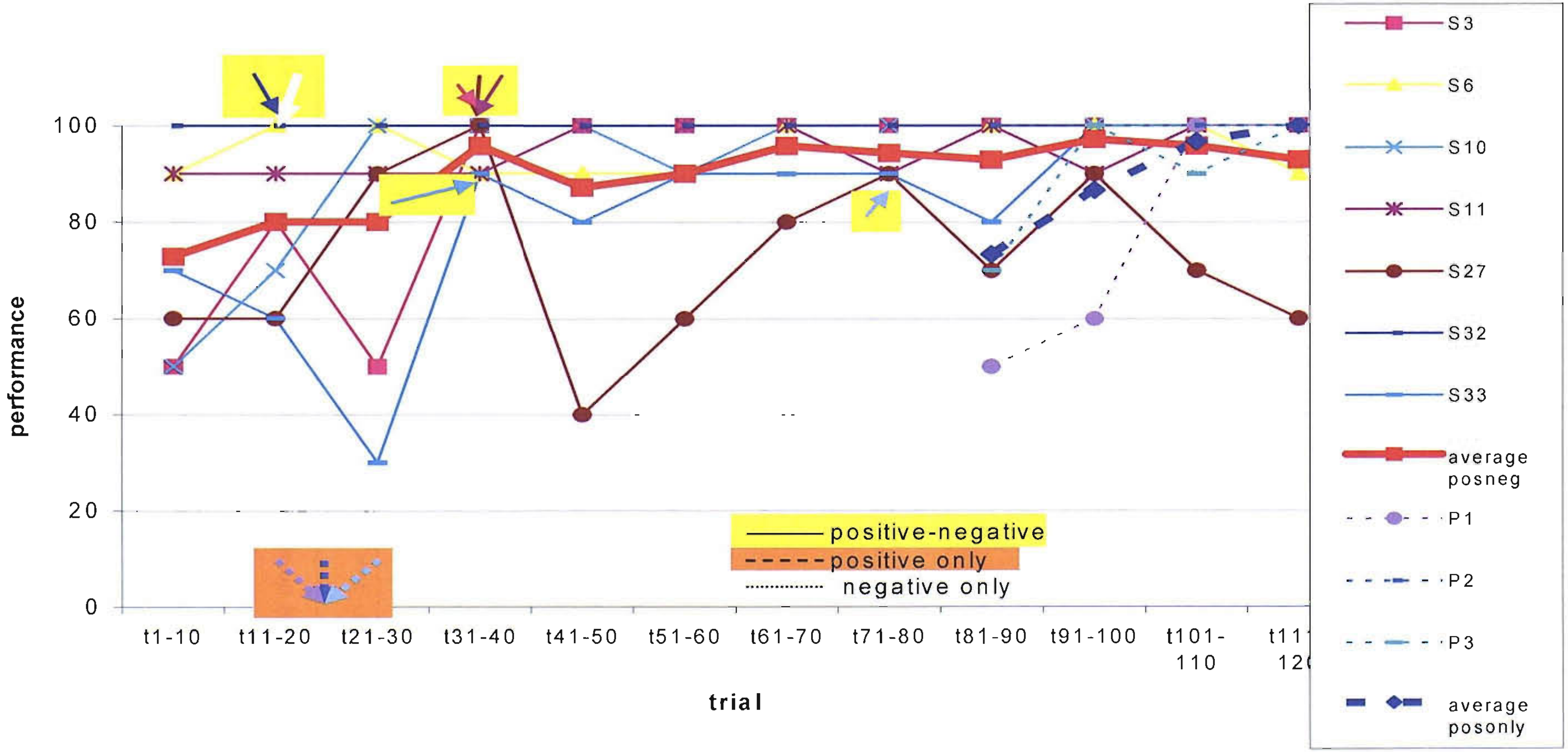


Figure 12.4 Learning curves of verbalisers in the positive-negative (continuous line), positive-only (dashed line), and negative-only (dotted line) group with verbalisation points for the easy rule

12.3 Medium rule

12.3.1 Positive-negative group

The positive-negative group was the same as the feedback group in Chapter 9. It is reproduced here for comparison purposes.

Nine Ss took part in this condition, 4 male, and 5 female. Ss were performing at an overall average of 80.28% correct in the test phase. The medium rule was precalibrated in the pilot phase to be learnable in about 40 trials.
In trials 31-40, Ss were performing at 73.33% correct.

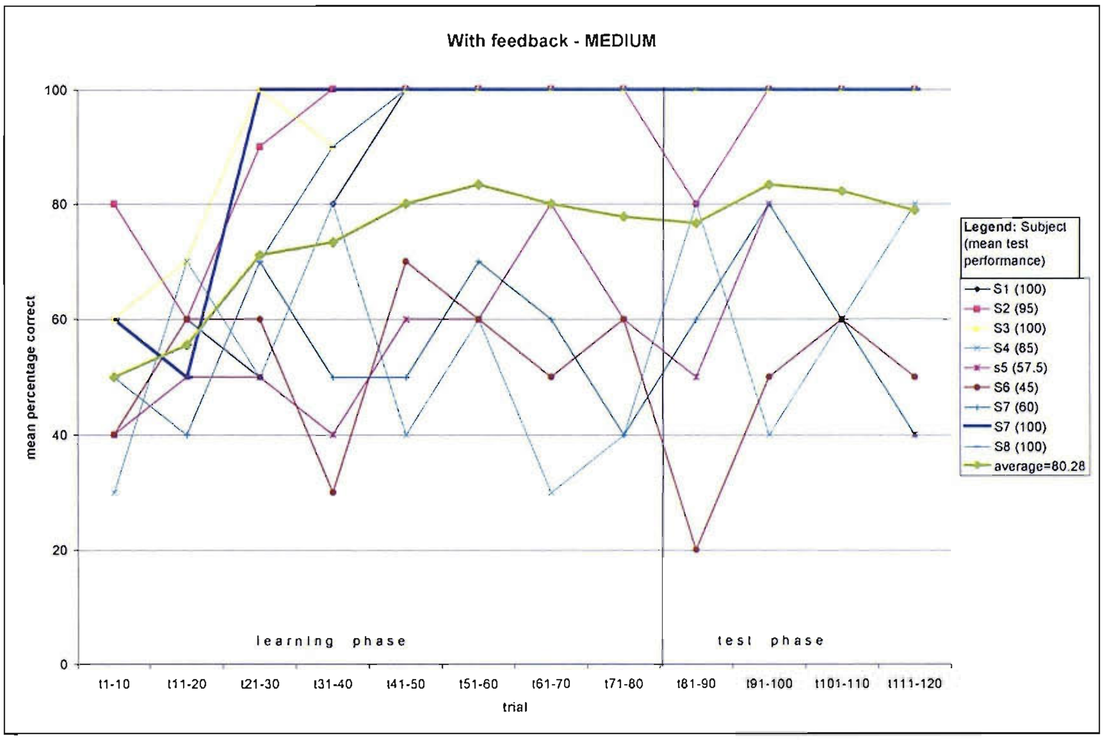


Figure 12.5 Learning curves of Ss in the feedback group in medium difficulty condition (=Figure 9.4)

Figure 12.5 shows the learning curves of the Ss with feedback in the medium difficulty condition, with their average in boldface. Most Ss seemed to learn this rule more gradually than the easy rule, first noticing the repeated syllables, and then later noticing the relevance of the position in the string. These Ss were performing at levels above chance when they had learnt the partial rule (repeated syllable), and their performance improved to near 100% correct when they had learnt the additional rule (position relevance). Some Ss only learned partial rules; for example, they noticed that it had something to do with the repeated syllable, but could not figure out the relevance of the position. These partial learners were performing at levels better than chance, but did not reach perfect performance.

12.3.2 Positive Instances Only

Four Ss took part in this experiment, 4 male, and 2 female. Average performance in the test phase was 62.9% correct.

Figure 12.6 shows the learning curves of all positive-only Ss in the medium condition, with the average performance shown in boldface. S1 did not know the rules and was performing at chance. S3 verbalised the rule after 20 test trials and was performing accordingly. S4 recognised a partial rule after 20 learning trials, and the full rule after 20 test trials. His performance corresponded to this, as he was only performing well in the last 20 trials. S13 could verbalise the rule after 20 learning trials and was performing at top level in the test phase. Although S14 verbalised the rule after 20 learning trials, his performance did not match this and his overall average in the test phase was only 42.5% correct. This S seemed to have changed from the correct rule to an incorrect rule. S17 did not learn the rule and was performing at chance level in the test phase.

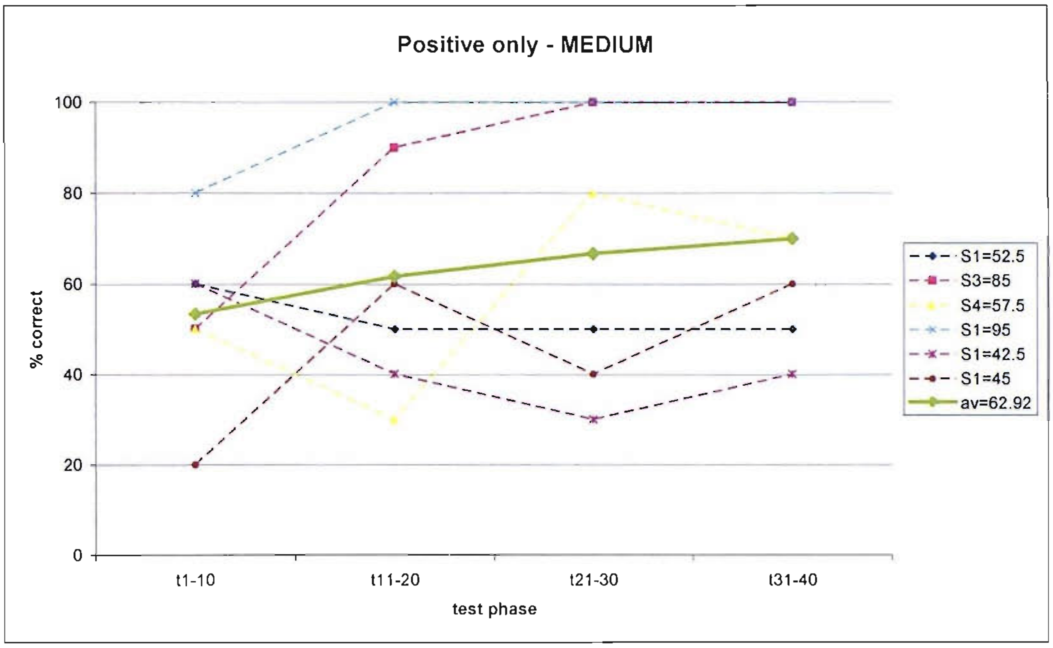


Figure 12.6: Learning curves of Ss with positive instances only in the medium condition

12.3.3 Negative Instances Only

3 Ss took part in this experiment, 2 male and 1 female. The average test performance was 46.7% correct.

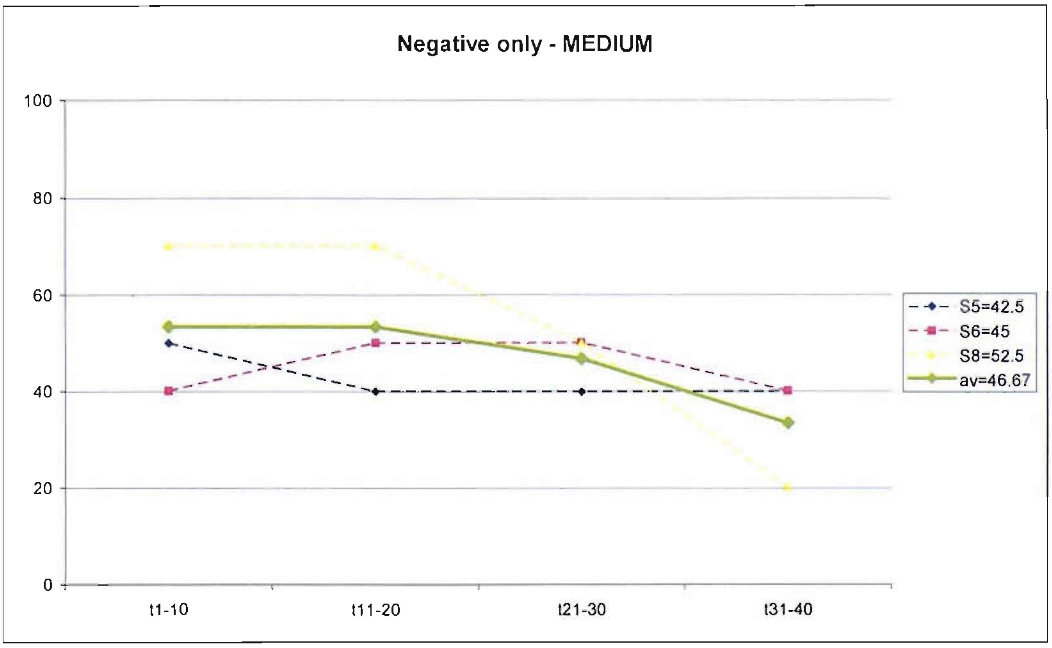


Figure 12.7: Learning curves of Ss in the negative-only group in the medium condition

One of the negative-only Ss in the medium condition could name a partial rule, but his performance was at about chance level. The other two Ss did not learn the rules explicitly or implicitly.

12.3.4 Verbalisation

In Figure 12.8 the learning curves of verbalisers in the positive-negative group are shown with continuous lines, positive-only verbalisers with dashed lines, and negative-only verbalisers with dotted lines; averages are in boldface. The verbalisation times are indicated by highlighted arrows for each S. Verbalisers in the positive-negative group were verbalising the rule after about 40 learning trials and performing at high levels. Some verbalisers in the positive-only group were already verbalising the rule in the training phase, while others only verbalised the rule in the test phase. However, not all the verbalisers in the positive only group were performing at very high levels. There were no verbalisers in the negative-only group.

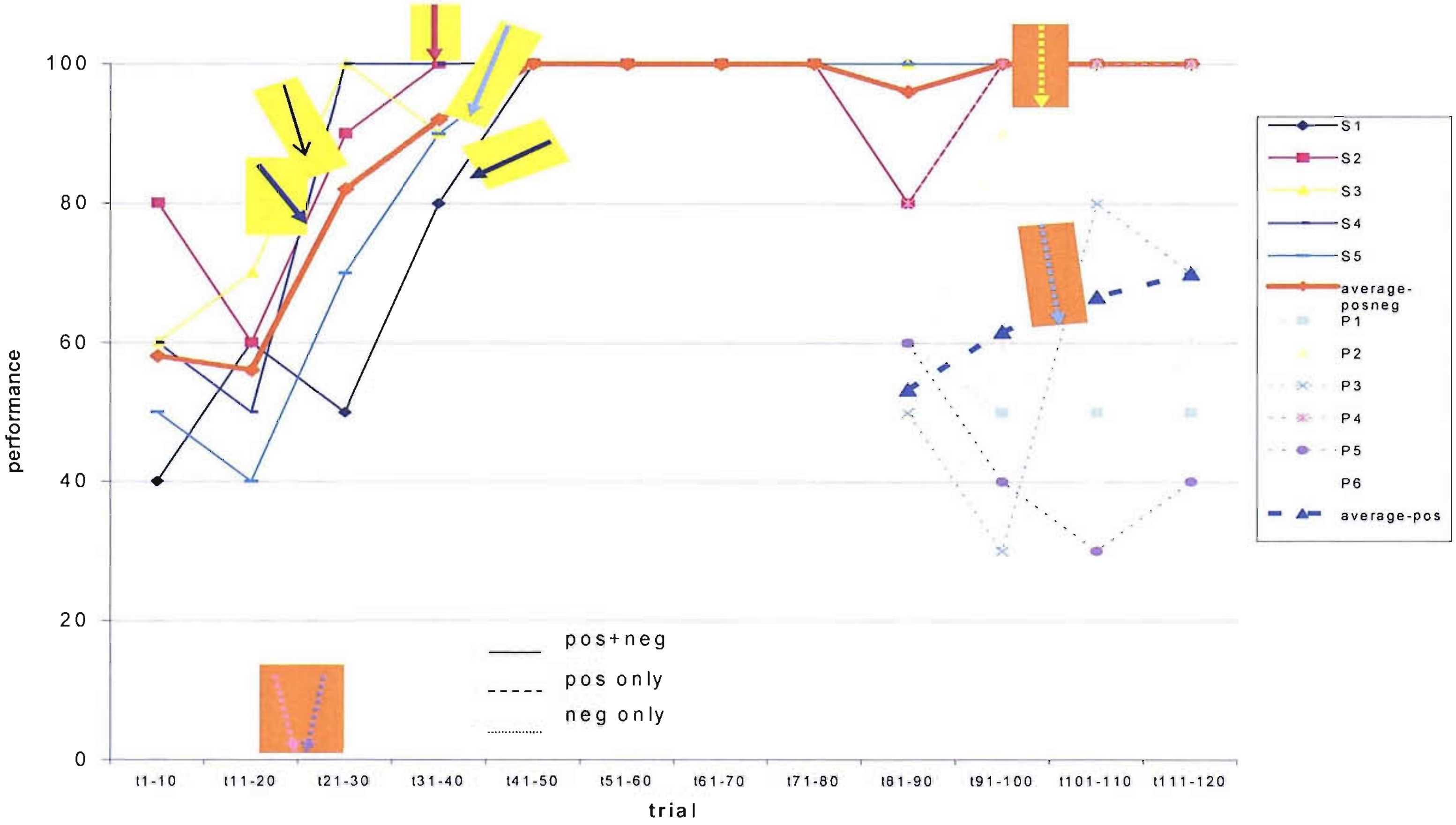


Figure 12.8 Medium rule: Learning curves of verbalisers in the positive-negative (continuous line), positive-only (dashed line), and negative-only (dotted line) group, with verbalisation points.

12.4 Hard rule

12.4.1 Positive-negative group

The positive-negative group was the same as the Feedback condition in Chapter 9. For comparison purposes it is reproduced here.

Thirteen Ss took part in this condition, 9 male, 4 female. Overall test performance was 63.46% correct. In the pilot studies the hard rule was predetermined to be learnable in about 80 trials. In trials 71-80, Ss were performing at a level of 68.46% correct.

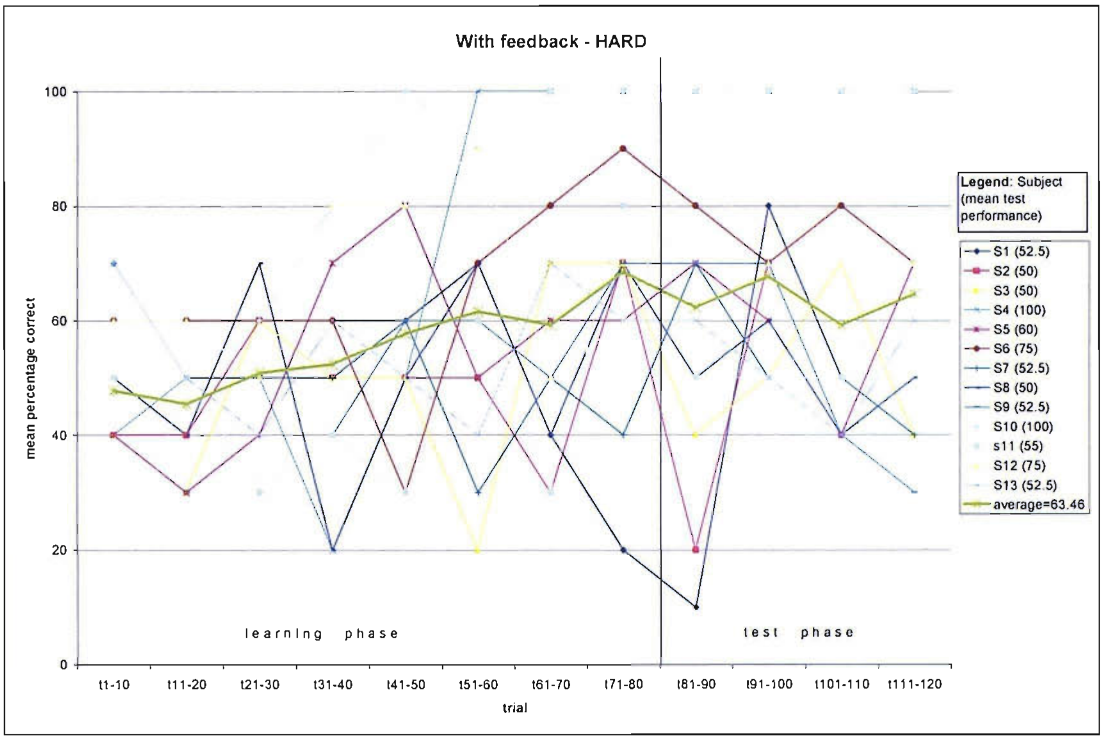


Figure 12.9 Learning curves of Ss in feedback group in the hard difficulty condition (=Figure 9.7)

Figure 12.9 (=Figure 9.7) shows the learning curves of the Ss with feedback in the hard condition. The boldface line is the average learning curve. Two Ss learned the rule explicitly and could describe it in words. Most Ss learned partial rules, for example "the same syllable may not repeat directly (like Sa Sa)". Half of the Ss could verbalise the correct rule, or partial rules, while half could not explicitly state the rule.

12.4.1 Positive Instances Only

Six Ss took part in this experiment, 3 male and 3 female. Average performance was 62.1% correct in the test phase.

Figure 12.10 shows the learning curves of all positive-only Ss in the hard condition, with the average performance shown in boldface. S5 did not learn the rule and was performing at chance in the test phase. S7 verbalised a partial rule after 40 test trials, but in general did not seem to have learnt the rule, as also shown by his performance of slightly above chance (60%). S12 stated a partial rule after 20 test trials and the full rule after 40 test trials (i.e. at the end of the experiment). This S's performance also indicated that he had learnt the rule. S16 learnt the rule after 20 test trials; his performance improved in the last 20 trials. S18 did not learn the rules and was performing at chance. S19 did not write down any comments and was performing at chance.

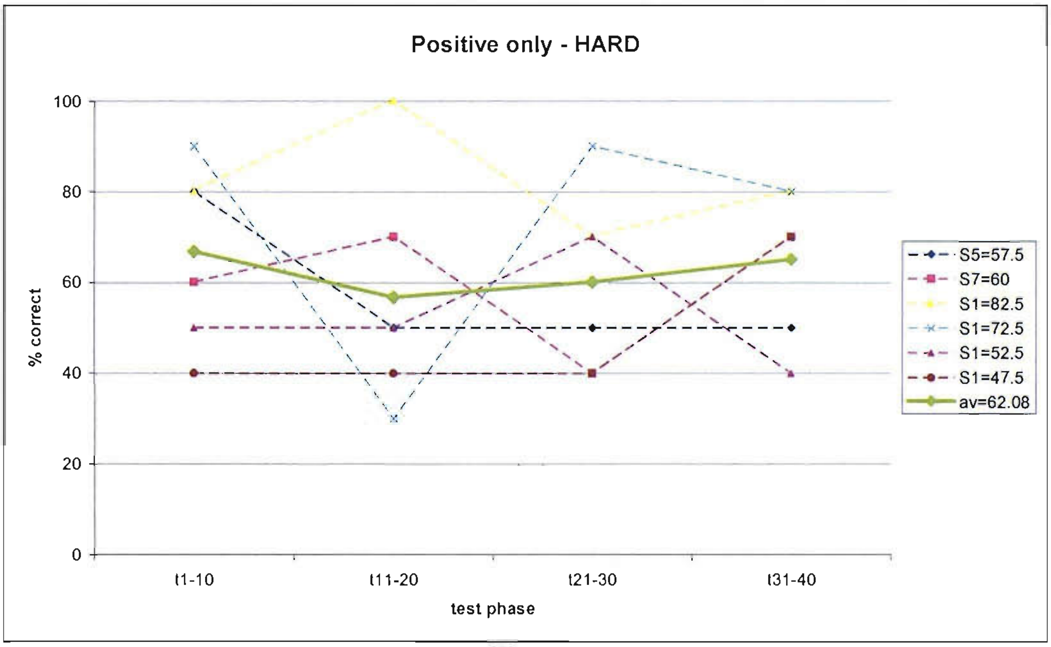


Figure 12.10: Learning curves of Ss with positive instances only in the hard condition

12.4.2 Negative Instances Only

3 Ss took part in this experiment, 1 male and 2 female. The average performance was 61.7% correct in the test phase.

Two Ss wrote down partial rules, and their overall test performance was also above chance, although S2's performance dropped through the test phase from 70% correct to 30% correct. S10 wrote down some incorrect rules, but his performance rose from 30% correct to 80% correct in the test phase

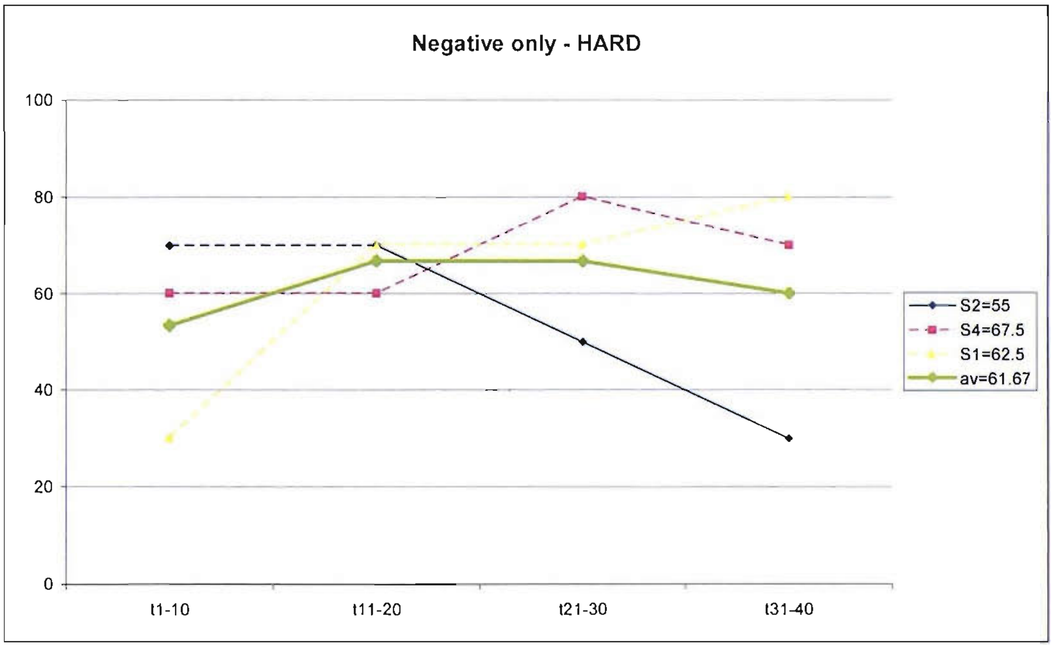


Figure 12.11: Learning curves of Ss in the negative-only group in the hard condition

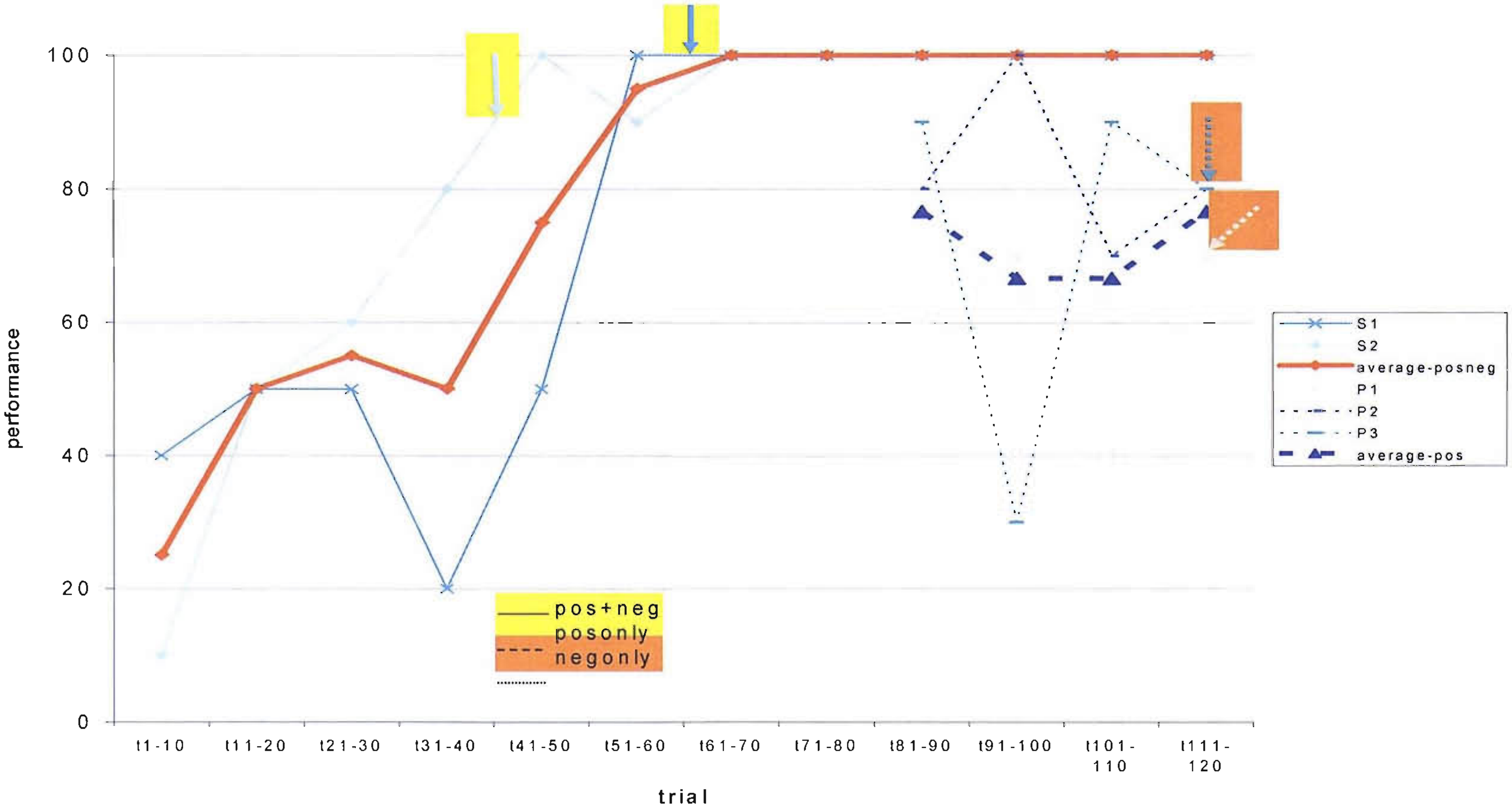


Figure 12.12 Hard rule: Learning curves of verbalisers in the positive-negative (continuous line), positive-only (dashed line), and negative-only (dotted line) group, with verbalisation points.

12.4.3 Verbalisation

Figure 12.12 shows the learning curves of verbalisers in the positive-negative group with continuous lines, positive-only verbalisers with dashed lines, and negative-only verbalisers dotted lines. The average learning curves of each group are in boldface. Time of verbalisation is indicated with highlighted arrows.

Verbalisers in the positive-negative group verbalised the rule earlier in the experiment than verbalisers in the positive-only group. There were no verbalisers in the negative-only group. Verbalisers in the positive-only group were not performing as well as verbalisers in the positive-negative group.

12.5 Results

12.5.1 Performance

The results of a two-way ANOVA with test performance as dependent variable and type of instance and difficulty as independent variables revealed a significant main effect of instantiation type ($F_{2,61}$=7.820, p<0.01). Ss in the positive-negative, positive-only and negative-only groups were behaving differently, as can also be seen in Table 12.1 in numerical form and in Figure 12.13 in graphical form.

	easy	**Medium**	**hard**
positive-negative	89.25	80.28	63.46
positive only	71.07	62.92	62.08
negative only	52.5	46.67	61.67

Table 12.1 Mean percentages of correct answers in the test phase

A positive correlation between test performance and difficulty level was observed (r=0.276, p>0.05). Ss in the easy condition were performing better than Ss in the harder conditions.

Interestingly, Ss in the hard condition seemed to be performing at the same level of accuracy regardless of which instance group they were in. In the easy and medium condition, Ss in the positive-negative group were performing at the highest level, Ss in the positive-only group were performing at above-chance levels, while Ss in the negative-only group were performing at chance level.

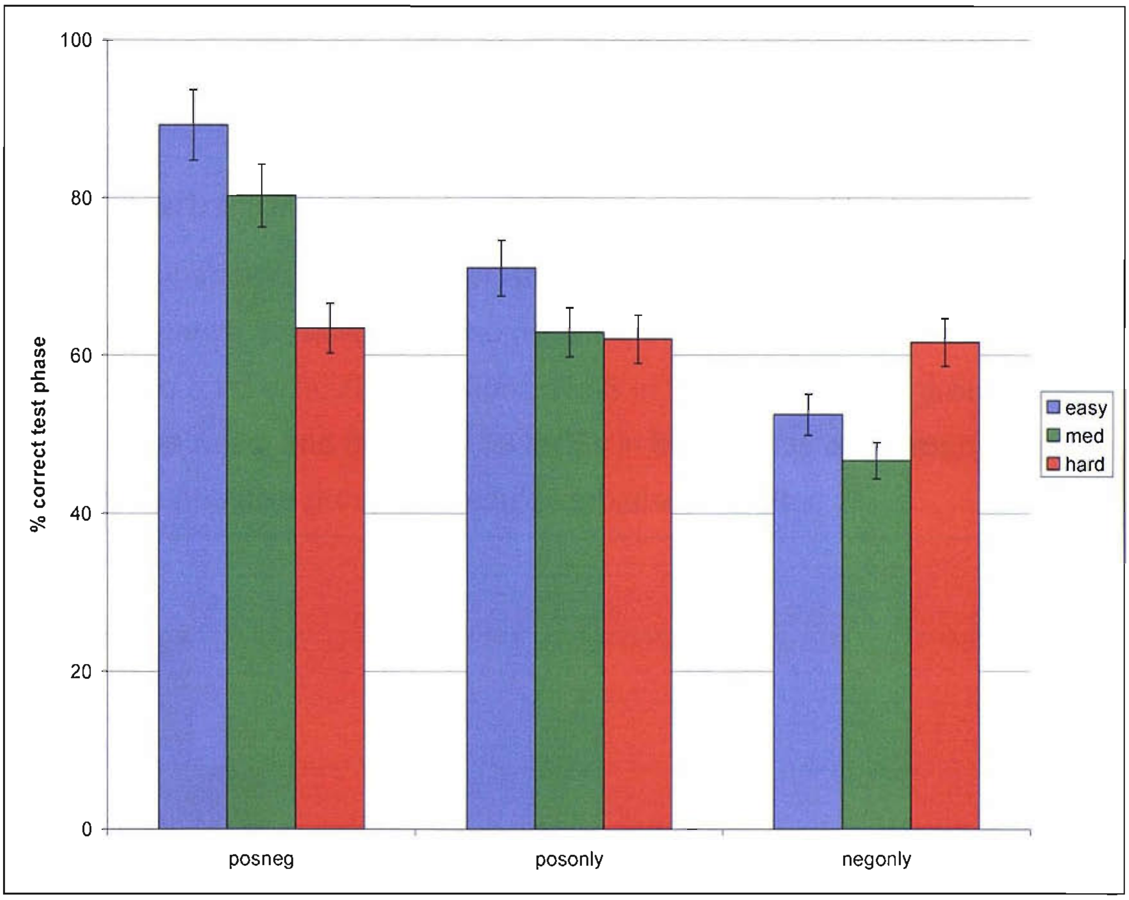


Figure 12.13: Mean percentage of correct responses in the test phase for each of the 3 difficulty conditions for positive-negative group, positive-only group, and the negative-only group

Subjects in the negative-only condition were performing at a lower level than subjects in the positive-only condition, and negative-only Ss were not able to verbalise the rules. It has been shown that negative rules are harder to learn than positive rules (Hovland & Weiss, 1953; Shumway, 1983). Although Ss in both the positive-only and the negative-only condition are completing the same task there are many more instances of the negative rule than of the positive rule, which would explain why it was harder to learn and why Ss are performing worse. Table 12.2 shows the numbers of possible positive and negative instances for each of the difficulty condition. For all condition, there are more possible negative instances than positive instances.

	Positive (Ruleful)	Negative (Unruleful)
Easy	92,252,160	980,179,200
Medium	23,063,040	69,189,120
Hard	34,594,560	115,315,200

Table 12.2 Number of possible positive and negative instances

12.5.2 Verbalisation

Figure 12.14 shows the number of Ss who could verbalise the rules in the positive-negative, positive only and negative only conditions, for the easy, medium and hard difficulty conditions. No S in the negative only group could verbalise the rules, and there are fewer Ss in the positive only group than in the positive-negative group who could verbalise the rules.

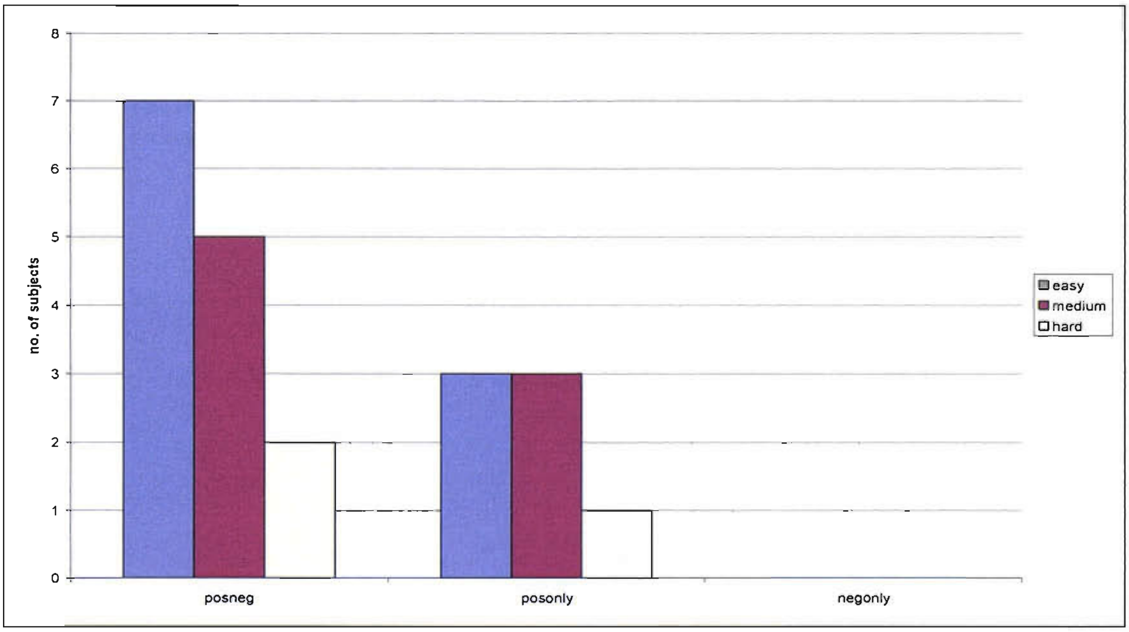


Figure 12.14 No. of subjects in the positive-negative, positive only, and negative only group who could verbalise the rules for the three difficulty conditions

A two-way ANOVA with verbalisation as dependent variable and instantiation and difficulty as independent variables showed a significant main effect of instantiation ($F_{2,61}$=3.400, $p<0.05$). Ss in the different instance groups were performing significantly differently on verbalisation. This is presumably because more Ss in the positive-negative group were performing well and verbalising the rule, than in the groups with only one type of stimulus (either positive or negative). See Figures 12.15 for the easy condition, Figure 12.16 for the medium condition, and 12.17 for the hard condition.

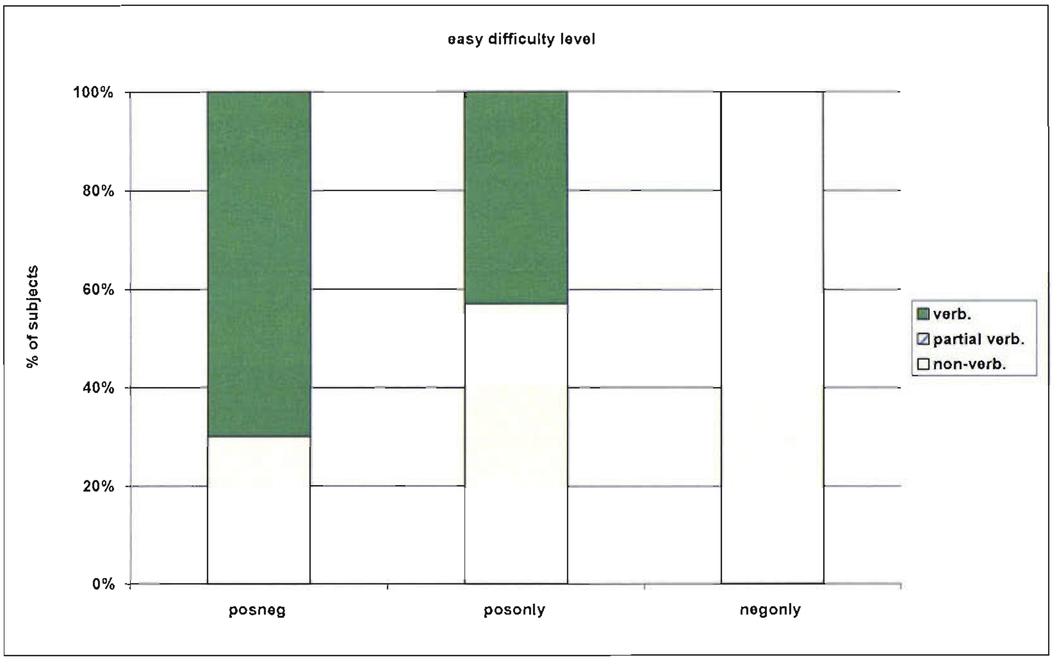


Figure 12.15: Percentage of Ss who could verbalise, partially verbalise and not verbalise in easy condition

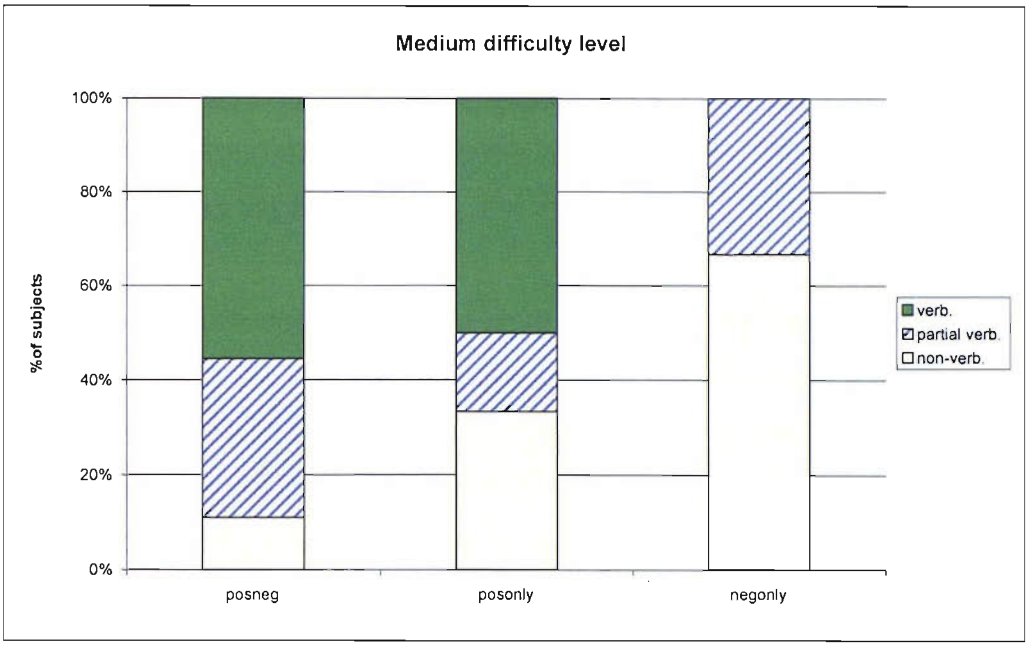


Figure 12.16: Percentage of Ss who could verbalise, partially verbalise and not verbalise in medium difficulty condition

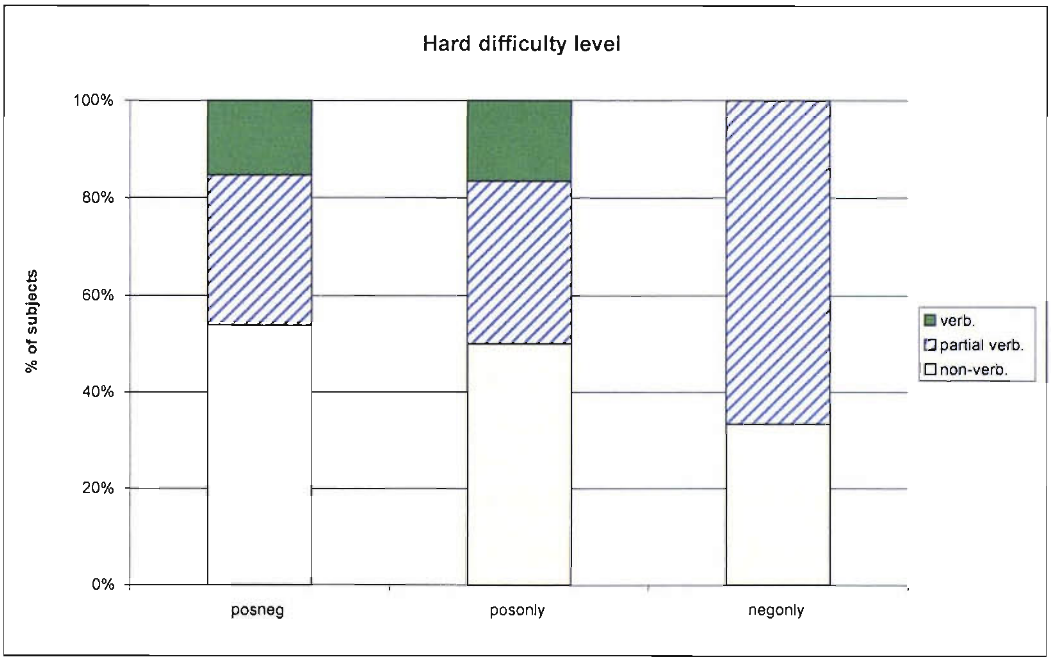


Figure 12.17: Percentage of Ss who could verbalise, partially verbalise and not verbalise in hard condition

There was a positive correlation between verbalisation and test performance (r=7.90, p<0.01). Ss who were performing well in the test phase could also verbalise the rules.

12.6 Conclusions

In this experiment, as in previous experiments, those Ss who were performing the task well could also verbalise the rules correctly. The learning seen in this experiment was explicit learning, not implicit.

Ss in the positive-negative group were performing at the highest level of accuracy; Ss in the positive-only group were performing at a level above chance, while Ss in the negative-only group were performing at chance. The hypothesis that Ss in the positive-only group would not be able to learn the rules successfully must therefore be rejected. It seems that the positive rules were so simple, that they were learnable from positive instances only.

The negative rules used in these experiments were the inverse of the positive rules. The numbers of possible negative instances were bigger for all difficulty conditions. Thus, the difficulty of learning the positive and the negative rule was not equal. This would explain why Ss did not learn the negative rule, and is also support for the fact that with a more difficult rule, Ss do not learn the rule from one type of instance only. Further research could investigate equating the number of possible instances for the positive and negative rule, which should yield similar performance on both rules. Equally, providing Ss with an proportionally equal number of instances should yield similar performance on both rules.

It is noteworthy that Ss in the positive-only group were much less confident of their verbalisations than Ss in the positive-negative group. Positive-only Ss often inserted words indicating uncertainty in their reports, such as “perhaps”, “maybe”, “not sure”, “possibly”, etc. Several Ss also abandoned the correctly learnt rule for an incorrect rule, and were then classifying strings according to that incorrect rule. This is an example of the credit-blame assignment problem (see Section 2.2). These Ss did not know which hypothesis to “blame” and which to “credit”. It can be seen in Figures 12.4, 12.8, and 12.12, that verbalisers in the positive-only condition were not necessarily classifying the strings correctly, as they were so unsure of their rules.

13 Discussion and Conclusions

This thesis is an attempt to find some meaningful regularities in the vast array of confusing findings in the field of implicit learning, in particular of artificial grammar learning (AGL). Research in AGL has its origins in the field of natural language learning, in which the ease with which young children "implicitly" pick up their native language from positive evidence only was acknowledged to be a remarkable feat worthy of research. Artificial grammars (AGs) were chosen as a means of studying implicit learning because like natural grammars they are sets of complex rules. In traditional AGL experiments, Ss are presented with positive instances generated by AGs, but are not told that the strings all follow rules. Instead they are instructed to do a different task (e.g. memorise the strings) and are only later (after this learning phase) told about the underlying rule(s). In a subsequent test phase, Ss have to differentiate new positive instances from negative ones, and the usual findings are that Ss are performing at a level significantly above chance, without, however, being able to verbalise any of the rules. Researchers (e.g. Reber, 1967) concluded that "implicit learning", i.e. learning without consciousness, was taking place.

The initial ties to natural language learning were soon dropped as it became evident with the Chomskian revolution that natural language learning was underdetermined by the data that the language learning child encounters and produces, hence some form of innate Universal Grammar must exist. This meant that AGL had virtually nothing in common with natural language learning. The main aim of AGL research, after this realisation that it could not shed much light on natural language learning, was to test Ss in tasks that did not require them to learn the rules of the AG deliberately and explicitly, to see whether they could learn anything at all about the underlying structure of the stimuli under such passive conditions. Any improvement in performance was taken to mean that learning was unconscious or implicit. Unfortunately, however, the use of very complex rules and of positive evidence only, which were initially factors pertinent specifically to natural language, remained an inherent part of AGL research.

In the experiment reported in chapter 4 of this thesis, replicating the traditional AGL paradigm, with positive instances only, it was found that Ss did not learn anything – explicitly or implicitly – about the underlying rules, but rather, were basing their classification of test stimuli on the similarity (in terms of shared parts) between each individual test stimulus and the training stimuli with which Ss were already familiar. It was also found that although Ss did not have any awareness of the rules of the AG, they seemed to have some sense of how well they were performing in classifying the test stimuli, because they were more confident with stimuli that were similar to the specific training stimuli they had previously encountered (i.e., when they had some of the same parts). In a further experiment reported in chapter 6, it was found that presenting Ss with negative instances as well as positive instances in the training phase in the traditional AGL paradigm interfered with rather than aiding learning, and that Ss subsequently performed at chance in the test phase. Adding negative instances merely increased the amount of information Ss had to process and remember, because not only did they now have to remember parts of the stimuli they had seen before, but they also had to try to remember the colour (e.g. green for positive stimuli, red for negative stimuli) of the stimulus.

Apart from always using positive evidence only and very complex rules in traditional AGL research, no one had yet taken the essential antecedent step of ensuring that the AG that was being used was *learnable* in the first place. In the pilot studies of the main experimental series in this thesis, it became clear that only extremely simple rules are learnable to a reasonable level of correctness within the usual 80 learning trials used in this kind of experiment. This finding already calls into question what experiments using the very complex rules of traditional AGL experiments could be expected to reveal about the learning of rules. In these pilot studies, Ss were in optimal learning conditions in which they were told there was a rule that they should try to learn and were presented with ruleful (positive) and unruleful (negative) instances, followed by corrective feedback on each response. If such Ss could only learn very simple rules, it seems highly unlikely that Ss who were not told about the existence of rules, were presented only with positive instances – and were also performing a completely different accompanying task at the same time – would learn any rules at all, implicitly or explicitly.

The main finding of this thesis is that Ss who were performing at a high level in the test phase were also very likely to be able to verbalise the rules. In other words, when there was any learning at all in these experiments, it was explicit learning. There was no indication whatsoever of any implicit learning at the end of the experiments reported here. No Ss who were verbalising incorrect rules were performing better than chance. All Ss who were performing better than chance and writing down their comments (unfortunately not all Ss wrote down comments even though they were instructed to do so), were able to verbalise either the full correct rules, or partial rules. Ss who verbalised partial rules were usually performing above chance, but not at 100%. This makes sense, as Ss using only a partial rule would be classifying correctly those stimuli that adhered to the partial rule, but they could be classifying other stimuli incorrectly: those that either did not conform to their correct partial rule, or those that conformed to their correct partial rule but not to the partial rule they had not yet learnt. There was one S who indicated that they thought they may know the rule implicitly, although there was no way of testing this experimentally, and that same S later did explicitly state the rule. There is also the possibility that some of those Ss who were not writing anything down, but were performing well had learnt something implicitly, although this is speculation.

It was found that immediate corrective feedback after every individual trial (as opposed to delayed feedback on average performance after every 20 trials) had a beneficial effect on learning. Ss who received immediate corrective feedback learned the rule more quickly and thus reached high levels of performance earlier, and were more confident that they had learnt the correct rule(s) (evidenced by their comments). Delayed feedback also had a favourable effect on learning, as some Ss in the no feedback groups also learnt the rule(s) or partial rules. Their learning was slower than feedback Ss and they were much less confident about the correctness of their rule(s), to the point of even abandoning the correct rule(s) in some instances. Ss were struggling with the credit/blame assignment problem: they did not know which rule(s) to credit with their correct responses and which rule(s) to blame for incorrect responses.

Contrary to our prediction, however, active responding did not seem to benefit learning. Rather, what seemed to generate better and faster learning was receiving passive feedback, without having to make any overt response about the rulefulness of the stimulus. This may have been because repeating the stimulus aloud (by pronouncing and clicking) was already a sufficiently "active" response. Having to make a further response about the rulefulness may have required an additional effort that slowed down (or even interfered with) learning.

Ss learned the rules almost as well and quickly when presented with positive instances only as when presented with both positive and negative instances. However, one major difference between the two groups was that Ss in the positive-only group were much less confident about the rule, even when it was correct, and they sometimes even switched to incorrect rules, similar to the effect seen with the no-feedback groups. In contrast, Ss in the positive-negative group hardly ever expressed uncertainty once they had articulated the correct rules, and never switched to incorrect rules. One possible explanation for this finding is that the rules were so simple that Ss could pick them up from positive evidence only, but they could not be sure that the rules were correct, since they did not see any negative evidence to bolster their hypotheses and dispel any doubts. The fact that Ss in the negative-only group did not learn the rules either explicitly (verbalisation) or implicitly (performance) indicates that the negative rules were more difficult to learn than the positive rules, and supports the inference that learning more difficult rules is itself more difficult when it is based on one type of instance only. There were many more possible instances of the negative rule than of the positive rule, which suggests that it would take longer to learn. It is likely, that Ss could have learnt the negative rule given more trials.

The type of instructions Ss received also had a large effect on the learning of the rules. Ss who were told the rules in advance (forearmed condition) were performing at top level from the beginning of the experiment. Most Ss who were told that there were underlying rules that they should try to learn (forewarned condition) learned the rules within 80 learning trials. On the other hand, Ss who were not told that there were underlying rules (rule-blind condition) did not learn anything about the rules and were performing at chance in the test phase.

The condition in which Ss were not told anything about the underlying rules is similar to traditional AGL experiments. In the experiments in this thesis, however, Ss could not rely on the memorisation of parts (which is the probable basis for their above-chance performance in traditional AGL research), because the strings consist of 80% random syllables (two syllables that conform to the rule, and six random syllables). The results show that most Ss in the rule-blind condition (implicit learning) did not learn to identify which strings were ruleful. In addition, those few Ss who did learn in the rule-blind condition were fully able to verbalise the rules. This indicates that Ss were learning the rules explicitly, and were fully aware of what they were.

In the experiments in this thesis, all Ss were explicitly asked in every break – so after every 20 trials – to write down any rule(s) they thought they knew, any strategies they were using to help them classify the strings, and anything else they were thinking and doing, no matter how irrelevant they thought it might be[22]. We cannot know if those Ss who did not write anything down learnt the rule(s) implicitly or explicitly. In addition, we cannot know if Ss who explicitly verbalised the rules, learnt the rule(s) implicitly first. This was implied by only one S and it could not be empirically verified – due to the nature of the particular experiment S was doing. However, most Ss who were performing at above-chance levels, and who did write something down, could explicitly verbalise the rule(s), or at least partial rule(s). Ss who learnt partial rules were

[22] *Apart from the rule-blind condition in which Ss did not know about the existence of rules, and were thus not asked if they knew the rules, but only whether they were using any strategies to help them with their task and for any other comments.*

performing at above-chance levels of around 70% correct, as they were getting those trials correct, which conformed to their partial rules. Ss who were verbalising incorrect rules were performing at chance levels. There seemed to be no evidence of implicit learning at all in the experiments presented in this thesis, even with our ultra-simple rules. This finding is in contradiction with classical implicit learning theories, according to which Ss are learning abstract rules without awareness. Since Ss in our experiments – with very simple rules and both positive and negative evidence – did not learn anything implicitly (and when they learned at all, always learned it explicitly) it would seem that previously reported "implicit learning" effects in these types of experiments could only have been based on the recurrence of remembered parts of the stimuli.

Despite these laboratory findings, implicit learning may still occur in the real world, perhaps with other kinds of stimuli, instead of sequences of AG syllables. Serial processing of quasi-verbal stimuli might be especially conducive to explicit processing. It is possible that implicit learning of "rules" in the form of parallel feature detection in non-sequential stimuli (e.g. visual displays) does occur. Implicit learning of social behaviours is another area in which the unconscious learning of "rules" is plausible. However, the findings in this thesis suggest that the sequentially presented AG paradigm that was expressly designed for the study of implicit learning fails to produce implicit learning, with the simplest of rules. Hence the traditional AG learning effects that probably arose because (1) the rules were too hard to be learnable at all, and hence (2) Ss had no choice but to memorise recurring parts. Inasmuch as there was an uncontrolled correlation between recurring parts among the positive-only stimuli in the training strings and correct performance on the test stimuli, the "implicit learning" was merely an artefact that vanished once both positive and negative stimuli were considered, and the rule was made simple enough to learn (explicitly) and verbalise. There is still considerable scope for studying implicit and explicit learning with stimuli and rules of increasing difficulty, although to be realistic, the training trials may need to be extended well beyond the usual experimental subject's hour's worth

References

- Altmann, G., Dienes, Z., & Goode, A. (1995). On the modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **21**, 899-912.

- Bourne, L.E., Jr. (1970). Knowing and using concepts. *Psychological Review*, **77**, 546-556.

- Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. In: E. Rosch & B.B. Lloyd (Eds.), *Cognition and Categorization* (pp. 169-211). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

- Brooks, L.R. & Vokey, J.R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Matthews et al (1989). . *Journal of Experimental Psychology: General,* **120**, 316-323.

- Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J.Hayes (Ed.), *Cognition and the development of language* (pp.155-207). New York: Wiley.

- Bruner, J.S. & Goodnow, J.L., & Austin, G.A. (1956). *A Study of Thinking*, New York: Wiley.

- Chan, C. (1992). Implicit cognitive processes: Theoretical issues and applications in computer systems design. Unpublished D.Phil. thesis, University of Oxford.

- Cheesman, J, & Merikle, P.M. (1984). Priming with and without awareness. *Perception and Psychophysics*, **36**, 387-395.

- Cheesman, J, & Merikle, P.M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology,* **40**, 343-367.

- Chomsky, N. (1968). *Language and Mind.* New York: Harcourt, Brace.

- Chomsky, N. (1980). *Rules and representations.* New York: Columbia University Press

- Chomsky, N. (1987). On the nature, use, and acquisition of language. In: *Handbook of Child Language Acquisition,* W.C. Ritchie & T.K. Bhatia (Eds.), 1999, Academic Press, Ltd., Ch. 2, pp. 33-54.

- Dienes, Z., Altmann, G.T.M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **21**, 1322-1338.

- Dienes, Z. & Altmann, G. (1997). Transfer of implicit knowledge across domains: How implicit and how abstract? In: *How implicit is implicit learning?*, D. Berry (Ed.). Oxford University Press.

- Dienes, Z., Broadbent, D.E., & Berry, D. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **17**, 875-887.

- Dienes, Z. & Perner, J. (1996). Implicit knowledge in people and connectionist networks. In *Implicit cognition.* (Ed. G. Underwood), Oxford University Press.

- Dulany, D.E., Carlson, R., & Dewey, G. (1984). A case of syntactical learning and judgement: How concrete and how abstract? *Journal of Experimental Psychology: General,* **113,** 541-555.

- Gold, E. M. (1967). Language identification in the limit. *Information and Control*, **10**, 447-474.

- Hovland, C. I. & Weiss, W. (1953) Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology*, **45**, 175-182

- Jacoby, L.L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight – Limits on the ability to avoid unconscious influences from the past. *Journal of Personality and Social Psychology,* **56,** 326-338.

- Johnstone, T. & Shanks, D.R. (1999). Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997). . *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **25**, 524-531.

- Johnstone, T. & Shanks, D.R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology,* **42**, 61-112.

- Kassin, S.M., & Reber, A.S. (1979). Locus of control and the learning of an artificial language. *Journal of Research in Personality,* **13**, 111-118.

- Klahr, D., Chase, W.G., & Lovelace, E.A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **9**, 462-477.

- Krantz, J.H. & Dalal, R. (2000). Validity of web-based psychological research. In: M.H. Birnbaum, *Psychological Experiments on the Internet.* pp. 35-60. Academic Press: USA

- Laurence, S., & Margolis, E. (2001). The poverty of the stimulus Aagument. *The British Journal for the Philosophy of Science,* **52**, 217-276.

- Mathews, R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J.R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **15**, 1083-1100.

- Miller, G.A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology,* **56**, 485-491.

- Newell, B.R. & Bright, J.E.H. (2001). The relationship between the structural mere exposure effect and the implicit learning process. *The Quarterly Journal of Experimental Psychology,* **54(4)**, 1087-1104.

- Perruchet, P. & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General,* **119**, 264-275.

- 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 structures 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. (1985). *The Penguin Dictionary of Psychology.* Penguin Press.

- Reber, A.S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General,* **118**, 219-235.

- Reber, A.S. (1993). *Implicit learning and tacit knowledge.* Oxford, England: Oxford University Press.

- 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.

- Reber, A.S. & Allen, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: A functionalist interpretation. *Cognition*, **6**, 189-221.

- Reber, A.S., Kassin, S., 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.

- Redington, M. & Chater, N. (1996). Transfer in artificial grammar learning: A re-evaluation. *Journal of Experimental Psychology: General,* **125**, 123-138.

- Reips, U.-D. (1995). *AAAbacus, the first cognitive learning experiment on the WWW* [WWW document]. URL: http://www.psych.unizh.ch/genpsy/Ulf/Lab/archiv/aaabacus.html

- Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages and solutions. In: M.H. Birnbaum, *Psychological Experiments on the Internet.* pp. 89-117. Academic Press: USA

- Servan-Schreiber, E. & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **16**, 592-608.

- Shanks, D.R., Johnstone, T., & Staggs, L. (1997). Abstraction processes in artificial grammar learning. *Quarterly Journal of Experimental Psychology,* **50A**, 216-252.

- Shanks, D.R. & St. John, M.F. (1994). Characteristics of dissociable human learning systems. *Behavioural and Brain Sciences,* **17**, 367-447.

- Shumway, R.J. (1983) Feature frequency and negative instances in concept learning. *American Educational Research Journal*, **20**, 451-459.

- St. John, M.F. (1996). *Computational differences between implicit and explicit learning: Evidence from learning crypto-grammars.* Unpublished manuscript.

- Stroulia, E. & Goel, A. (1996). A model-based approach to blame assignment: Revising the reasoning steps of problem solvers. In Proc. Thirteenth National Conference on Artificial Intelligence, pp. 959-964.

- Vokey, J.R. & Brooks, L.R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **18**, 328-344.

- Whittlesea, B.W.A. & Dorken, M.D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General,* **122**, 227-248.

- Zajonc, R.B. (1968). Attitudinal effects of mere exposures. *Journal of Personality and Social Psychology*, **9(2)**, 1–27.

Appendix A

A1 Servan-Schreiber & Anderson's theory of competitive chunking

Based on the results of their studies Servan-Schreiber and Anderson formulated a theory of competitive chunking (CC). In CC two things are known about every chunk: (1) what its immediate subchunks are, and (2) the chunk's strength, i.e. how often and how recently it has been used in the past. The strength construct in CC is identical to that in ACT* (Anderson, 1983) for declarative memory traces: A newly created chunk has a strength of one unit. Every time the chunk is used its strength is increased by one unit. A chunk's strength decreases with time however. Thus, at any point in time, the strength of a chunk is the sum of its successive individually decaying strengthenings:

$$\text{Strength} = \sum_{i} T_i - d$$

where T_i is the time elapsed since the ith strengthening, and d is the decay parameter ($0<d<1$). Once a chunk is created, it exists forever in long-term memory, and there is no bound on how much strength it can accumulate.

A1.1 The strengthening of existing chunks

According to CC the process proceeds from the simplest, elementary chunks, that is, those chunks that the system never had to learn, in a recursive cycle. When a stimulus is perceived the system matches the elementary chunks to the stimulus forming the elementary percept. Then the system matches its more complex chunks to this percept forming a new percept. The system continues to match more complex chunks to the percepts until it has no more chunks available. Thus, the number of chunks gets smaller with each cycle of the perception process.

For example, if we give the system the following stimulus:

Sa Da Ga Ba – Pa Ka Ta Ra

it would first perceive every letter, giving the elementary percept

S a D a G a B a – P a K a T a R a

Then it would use its more complex chunks to form the percepts

(Sa) (Da) (Ga) (Ba) (Pa) (Ka) (Ta) (Ra)

Then for example

((Sa) (Da)) ((Ga) (Ba)) ((Pa)(Ka)(Ta)) (Ra)

The number of chunks in the final percept is an important variable in CC called *nchunks*. *nchunks* is a measure of how compact the representation of a stimulus is, and thus, of how familiar the stimulus is perceived to be. In the above example the number of chunks in the elementary percept is 16, then 8, then 4 in the final percept.

Familiarity can be cautiously defined as

$$e^{1\text{-nchunks}}$$

where the familiarity of a stimulus ranges from 1 (maximally familiar) to an asymptotic 0 (maximally unfamiliar).

The probability that a chunk is retrieved depends exclusively on the strength of its subchunks, which Servan-Schreiber and Anderson call the *chunks support*. The average strength of a chunk's immediate subchunks is its support. A chunk's strength does not affect its own probability of being used but directly affects its superchunks' probabilities of being used. Thus, when the strength of a chunk decays, its superchunks are being forgotten, or conversely, when the strength of a chunk increases, its superchunks are being learned. In our example, if (Sa) and (Da) did not have enough strength, then their superchunk ((Sa)(Da)) may not be retrieved.

Whether a chunk is used and consequently strengthened, depends on the chunk's own strength, relative to the strength of its competitors (as there may be other chunks competing). Therefore both a chunk's strength and its support are critical to its being used.

A1.2 Creation of new chunks

The creation process follows on from the perception process. The final percept is the input and the output is a new chunk whose immediate subchunks are chunks in the final percept. If the final percept is for example

Chunk1 Chunk2 Chunk3

there are at least two potential new chunks that may be created: (Chunk1 Chunk2) and (Chunk2 Chunk3). The probability that a new chunk is actually proposed is assumed to be equivalent to the process of retrieving existing chunks for perceiving. The chunk with the largest support wins the chunk creation competition.

A1.3 Simulation

Servan-Schreiber & Anderson report two simulations of experimental data providing evidence for their theory of competitive chunking. The first simulation simulates Miller's (1958) data and illustrates the learning process, the second simulates their own data and illustrates how such a network can be used to perform the grammatical discrimination task.

In order for the simulation to provide data in the same form as the human data, the value of n-chunks was regarded as an act of rejection, i.e. the probability that a string would be rejected increased with the value of nchunks. The higher the value of nchunks, the less familiar a string was perceived to be, and the higher the probability that it would be rejected. In the example of above, if nchunks was 1 (i.e. the stimulus is encoded into a single chunk), the stimulus is maximally familiar and less likely to be rejected. If nchunks is 3 (i.e. the stimulus is encoded into three separate chunks), the stimulus is less familiar and thus more likely to be rejected.

Further assumptions were that letters were elementary chunks with the following constraints:

(a) the subchunks of a chunk must be adjacent (consistent with the Gestalt principle of proximity)

(b) The next level of chunks called “word chunks” can have at most 3 subchunks (reflecting the finding that the preferred word chunk size of Ss is 3)

(c) The next higher level, the "phrase chunks" can have at most 2 subchunks (so that length constraints become more severe as the complexity of chunks increases)

It was also assumed that each string of letters also includes “extremity markers” that signal the beginning and end of that string: the string TTXVPS was represented as “beginTTXVPSend” in the simulation. This was to help explain why Ss are apparently more sensitive to grammatical violations at the extremities of the string rather than those occurring in the middle (Reber, 1967; Servan-Schreiber & Anderson, 1990).

Many different values were tested for the two parameters c (competition parameter) and d (decay parameter). It was found that with the values of c=d =0.5 the coefficient of correlation between the simulated and the human data was 0.935. In addition the simulation was able to reproduce 87% of the variance.

The data from the simulation showed that forming chunks in the way Servan-Schreiber and Anderson suggest in their theory, was efficient enough to be able to discriminate between positive and negative letter strings. Ss were basing their classification on an overall familiarity measure. The more chunks present in the novel examples, the more likely for the string to be called “good”.