Journal of Experimental Psychology:
Learning, Memory, and Cognition
1991, Vol. 17, No. 5, 875-887

0278-7393/91/$3.00

Implicit and Explicit Knowledge Bases in Artificial Grammar Learning

Zoltan Dienes, Donald Broadbent, and Dianne Berry
University of Oxford, Oxford, England

Two experiments examined the claim for distinct implicit and explicit learning modes in the artificial grammar-learning task (Reber, 1967, 1989). Subjects initially attempted to memorize strings of letters generated by a finite-state grammar and then classified new grammatical and nongrammatical strings. Experiment 1 showed that subjects' assessment of isolated parts of strings was sufficient to account for their classification performance but that the rules elicited in free report were not sufficient. Experiment 2 showed that performing a concurrent random number generation task under different priorities interfered with free report and classification performance equally. Furthermore, giving different groups of subjects incidental or intentional learning instructions did not affect classification or free report.

There appear to be many examples in everyday life of people learning to respond appropriately according to criteria that can readily state, for example, in learning the rules of algebra. This, however, is not always so. There also appear to be cases of people learning to respond in some rulelike way without being able to say what the rules are that govern their behavior. For example, we learn to recognize and produce grammatical utterances without being able to say what the rules of grammar are.

Several authors have argued that people can learn complex tasks according to distinct implicit and explicit learning modes (e.g., Berry & Broadbent, 1984, 1988; Reber, 1967, 1989). The modes are distinguished both by the conditions that elicit them and by the type of knowledge that they result in. Implicit rather than explicit learning is claimed to occur, especially under incidental conditions and when the crucial information is nonsalient; purportedly, the resulting implicit but not explicit knowledge is largely unconscious or nonverbalizable (see, e.g., Reber, 1989).

One paradigm that has been used extensively to investigate the acquisition of implicit knowledge is artificial grammar learning (e.g., Mathews, Buss, Stanley, Blanchard-Fields, Cho, & Druhan, 1989; Reber, 1967, 1976, 1989; Reber & Allen, 1978; Reber & Lewis, 1977). In this paradigm, subjects typically memorize strings of letters that appear arbitrary but are actually generated by a finite-state grammar. Figure 1 shows a typical finite-state grammar. Subjects are then informed of the existence of the complex set of rules that constrain letter order and are asked to classify new grammatical and nongrammatical strings. Subjects' typical classification performance—about 65%—indicates that they have acquired substantial knowledge about the grammar.

This research was supported by the Economic and Social Research Council. We gratefully acknowledge Louis Manza, Robert Mathews, Pierre Perruchet, Arthur Reber, and an anonymous reviewer for valuable comments on the manuscript.

Correspondence concerning this article should be addressed to Zoltan Dienes, who is now at the School of Experimental Psychology, University of Sussex, Brighton, Sussex BN1 9QG England.

The purpose of this article is to examine two claims about the knowledge acquired in the artificial grammar-learning paradigm. First, Reber (1967, 1989) has claimed that a considerable portion of the knowledge is probably unavailable to consciousness. Presumably, this implies that the knowledge is difficult to elicit in some ways; that is, the classification knowledge is stored in a relatively specific data base. The strategy of this article will be to explore what tasks can and what tasks cannot elicit the knowledge to characterize it further. Reber (1989) has also hinted that even if implicit knowledge is found to be accessible to consciousness, it may in fact not normally be used consciously. In that case, criteria other than ultimate accessibility are needed to distinguish implicit from explicit knowledge. Such criteria may be suggested by a second claim made by a number of investigators (e.g., Broadbent, 1989; Hayes, 1987; Reber, 1989) that the knowledge is different from explicit knowledge in that it has distinctive properties of storage or retrieval. The strategy of this article will be to explore experimental manipulations that might plausibly influence the storage (or retrieval) of implicit and explicit knowledge differentially.

Previous researchers have investigated the first claim—that the knowledge is unconscious—by seeing if a task regarded as measuring explicit knowledge can elicit the knowledge used for classification performance. Free report is clearly a measure of explicit knowledge of the contents of the report, but it can plausibly be regarded as insensitive and incomplete compared with recognition or forced-choice measures (e.g., Brewer, 1974; Brody, 1989; Tulving, 1983; cf. Mathews, 1990). For example, Tulving (1983) presented evidence that free recall rather than recognition required more informational overlap between the stored trace and retrieval cues. Further, in the conditioning-without-awareness literature, it has been found that the subject may show knowledge of the experimental contingencies when asked specific questions even when such knowledge is not shown by free report (for a review, see Brewer, 1974). Whereas free report gives the subject the option of not responding, forced-choice questions do not. On the other hand, when forced-choice measures are used, it is not clear when the test is a test of implicit knowledge and when

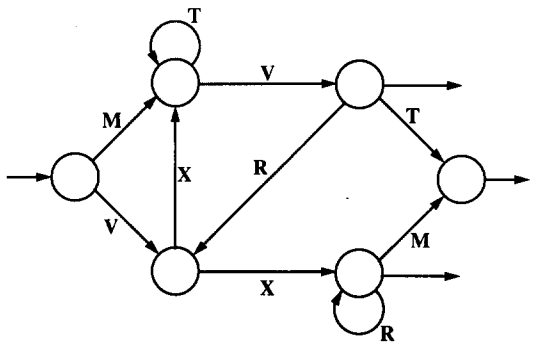

Figure 1. Example of a finite-state grammar.

it is a test of explicit knowledge (compare Dulany, Carlson, & Dewey, 1985, and Reber, Allen, & Regan, 1985). For example, classification performance is regarded as a measure of implicit knowledge in the artificial grammar-learning task (Reber, 1989), but Willingham, Nissen, and Bullemer (1989) regarded classification performance as a measure of explicit knowledge on a sequential reaction time (RT) task. Deductively, there are no correct answers to the question of whether a measure is implicit or explicit. The most productive way to proceed may be to characterize experimentally what the subjects can do with their knowledge of the artificial grammar by seeing what measures can elicit the knowledge. Presumably, to say that the knowledge is implicit or unconscious implies that there will be tasks that cannot fully elicit the knowledge.

Reber has often asked for introspections from subjects regarding classification (e.g., Allen & Reber, 1980; Reber & Allen, 1978; Reber & Lewis, 1977). However, he has rarely given a detailed analysis of the rules reported by subjects. Reber and Allen concluded that although subjects emerged with a small but solid body of articulated knowledge, they still could not tell all that they knew. Unfortunately, the appropriate analyses were not made to justify this conclusion. It is quite possible that if the rules elicited from subjects were applied to the strings, the predicted would match the actual classification performance. Reber and Lewis (1977) also compared subjects' rules and justifications with the subjects' classification performance. They cite several cases in which a subject retrospectively claimed a letter in a certain position was acceptable, but the subject correctly rejected a nongrammatical item in which the only violation was the letter in question. However, there is no evidence that the nongrammatical item did not violate another of the subject's rules, and there may be no inconsistency between stated rules and classification performance.

A more systematic procedure to investigate the validity of the subjects' freely stated rules was used by Mathews et al. (1989). Experimental subjects classified for 600 trials with feedback. After each 10-trial block, subjects were asked to give complete instructions on how to classify (the free-report measure of their knowledge). The validity of these instructions was assessed by the classification performance of yoked subjects who were requested to follow the transcribed instructions. The yoked subjects always performed substantially worse than experimental subjects. Although these results are suggestive, there are some problems with using yoked subjects to assess the validity of the rules stated by experimental subjects. If the instructions contain exemplars or parts of exemplars, implicit learning on the yoked subjects' part may lead to an overestimation of the explicit rule content of the instructions. Conversely, application errors by the yoked subjects may lead to an underestimation of the validity of the instructions. Stanley, Mathews, Buss, and Kotler-Cope (1989) showed that giving subjects a rule for performing a dynamic control task almost perfectly led to only 60% performance. Subjects could not have systematically applied the rule. A more systematic way of assessing the validity of rules elicited in free report would be to use them directly to simulate classification performance. This latter procedure will be used in the experiments reported in this paper (see also Druhan & Mathews, 1989).

As argued above, forced-choice measures may allow a more complete characterization of the subjects' knowledge than free report. Dulany, Carlson, and Dewey (1984) employed a forced-choice measure that did elicit the knowledge underlying classification of grammatical and nongrammatical strings. They asked subjects during classification to score that part of a string that made it right if it was classified as grammatical or that part that violated the rules if it was classified as nongrammatical. The mean validity of the features scored for each subject predicted proportion of correct classifications without significant error. Thus, the scoring and classification tasks tapped the same data base with about the same sensitivity.

Dulany et al.'s (1984) results help to characterize the subjects' knowledge without invalidating the claim that the knowledge might be implicit and thus difficult to elicit with other plausible knowledge measures. That is, Dulany et al.'s results do not exhaust all the ways in which the knowledge may be difficult to elicit. The classification task and the scoring task of Dulany et al. show that subjects are able to recognize the well-formedness of complete strings and their embedded parts. But perhaps subjects' knowledge can be elicited only by complete exemplars. In this sense, the knowledge may be implicit; perhaps only explicit knowledge can apply to isolated elements of exemplars. This characterization of the subjects' knowledge might partly correspond to Reber's view that the knowledge is "*implicit* in our sense that [subjects] are not consciously aware of the aspects of the stimuli which lead them to their decision" (Reber & Allen, 1978, p. 218).

The hypothesis that the subjects' knowledge applies only to complete exemplars has yet to be fully explored. Perruchet and Pacteau (1990) found that subjects could successfully rate the grammaticality of isolated bigrams, but they also showed that such bigram knowledge could not fully account for classification performance (Perruchet & Pacteau, Experiment 2; see also Mathews, 1990). Servan-Schreiber and Anderson (1990) also presented evidence that subjects' knowledge comes in chunks of letters, but they did not investigate whether these chunks could be accessed in isolation.

Experiment 1 of the present article tested subjects' knowledge of part strings that were not embedded within whole strings. Subjects were asked which letters could occur after

stems that varied in length from zero letters upwards (the test of sequential letter dependencies, or SLD test). By asking the subject to formulate general rules with reference to the presented constituents of exemplars, and out of the context of a particular exemplar, the SLD test might correspond to some notion of an explicit knowledge test (e.g., Brody, 1989; Eriksen, 1962; Perruchet & Pacteau, 1990). On the other hand, by employing a recognition measure, the SLD task might correspond to some notion of an implicit knowledge test (e.g., Reber et al., 1985). This ambiguity highlights the need to define empirically what subjects are able to do with their knowledge and thus to refine what might be meant by implicit or explicit knowledge.

The SLD test is an extension of the bigram rating test used by Perruchet and Pacteau (1990). Perruchet and Pacteau found that the bigram ratings were sufficient to account for classification performance of exemplars in which the position of the bigram did not affect its grammaticality. The SLD test differs from the bigram rating task in that it allows an assessment of subjects' knowledge of the positional dependence of bigrams. It is known that subjects can use such knowledge of positional dependence in classification performance (Perruchet and Pacteau, 1990, Experiment 2). It is thus important to know whether this information can be elicited out of the context of particular exemplars.

Experiment 1 also looked at another issue. In most concept formation tasks, subjects are typically exposed to both positive and negative exemplars, but in artificial grammar learning, subjects have typically been exposed only to positive exemplars. This may be an important factor in inducing an implicit learning mode.[1] Brooks (1978) did use two categories in an artificial grammar-learning task by employing two grammars simultaneously. Subjects associated particular grammatical exemplars with English words belonging to one of two classes. The grammar from which the exemplar was generated could be determined by the class of word with which it was associated; indeed, subjects could later discriminate exemplars from the two grammars using this cue. Reber and Allen (1978) argued that the Brooks technique did inhibit the normal implicit abstraction process. An important aspect of the Brooks technique is its emphasis on learning specific exemplars, but it remains plausible that providing a distinction between two categories may help induce a strategy that inhibits implicit learning. Thus, Experiment 1 employed two groups: One saw only grammatical exemplars, and the other saw both grammatical and nongrammatical exemplars, and the types of exemplar were distinguished by being presented in different colors.

Experiment 1

Method

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

Design. Subjects were randomly allocated to one of two groups: (a) the grammatical group that saw only grammatical exemplars or (b) the mixed group that saw both grammatical and nongrammatical exemplars.

Materials and apparatus. The grammar used was the one used by Dulany et al. (1984), Perruchet and Pacteau (1990), and Reber and Allen (1978) (see Figure 1). The 20 grammatical acquisition exemplars and the 50 grammatical and nongrammatical test exemplars were the ones used by Dulany et al. (1984) and Perruchet and Pacteau (1990; Experiments 1 and 3) (see Table 1). Twenty nongrammatical acquisition exemplars were created, also shown in Table 1. Five were taken from the nongrammatical test exemplars, and the remaining 15 were made by substituting an inappropriate for an appropriate letter in an otherwise grammatical string. The position of violation covered letter positions one to six over the 15 exemplars.

During the acquisition phase, each exemplar was displayed on a color monitor by a Sinclair ZX Spectrum for 5 s, and the total set of exemplars was presented six times in a different random order each time. Randomization was constrained to avoid making the grammar salient. For the grammatical group, only the 20 grammatical acquisition exemplars were displayed. For the mixed group, all 40 acquisition exemplars were displayed. The grammatical items were displayed in black, and the nongrammatical exemplars were displayed in red; grammatical and nongrammatical exemplars alternated.

For the test of sequential letter dependencies, all possible grammatical stems were generated of length zero to five, with the constraint that the possible exemplars that were based on the stem could be no more than six letters long and that it must be possible for at least one letter to follow the stem. This produced a total of 32 stems, including the null stem. The stems were ordered so that previous stems did not contain later stems. Each stem was displayed in black by the Spectrum.

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

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

For the classification phase, the grammatical subjects were informed that the order of letters in each item was determined by a complex set of rules; the mixed subjects were informed that the order of letters in the black items followed a complex set of rules but that the order in the red items broke those rules in some way. All subjects were then told that they would now see some more items, only half of which followed the rules, and they were to decide which items followed the rules. Each exemplar was displayed in black until the subject pressed [1] to indicate grammatical or [0] to indicate nongrammatical. The 50 test exemplars were repeated once in a different random order.

After classifying all the items, subjects were given the free-report test and then the SLD test. In the free-report test, subjects were asked to indicate how they decided whether an item followed the rules, any strategies they used, and any rules that they thought the (black) items followed, even if they were not confident as to the correctness of the rules. Subjects were also asked to indicate any specific exemplars they could recall. Subjects were urged to be as complete as possible. Subjects' responses were recorded on tape. In the SLD test, subjects were shown stems, and the experimenter probed with possible next letters (M, V, X, R, and T). To each letter the subject said yes or no

[1] Indeed, natural language acquisition appears to occur largely by exposure to positive evidence (Brown & Hanlon, 1970). However, this situation may bear no necessary relation to adults learning finite-state grammars.

Table 1
Exemplars Presented in Acquisition and Test Periods

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

Note. Gramm. = grammatical; Nongramm. = nongrammatical.

and gave a confidence rating on a 5-point scale on which 1 indicated a guess and 5 indicated certainty.

Results

Classification performance. The proportions of items judged correctly by the grammatical and mixed groups were .65 (SE = .02) and .60 (SE = .01), respectively. The groups differed significantly, $t(38) = 2.35$, $p < .05$. These proportions are comparable to the proportion correct that was obtained by Dulany et al. (1984) (.63 for the implicit-sequential group).

For the two presentations of each exemplar, the mean proportions of judgments that were correct–correct (CC), error–correct (EC), correct–error (CE), error–error (EE), and the average of the two mixed cases (AV) are displayed in Table 2.

A 2 × 2 (Group [grammatical vs. mixed] × Error Type [EE vs. AV]) mixed-model analysis of variance (ANOVA) indicated significant main effects of group, $F(1, 38) = 4.35$, $p < .05$, and of error type, $F(1, 38) = 18.74$, $p < .001$. That is, the mixed rather than the grammatical group made a greater number of both error types. Also, subjects made more error–error than mixed error types, as found by Dulany et al. (1984).

SLD test. The proportions of correct responses to the SLD test by the grammatical and mixed groups were .64 (SE = .02) and .60 (SE = .01), respectively. The difference between the groups was marginally significant, $t(38) = 1.88$, p = .07.

Questions were classified according to the letter position probed, from the first to the sixth position. The mean proportions correct from the first to the sixth position were .68, .68, .72, .63, .61, and .59. We performed t tests that indicated that proportions correct for all positions were significantly above chance ($ps < .005$). Because a proportion correct of 1.00 required saying yes about 40% of the time, the proportion correct for each question type was compared with the expected chance proportion correct given the overall response bias of the subject and the particular response bias required by the question. This did not change the above results.

An analysis was conducted to determine whether knowledge of the positional dependence of bigrams could be elicited by the SLD test. There were 19 questions referring to admissible bigrams that were nongrammatical in the given position (bigrams at the beginning of the string were not included) and 60 questions referring to admissible bigrams that were grammatical in the given position. If subjects acquire knowledge of admissible bigrams but not their positional dependence, the tendency to respond "grammatical" should be the same in both cases. A 2 × 2 (Group [grammatical vs. mixed] × Question Type [nongrammatical vs. grammatical admissible bigram]) mixed-model ANOVA on proportion of grammatical responses indicated a significant effect only for question type, $F(1, 38) = 23.53$, $p < .0001$. The proportion of grammatical responses for nongrammatical and grammatical admissible bigrams was .54 and .68, respectively. That is, subjects were sensitive to the position of bigrams. The above proportions were compared with the subjects' tendency to respond "grammatical" over the whole SLD test (M = .54). This analysis indicated that performance on admissible nongrammatical bigrams was not significantly different from chance, $F < 1$ but that performance on admissible grammatical bigrams was, $F(1, 38) = 174.62$, $p < .0001$.

Classification performance and SLD. The Spearman's within-groups correlation across subjects between classification performance and proportion of questions correct on the SLD test was .29, p = .07 (a within-groups correlation was used to detect any association independent of that already demonstrated by the between-groups effects).[2]

Is the level of knowledge elicited by the SLD test sufficient to account for the level of classification performance achieved by subjects? Comparing the proportions correct on both tasks would not be an adequate way of answering this question, because proportion of correct answers on the SLD test is influenced by the subject's bias to answer yes. Two further measures were calculated to overcome the problem of bias: d' and a predicted performance that is based on the SLD test. The problem with d' is that a more difficult discrimination may be involved with one task rather than with the other simply because of the distractors chosen by the experimenter for that task; a problem with predicted performance is that it

[2] The within-groups correlation was calculated by subtracting the group mean from each subject's scores and then calculating the Spearman's correlation in the normal way.

Table 2
Consistency of Judgments

	Group				
	Gramm.		Mixed		
Judgment	*M*	*SD*	*M*	*SD*	Dulany et al.[a]
CC	0.50	0.08	0.44	0.08	0.47
EC	0.14	0.06	0.17	0.06	0.14
CE	0.16	0.05	0.16	0.06	0.18
AV	0.15	0.04	0.16	0.04	0.16
EE	0.21	0.08	0.23	0.06	0.21

Note. CC = correct–correct; EC = error–correct; CE = correct–error; EE = error–error; AV = mixed cases; Gramm. = grammatical.
[a] The means for Dulany et al. (1984) are for the implicit-sequential group.

might be based on a transformation not actually used by subjects. The approach taken in this article will be to provide converging evidence that is based on both measures.

The *d'* was calculated for each subject for both classification performance and SLD performance. If both the classification and SLD tasks elicited equivalent levels of knowledge, then the *d'*s for the two tasks should be similar. A 2 × 2 (Group [grammatical vs. mixed] × Task [classification vs. SLD]) mixed-model ANOVA indicated only a significant effect of group, $F(1, 38) = 9.22$, $p < .005$. That is, *d'* was higher for grammatical (0.76) rather than mixed (0.54) subjects. The *F* for task was 3.88, which was marginally significant, $p = .06$. That is, subjects tended to have greater *d'*s for the SLD (0.70) rather than classification task (0.60). The *F* for the interaction was 0.27. The Spearman's within-groups correlation between the *d'*s for the classification and SLD tasks was .29, $p = .07$.

A second method used to determine if the classification and SLD tasks elicited equivalent levels of knowledge was to apply a transformation to the SLD responses to yield a predicted classification performance. For each test exemplar for each subject, an average SLD response was calculated by adding the subject's confidence on the SLD test of each successive letter in the exemplar if the subject regarded the letter as grammatical, subtracting the confidence if the subject regarded the letter as nongrammatical, and then dividing by the number of letters in the exemplar. Call this final figure the sum for the exemplar.[3]

To calculate a predicted performance for each subject (PPSUM), an exemplar was classified as grammatical if the sum for that exemplar was greater than the mean sum (for that subject) and as nongrammatical if the sum was lower than the mean sum. The means of PPSUM for the grammatical and mixed groups were 0.63 ($SE = 0.02$) and 0.59 ($SE = 0.02$). A 2 × 2 (Group [grammatical vs. mixed] × Performance Type [PPSUM vs. performance]) mixed-model ANOVA indicated only a significant effect of group, $F(1, 38) = 6.35$, $p < .05$. That is, both PPSUM and classification performance were lower for the mixed rather than for the grammatical group. The *F* for performance type was 0.91, and the *F* for the interaction was 0.06.

Classification performance and free report. The rules elicited by free report were used to classify each exemplar in turn, to produce a predicted performance (PPFR) for each subject. Only rules that could be clearly applied were used; for example, "See if sounding the string out makes a word" was not used because no strings exactly specified a word. *Rules* were statements that could be used to either assign grammatical or nongrammatical status to an exemplar. If no rule applied to an exemplar, it was assigned a probability correct of .5. The mean values for PPFR[4] for the grammatical and mixed groups were 0.55 ($SE = 0.01$) and 0.52 ($SE = 0.01$), respectively. A 2 × 2 (Group [grammatical vs. mixed] × Test [PPFR vs. performance]) mixed-model ANOVA indicated significant main effects of group, $F(1, 38) = 10.70$, $p < .005$, and of test, $F(1, 38) = 50.02$, $p < .001$. That is, the grammatical rather than the mixed group was better on both tests. Also, the rules elicited from free report underpredicted actual performance. The within-groups Spearman correlation between PPFR and performance was .03; this correlation did not differ significantly between groups.

Control subjects. Six control subjects were run only on the SLD test to determine baseline performance. The proportion of correct responses was .50 ($SE = .03$), not different from chance. The mean *d'* and PPSUM were 0.06 ($SE = 0.04$) and 0.53 ($SE = 0.02$), respectively. Neither of these measures was significantly different from chance.

Discussion

The results of Experiment 1 addressed two main questions: (a) Can performance be accounted for by either free report or the SLD test? (b) Is the type or quantity of learning affected by the presence of nongrammatical exemplars?

Three results pointed to a correspondence between classification performance and ability to answer the SLD test. First, there was a positive correlation between classification performance and correct responses on the SLD test. Second, subjects' *d'*s for the classification and SLD tasks were similar. (The *d'*s were marginally higher for the SLD rather than for the classification task, but, to anticipate, Experiment 2 indicated that this effect was not reliable.) Third, there was a close

[3] For the nongrammatical exemplars, a procedure was required for dealing with letters subsequent to the position of the nongrammatical letter because the SLD test contained only grammatical stems. The procedure adopted was to ignore the letter immediately after the nongrammatical letter and thereafter treat the stem as the corresponding grammatical stem. This procedure should, if anything, underestimate the assessed nongrammaticality of a nongrammatical exemplar. Only the 19 nongrammatical exemplars produced by a single substitution were used. To balance numbers, only the first 19 grammatical exemplars listed in Table 1 were used.

[4] Three independent judges calculated PPFR scores for 20 subjects. The intraclass correlation was .88, which indicates that the scoring procedure was highly reliable. (Note that the intraclass correlation is sensitive to absolute differences between the judges as well as different relative orderings.)

match between classification performance and predicted performance that was based on answers to the SLD test. To calculate a predicted performance, we constructed a model in which subjects classified on the basis of their average SLD response for each exemplar. There was no significant difference between classification performance and predicted classification performance; the drop in classification performance of the grammatical as compared with the mixed group was matched by an identical drop in predicted classification performance. Thus, there is evidence that the SLD test can access the knowledge underlying classification performance with about the same degree of sensitivity as classification performance. This result conceptually replicates Perruchet and Pacteau's (1990; Experiment 3) finding that bigram ratings could predict classification performance with these same strings. If rating bigrams could predict classification performance, then the SLD task should do so as well.

Because the knowledge base can be adequately accessed by the SLD test, it is clear that subjects can recognize the well-formedness of elements of exemplars in isolation. This is shown most strikingly by the high levels of performance on the SLD test for stems of length zero and one ($M = .68$), where elements are maximally isolated. Also, increasing the length of the stem does not benefit SLD performance. In these comparisons, stem length is confounded with letter position; that is, later rather than earlier positions provide more information. Nonetheless, the results do rule out the hypothesis that the presence of a complete exemplar is necessary to elicit knowledge of the grammar.

The data from the SLD test also shed light on the accessibility of the associations formed by subjects. Perruchet and Pacteau (1990) found that subjects' ratings of isolated bigrams could account for bigram knowledge in classification performance, but whether subjects' ratings of isolated elements could elicit subjects' knowledge of the positional dependence of bigrams was left open. The results of the current experiment indicated that ratings on the SLD test could elicit knowledge of positional dependence.

Experiment 1 found that free report did not elicit the same amount of knowledge as classification performance, which supports the claims of Reber (1989) and Mathews et al. (1989). Also, free-report performance did not correlate with classification performance. This failure of elicitation is suggestive evidence for an implicit knowledge base underlying classification performance, although an explanation in terms of a single explicit memory system is possible. A subject may have to remember a relatively large amount of information in the artificial grammar-learning task. Note that classification involves recognition or lack of recognition of the presented associations. If recognition is seen as instigating a retrieval process that was possible but less reliable under free report (see, e.g., Brody, 1989; Davis, Sutherland, & Judd, 1961; Tulving, 1983), then the difference between classification performance and free report would be expected. Also note that the free-report technique used in Experiment 1 required only a single-shot retrospective report. In sum, although the inadequacy with which free report accesses the knowledge base is suggestive evidence of implicit knowledge, the finding can be reexplained in terms of a single knowledge type—explicit knowledge. It is clear that further evidence should be sought and the replicability of the current evidence determined.

The presence of nongrammatical exemplars interfered with both performance and free report. Subjects reported that they found it hard to remember what appeared in black and what in red, and so they may have confused correct and incorrect information. Did this interfere with both implicit and explicit learning? If free report is taken as a measure of explicit learning and performance as a measure of implicit learning, then the presence of contrast did not appear to affect implicit or explicit learning differentially: The Group × Test Type (performance vs. free report) interaction was not significant (though perhaps a greater influence of contrast on free report was hidden by a floor effect).

Experiment 2

The aim of Experiment 2 was to investigate the possibility of distinct implicit and explicit types of knowledge associated with artificial grammar learning by using a dual-task methodology. If implicit and explicit knowledge are of a different type, then it should be possible to influence one but not the other with a suitable manipulation.

A suitable manipulation is suggested by the results of Hayes (1987; and in Broadbent, 1989); he found that random number generation (RNG) interfered with artificial grammar learning when subjects were given intentional learning instructions but did not interfere under standard incidental memory instructions. One interesting interpretation of these results is that RNG interferes with the formation of explicit but not implicit knowledge, so under standard incidental instructions, RNG might interfere with tests of explicit but not implicit knowledge.

In a pilot experiment, 12 subjects were instructed to generate random numbers while they were memorizing grammatical exemplars. The aim was to see if this differentially interfered with subsequent free report rather than with classification or SLD performance. In fact, all knowledge measures were significantly lower in the pilot study than in the grammatical group of Experiment 1 (all $ps < .05$): Classification performance was 0.56, the proportion of correct responses on the SLD test was .53, and the PPFR was 0.52.

Thus, Hayes's (1987) finding of a lack of interference of RNG on performance under standard memory instructions was not replicated. One key procedural difference between the pilot study and Hayes was that Hayes did not explicitly inform subjects of the priority to be given to the two tasks; in the pilot study, subjects were told to give the RNG task priority. Hayes might have subtly communicated different priorities to the incidental and intentional groups. Alternatively, given ambiguous instructions, subjects may have adjusted priorities according to the perceived difficulty of the grammar-learning task. When subjects were asked to test hypotheses under dual-task conditions, they may have focused on only one or two attributes of each string to comply with the experimenter's instructions. When subjects were asked to

memorize seemingly arbitrary sets of letters, they may have attempted to take in all attributes of each string to comply with the experimenter's instructions. According to this view, only the subjects given intentional instructions in the Hayes study fully engaged in the RNG task, and this interfered with all types of learning.

A proper test of this view requires that instructions and priorities are orthogonally manipulated. Thus, Experiment 2 investigated the effect of single- versus dual-task conditions and of priority instructions on incidentally and intentionally instructed subjects. If priority has an effect on classification performance, then the results of Hayes (1987; in Broadbent, 1989) contain a potential artifact, but, even so, if the effect of priority is different for incidentally and intentionally instructed subjects, there is still good evidence for separate implicit and explicit learning modes. If priority has no effect on classification performance, then the absence of dual-task interference in the incidental rather than the intentional conditions in Hayes' study cannot be based on a priority artifact. In this case, the comparison between single- and dual-task subjects allows a procedural replication both of Hayes and of the pilot study to determine the replicability of both sets of results.

In addition to assessing the possibility of different learning modes, Experiment 2 also allows an assessment of the possibility of different knowledge types. If effects of priority or dual-versus single-task are greater for free report than for classification performance and SLD, there would be evidence for different implicit and explicit knowledge types.

The data from Experiment 2 could also shed light on the nature of the associations formed under dual- rather than single-task conditions. Cohen, Ivry, and Keele (1990) argued that diverting attention from learning a structured sequence interfered with learning the positional dependence of bigrams but not with learning bigrams per se. Cohen et al. used an RT task that involved responding to the location in which a stimulus appeared; Experiment 2 explored the generality of this effect with the artificial grammar-learning paradigm.

Method

Design. Experiment 2 used a 2 × 3 (Instructions [incidental vs. intentional] × Condition [single task vs. dual task with grammar high priority vs. dual task with grammar low priority]) between-subjects design, with an equal number of subjects in each of the six cells.

Subjects. The subjects were 60 paid volunteers, aged between 18 and 45, from the Oxford University subject panel. No subject had participated in previous grammar-learning experiments.

Materials and apparatus. These were the same as those used in Experiment 1 (grammatical group).

Procedure. All subjects were exposed to the exemplars with the same displays used in Experiment 1 (grammatical group) and then performed the classification, free report, and SLD tasks, respectively. Half of the subjects performed 5 min of RNG alone before being exposed to the exemplars, and half of the subjects performed 5 min of RNG alone after the SLD task. No further RNG was performed by the single-task subjects.

For the RNG task, a metronome was set to give a click every 2 s; subjects were told to give a digit between 0 and 9 every time they heard a click and to make the sequence of digits as random as possible. It was pointed out that each digit should on average occur equally often and be equally likely to follow any digit. Subjects were discouraged from repeating well-learned sequences such as 12345 or telephone numbers. It was suggested that they could imagine a hat containing 10 pieces of paper, 1 for each digit, and that at every click, they could imagine drawing a piece of paper out of the hat, reading it, and then replacing it.

The dual-task subjects performed RNG while being exposed to the exemplars and were given priority instructions as to which task to emphasize. They were told to concentrate on the primary task, to perform it as well as they could, and not to let the secondary task interfere; they were to attend to the secondary task to the extent that they were performing the primary task as well as they could.

Incidental subjects were given the same instructions that were used in Experiment 1; they were simply asked to memorize the exemplars. In addition, intentional subjects were asked to search for rules. The instructions were taken verbatim from Reber (1976):

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

Results

Classification performance. A 2 × 3 (Instruction [incidental vs. intentional] × Condition [single task vs. grammar learning–high priority vs. grammar learning–low priority]) ANOVA on classification performance indicated only an effect of condition, $F(2, 54) = 5.08$, $p < .01$. The F for instruction was 1.67, and the F for the interaction was 1.02. The effect for condition was further analyzed by means of two orthogonal contrasts. The first contrast tested for a dual- versus single-task effect by comparing the classification performance of the single-task groups (0.69) with the average of the low- and high-priority groups (0.62), $F(1, 54) = 10.00$, $p < .01$. That is, dual- rather than single-task conditions interfered with classification performance. The second contrast compared the high- (0.63) and low-priority groups (0.62), $F(1, 54) < 1$. In short, performing under dual- rather than single-task conditions interfered with classification performance, but there was no effect of priority and no interaction of condition with instructions.

A 2 × 3 (Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority]) ANOVA on d' yielded similar results. The ANOVA indicated only an effect of condition, $F(2, 54) = 5.69$, $p < .01$. The F for instruction was 1.45, and the F for the interaction was 1.69. The mean d's were 0.93 for single-task conditions, 0.60 for high-priority conditions, and 0.54 for low-priority conditions.

The mean proportions of judgments that were incorrectly classified on both presentations (EE) or on only one (AV) are displayed in Table 3. A 2 × 2 × 3 (Error Type [EE vs. AV] ×

Table 3
Consistency of Judgments

	Condition					
	Single task		High priority		Low priority	
Instructions	*M*	*SD*	*M*	*SD*	*M*	*SD*
Implicit						
EE	0.16	0.08	0.19	0.07	0.21	0.06
AV	0.15	0.03	0.15	0.06	0.16	0.06
Explicit						
EE	0.18	0.10	0.25	0.08	0.24	0.06
AV	0.14	0.04	0.20	0.04	0.15	0.04

Note. EE = error–error; AV = mixed cases.

Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority)] mixed-model ANOVA indicated significant main effects of error type, $F(1, 54) = 22.80$, $p < .0001$, and of condition, $F(2, 54) = 3.92$, $p < .05$. That is, subjects made more error–error than mixed–error types, as found in Experiment 1, and by Dulany et al. (1984). The effect of condition was not analyzed further.

SLD test. A 2 × 3 (Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority]) ANOVA on the proportions of correct responses to the SLD test indicated only an effect of condition, $F(2, 54) = 12.83$, $p < .0001$. The F for instruction was 0.05, and the F for the interaction was 0.46. As for classification performance, the effect for condition was further analyzed by means of two orthogonal contrasts. The contrast for a dual- (0.66) versus single-task (0.57) effect was highly significant, $F(1, 54) = 25.40$, $p < .0001$. That is, dual- rather than single-task conditions interfered with SLD responding. The contrast for priority (0.57 vs. 0.56) was nonsignificant, $F(1, 54) = 0.15$. In summary, the results mirrored those for classification performance.

We conducted t tests that indicated that all of the proportions correct for SLD questions at all positions probed differed significantly from chance, $ps < .0005$.

An ANOVA was conducted on the positional dependence of subjects' bigram knowledge. A 2 × 3 × 2 (Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority] × Question Type [nongrammatical vs. grammatical admissible bigram]) ANOVA on proportion of grammatical responses indicated a marginal effect of condition, $F(2, 54) = 2.76$, $p = .07$, a marginal effect of question type, $F(1, 54) = 3.21$, $p = .08$, and a marginal interaction between condition and question type, $F(2, 54) = 2.40$, $p = .10$. The interaction was analyzed further by orthogonal partial interactions. The effect of question type (i.e., sensitivity to the positional dependence of bigrams) was greater under single-task (0.65 for nongrammatical and 0.75 for grammatical admissible bigrams) than under dual-task conditions (0.64 for nongrammatical and 0.65 for grammatical admissible bigrams), $F(1, 54) = 3.91$, $p < .05$, one-tailed. The effect of question type did not differ under low- rather than high-priority conditions, $F < 1$. In short, as found by Cohen et al. (1990) for a RT task, sensitivity to the positional dependence of bigrams was disrupted under dual- rather than single-task conditions. Under dual-task conditions, there was no sensitivity to the positional dependence of bigrams, $F < 1$, but there was sensitivity under single-task conditions, $F(1, 54) = 6.49$, $p < .05$.

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

The d's for the SLD and classification tasks were very similar (overall, 0.70 vs. 0.69). A 2 × 2 × 3 (Task [SLD vs. classification] × Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority]) mixed-model ANOVA indicated only a significant effect of condition, $F(2, 54) = 14.74$, $p < .0001$. The Spearman's within-groups correlation between d' for the SLD and classification tasks was .44, $p < .01$.

The SLD responses were transformed as in Experiment 1 to produce a sum for each exemplar for each subject. A 2 × 2 × 3 (Performance Type [PPSUM vs. actual classification performance] × Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority]) mixed-model ANOVA indicated only a significant effect of condition, $F(2, 54) = 7.81$, $p = .001$. The effect for condition was further analyzed by means of the two orthogonal contrasts. The contrast for a dual- versus a single-task effect was highly significant, $F(1, 54) = 19.75$, $p < .001$. That is, dual- rather than single-task responding interfered with both knowledge measures (0.67 vs. 0.60 for PPSUM, and 0.69 vs. 0.62 for classification performance). The contrast for priority was nonsignificant (0.60 vs. 0.61 for PPSUM, and 0.63 vs. 0.62 for classification performance), $F(1, 54) = 0.22$.

Classification performance and free report. A 2 × 2 × 3 (Test Type [PPFR vs. actual classification performance] × Instruction [incidental vs. intentional] × Condition [single-task vs. grammar learning–high priority vs. grammar learning–low priority]) mixed-model ANOVA indicated significant effects of test type, $F(1, 54) = 160.83$, $p < .0001$, and of condition, $F(2, 54) = 7.81$, $p = .001$. That is, PPFR (0.54) substantially underpredicted actual classification performance (0.65). The effect of condition was analyzed by orthogonal contrasts. The contrast for single versus dual task was highly significant, $F(1, 54) = 13.25$, $p < .005$. That is, dual- rather than single-task conditions interfered with both knowledge measures (0.52 vs. 0.57 for PPFR). The contrast for priority was nonsignificant, $F(1, 54) = 0.02$.

The Spearman's within-groups correlation between PPFR and classification performance was 0.32, $p < .05$. This correlation was smaller and not significant in Experiment 1 (but had the same magnitude in the pilot study). In Experiment 2, the correlation was just as strong under single-task compared with dual-task conditions (.47 compared with .20) and just as strong under incidental compared with intentional conditions (.26 compared with .35). The lack of correlation in Experiment 1 may, therefore, have been due to lack of power.

Random number generation. The measures of randomness were first-order entropy (corrected according to Miller,

1955), and first-, second-, and third-order ϕ;[5] and also, as recommended by Baddeley (1966), the proportion of arithmetic sequences, the proportion of bigrams repeated at least once, and the number of digits produced.

A Hotelling's T^2 comparing single- and dual-task conditions was 363.81, $F(6, 31) = 52.21$, $p < .0001$. Four measures were univariately significant, all $ps < .0001$. Subjects under dual- rather than single-task conditions produced relatively more repetitions than alternations at the second and third order of dependency (consistent with Truijens, Trumbo, & Wagenaar, 1976), produced more arithmetic sequences, and used the same bigrams more often.

Considering now only data collected under dual-task conditions, the Hotelling's T^2 for priority was 17.03, $F(7, 30) = 2.03$, $p = .084$. Two measures were univariately significant, $ps \leq .05$: When subjects were asked to give the RNG task greatest priority, they produced more digits and distributed them over the 10 response categories more evenly. The multivariate and univariate analyses indicated no significant effect for instructions or for the Instructions × Priority interaction.

Possible trade-off between classification and RNG performance was further investigated by the correlations between classification and RNG performance under dual-task conditions, with single-task RNG performance partialed out to control for preexisting ability factors[6] affecting both tasks. Only one of 21 correlations achieved significance at the .05 level.

Discussion

The aim of Experiment 2 was to investigate the possibility of different modes of learning and different types of knowledge in artificial grammar learning by systematically exploring the influence of concurrent random number generation. Experiment 2 provided data relevant to the influence of different task priorities under dual-task conditions and to the influence of performing under dual- versus single-task conditions. Experiment 2 also allowed an attempted replication of the important findings of Experiment 1; this last issue will be dealt with first.

As in Experiment 1, the results of Experiment 2 indicated a correspondence between classification performance and ability to answer the SLD test. There was a significant correlation between classification performance and correct responses on the SLD test. The d's for the two tasks were similar. Further, there was a close match between classification performance and predicted performance that was based on answers to the SLD test, using a linear transformation. Also, there was a close matching of average PPSUM and classification performance across groups with different levels of classification performance. Because Experiment 2 used a different manipulation than Experiment 1 to influence classification performance, the close matching of classification performance and PPSUM in Experiment 2 considerably strengthens the case for both the psychological validity of the linear transformation and the sensitivity of the SLD test to the classification knowledge base.

As in Experiment 1, free report did not fully elicit the knowledge underlying classification performance. This provides tentative evidence that classification performance relied on implicit knowledge. However, in this experiment, in contrast to Experiment 2, free report did correlate with classification performance. To investigate the possibility of different knowledge types further, Experiment 2 explored the effect of a dual task on the different knowledge measures.

No measure of artificial grammar learning—neither classification performance, SLD responding, nor free report—showed an effect of priority under dual-task conditions. This was the case even though the priority manipulation was effective in changing subjects' RNG performance. The effect of priority on only one of the tasks was confirmed by the absence of significant correlations between dual-task grammar learning and RNG performance. Thus, the apparent decrease in resources applied to RNG by subjects asked to give RNG low rather than high priority was not matched by an increase in resources effectively applied to artificial grammar learning. One possibility is that although subjects attempted to apply more resources to artificial grammar learning, they simply did not know how to do so effectively. This possibility is consistent with the finding, discussed below, that intentional rather than incidental instructions had no impact on subjects' performance: Knowing that there are rules to be found and attempting to find them did not benefit effective acquisition of the rules.

In terms of interpreting the results of Hayes (1987; and in Broadbent, 1989), the lack of effect of priority on artificial grammar learning implies that Hayes's results were not subject to a priority artifact. In fact, the power of the current design for detecting a priority effect on classification performance as large as the effect found by Hayes that was due to incidental versus intentional instructions under dual-task conditions was 0.85. However, the data from Experiment 2 do suggest why Hayes, unlike the current experiment, failed to find a dual-task effect for incidental subjects. The means from Experiment 2 were used to estimate the population difference in classification performance between single- and dual-task conditions (seven exemplars). An analysis indicated that the procedure of Hayes (1987) had a power of only 0.3; that is, it is more likely than not that Hayes's procedure would fail to detect the effect of the dual task on incidentally instructed subjects.

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

Analysis of the SLD responses suggests how the dual- rather than single-task conditions interfered with forming associations in the grammar-learning task. Specifically, under single-task conditions, subjects were sensitive to the positional dependence of bigrams, as in Experiment 1 (and Perruchet &

[5] For the calculation of ϕ, see Wagenaar (1970), and of first-order entropy, see Attneave (1959).

[6] Except for a general time-sharing ability.

Pacteau, 1990, Experiment 2). However, under dual-task conditions, subjects were not sensitive—subjects were as likely to respond "grammatical" to an admissible bigram in a grammatical as in a nongrammatical position. These results with the artificial grammar-learning paradigm parallel the findings of Cohen et al. (1990) with their RT task; namely, dual-task conditions interfere with learning the positional dependence of bigrams. Interestingly, Baddeley (1986) regarded the articulatory loop as particularly involved in coding order information in complex stimuli. The relation of the articulatory loop to the coding of the positional dependence of bigrams in artificial grammar learning and on repeating RT tasks (Cohen et al. 1990) requires further investigation.

The results of Experiment 2 are consistent with a single mode for learning artificial grammars. An early result by Reber (1976) indicated that giving subjects intentional rule-search instructions, rather than incidental memory instructions, deteriorated classification performance. Although this suggested that subjects could approach artificial grammar learning in distinct implicit or explicit modes, Reber, Kassin, Lewis, and Cantor (1980)[7] and several more recent studies have failed to find an effect of intentional instructions on nonsalient stimuli (Dulany et al., 1984; Hayes, 1987; Mathews et al., 1989; Perruchet & Pacteau, 1990). Experiment 2 provided data consistent with the latter studies: Asking subjects to approach the task in an intentional rather than an incidental manner had no effect on the classification and SLD tasks; further, it did not lead to any better (or worse) explicit knowledge, as indexed by free report. (The absence of an instructional effect on the SLD test is consistent with Perruchet and Pacteau's [1990] finding that there was no instructional effect on their bigram rating task.) Note that Experiment 2 could detect a population difference in classification performance between incidental and intentional subjects of the size found by Reber (1976) with a power greater than 0.98. Thus, the failure to replicate Reber can be accepted as valid with some confidence, especially in light of the consistency of this null result with all the later studies. We summarize the evidence with respect to learning modes by saying that both the absence of an interaction between incidental versus intentional instructions and dual-task conditions, discussed previously, and the absence of a main effect of incidental versus intentional instructions are consistent with subjects approaching artificial grammar learning with a single learning mode.

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

General Discussion

This article reported two experiments that compared classification performance with a structured knowledge test—the SLD test—and with free report, over a range of experimental manipulations. Interest focused on the extent to which classification knowledge was elicited by the SLD task, whether distinct learning modes could be applied to artificial grammar learning, and the extent to which the knowledge underlying classification performance could be regarded as implicit. These issues are discussed in turn.

Classification Performance and SLD

Both experiments indicated a positive relation between classification performance and number of correct SLD responses; the Spearman's correlation across subjects over all experiments was .42, $p < .001$.[8] Similarly, the Spearman's correlation over all experiments between d' on the SLD and classification tasks was .52, $p < .001$ (see Footnote 8). Thus, there is evidence that the knowledge tests tapped the same knowledge base. But did they do so equally efficiently?

Evidence for the classification and SLD tasks accessing the same knowledge base with equal efficiency is provided by the similar d's for the two tasks over all the experiments: 0.61 and 0.60, respectively. A PPSUM was calculated for each subject. The PPSUM was based on the subject's average SLD response to each sum. For each subject, the Spearman's correlation over exemplars between the sum and the tendency to say "grammatical" on the classification task was calculated. These correlations were transformed according to Fisher's z. The mean value (converted back to a correlation) was .24, which was different from zero, $t(111) = 11.04$, $p < .0001$. That is, each subject's responses on the SLD test predicted which items that subject said "grammatical" to. The correlation is small, but the more guessing the subject engages in on either the SLD or the classification task, the smaller the correlation would be (the small difference between EE and AV is evidence for guessing on the classification task; see Reber, 1989). If the correlation is small simply because of guessing, then it should be substantially increased by averaging over larger numbers of responses. Indeed, the Spearman's correlation over the nine group means for all experiments between PPSUM and classification performance was .91, $p < .01$, and not significantly different from 1.0. In all experiments, PPSUM closely matched actual classification performance across a range of manipulations that affected classification performance. That is, the SLD test could access the classification knowledge base with a sensitivity equal to classification performance. This is consistent with Perruchet and Pacteau's (1990) finding that bigram ratings could predict classification performance with these strings.

Subjects are able to recognize the well-formedness of exemplars or elements of exemplars. The elements can be in isolation, as shown by the high levels of SLD responding for small stem lengths. That is, the overall sense of grammaticality of an exemplar is analyzable by the subject when directly probed; the subject is capable of formulating general rules

[7] The significance of this contrast was not reported by Reber et al. (1980). However, it can be calculated from the data reported in the article to be $p > .10$.

[8] This includes both within-group and between-group covariation.

that capture the perceived grammaticality of an exemplar. While classification knowledge may be seen as implicit in some sense, it is not "*implicit* in our sense that [subjects] are not consciously aware of the aspects of the stimuli which lead them to their decision" (Reber & Allen, 1978, p. 218).

Learning Modes and Knowledge Bases

Reber (1976) argued that subjects could approach the task with either an implicit or explicit learning mode; when subjects were asked to search for rules, their classification performance deteriorated compared with the performance of subjects who were simply asked to memorize the strings. Although this early result was encouraging, there has been a consistent failure to replicate, both in Reber's laboratory (Reber et al., 1980) and in a number of other laboratories (e.g., Dulany et al., 1984; Hayes, 1987; Mathews et al., 1989). Experiment 2 provided data consistent with the latter studies; intentional rather than incidental instructions had no influence on classification performance.

Hayes (1987; and in Broadbent, 1989) seemed to provide a way of reconciling these findings with the existence of two learning modes: The effect of intentional instructions might show itself only under dual-task conditions. However, Experiment 2 demonstrated that the RNG task used by Hayes interfered equally with incidental and intentional subjects. Experiment 2 also indicated that low power probably prevented Hayes from detecting a dual-task effect on incidental subjects. In sum, the evidence from both intentional versus incidental instructions and the use of dual tasks is consistent with subjects approaching artificial grammar learning with a single mode of learning. Future research could usefully investigate the effect of stronger experimental manipulations in inducing different learning modes (Reber, 1989). Future research could also investigate the influence of RNG and intentional versus incidental instructions on tasks that may involve a greater implicit component (e.g., the RT tasks of Cohen et al., 1990).

The data of Experiments 1 and 2 were also consistent with a single knowledge base underlying different measures of artificial grammar learning. The strong relation between classification performance and the SLD test was discussed in the last section. Further, the Spearman's correlation over the nine groups (of Experiments 1 and 2 and the pilot study) between classification performance and PPFR was .75, $p < .05$, and the Spearman's correlation between PPSUM and PPFR was .91, $p < .01$. That is, changes in the group mean of one knowledge measure was mirrored by similar changes in the mean of either of the other measures. Classification performance and PPSUM had very similar group means; PPFR consistently underpredicted classification performance. These results are consistent with a single knowledge base, tapped with equal sensitivity by classification performance and the SLD test, but only inadequately by free report. This conclusion is also supported by results from Mathews et al. (1989). They found that yoked subjects who used the free report instructions from experimental subjects exposed to the grammar always classified fewer correct items than the experimental subjects, consistent with the insensitivity of free report to the knowledge base. They also found that an increase in classification performance by the experimental subjects was matched by an increase by yoked subjects, consistent with a single knowledge base underlying classification performance and free report. Future research needs to investigate tasks that may have a greater implicit component.

Is the Knowledge Implicit?

If subjects learn the artificial grammar task in a single mode and acquire a single knowledge base, is the knowledge implicit? The classification knowledge was not completely elicited by an immediate free-report test. This is intriguing evidence for regarding the knowledge as implicit because there are other cases (e.g., Mathews, Buss, Chinn, & Stanley, 1988; Schwartz, 1966) in which classification knowledge is adequately elicited by free report. However, a single dissociation is only weak evidence for separate processes (see, e.g., Dunn & Kirsner, 1988). One important difference between artificial grammar learning and other typical concept formation tasks (e.g., Schwartz, 1966) is the number of associations that may need to be stored. This factor could explain the failure of free report in these experiments without invoking different implicit and explicit knowledge types.

In part, classification can be regarded as a recognition task. The associations presented in an exemplar can be recognized or not, and then some rule needs to be applied (e.g., a linear combination of the information, such as that used to derive PPSUM) to make an overall classification decision. Recognition is generally more sensitive than free report (see, e.g., Tulving, 1983), and the difference between free report and recognition occurs particularly for large set sizes (Davis, Sutherland, & Judd, 1961). In Experiments 1 and 2, the subject may usefully form up to $(n^2 + n)/2$ associations for an n letter exemplar (each letter with any other in the exemplar). Under these conditions, and especially because some of the associations may be low confidence, it is not surprising that not all of the associations are retrieved in free report in Experiments 1 and 2. However, this issue needs further empirical investigation. Future research needs to establish whether the low levels of free-report performance in artificial grammar learning reflect a problem of retrieving considerable low confidence knowledge in a short period of time or whether they reflect a deeper incompatibility between the mechanisms employed in free report and the type of knowledge stored.

The results of Experiment 2 gave some hints about the type of knowledge acquired in artificial grammar learning. The dual-task effect in the absence of a priority effect on grammar learning in Experiment 2 suggests that there may be some resource that is required for grammar learning and that is applied to RNG in an all-or-none way. One possible resource is the articulatory loop. This may be used in grammar learning to provide an articulatory encoding of the visual stimuli. Baddeley (1986) suggested that the articulatory loop was important in encoding order information in complex stimuli; thus, the articulatory loop may be important for artificial grammar learning in learning order information more complex than admissible bigrams. Consistently, analysis of the SLD responses indicated that dual- rather than single-task

conditions interfered with subjects' knowledge of the positional dependence of bigrams (compare Cohen et al. 1990). The relation of the articulatory loop to the coding of the positional dependence of bigrams in artificial grammar learning requires further investigation.

In summary, Experiments 1 and 2 found little evidence of distinct learning modes or knowledge types in artificial grammar learning. However, there are other aspects of artificial grammar learning that appear interesting; for example, it appears to be learned in a relatively passive way. Thus, asking subjects to search for rules did not enhance learning the rules of the grammar; also, shifting resources away from random number generation, presumably toward artificial grammar learning, under dual-task conditions did not improve grammar learning. Not all concept formation tasks are passive in this way. For example, Mathews et al. (1989) found that learning a biconditional rule (but not a finite-state grammar) was impaired by incidental rather than by intentional learning conditions. Similarly, Abrams and Reber (1988) found that psychiatric patients as compared with normal controls were impaired in learning a biconditional rule but not in learning a finite-state grammar. Further research needs to ascertain in what way the learning process and the resulting knowledge underlying artificial grammar learning can be regarded as implicit.

References

Abrams, M., & Reber, A. S. (1988). Implicit learning: Robustness in the face of psychiatric disorders. *Journal of Psycholinguistic Research, 17*, 425–439.

Allen, R., & Reber, A. S. (1980). Very long-term memory for tacit knowledge. *Cognition, 8*, 175–185.

Attneave, F. (1959). *Applications of information theory to psychology.* New York: Holt, Rinehart & Winston.

Baddeley, A. D. (1966). The capacity for generating information by randomisation. *Quarterly Journal of Experimental Psychology, 18*, 119–129.

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Clarendon Press.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36*, 209–231.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79*, 251–272.

Brewer, W. F. (1974). There is no convincing evidence for operant or classical conditioning in adult humans. In W. B. Weimer & D. S. Palermo (Eds.), *Cognition and the symbolic processes* (pp. 1–42). Hillsdale, NJ: Erlbaum.

Broadbent, D. E. (1989). Lasting representations and temporary processes. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 211–227). Hillsdale, NJ: Erlbaum.

Brody, N. (1989). Unconscious learning of rules: Comment on Reber's analysis of implicit learning. *Journal of Experimental Psychology: General, 118*, 236–238.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language* (Vol. 2, pp. 76–105). New York: Wiley.

Cohen, A., Ivry, R., & Keele, S. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 17–30.

Davis, R., Sutherland, N. S., & Judd, B. R. (1961). Information content in recognition and recall. *Journal of Experimental Psychology, 61*, 422–429.

Druhan, B., & Mathews, R. (1989). THIYOS: A classifier system model of implicit knowledge of artificial grammars. *Proceedings of the eleventh annual conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

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

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1985). On consciousness in syntactic learning and judgment: A reply to Reber, Allen, and Regan. *Journal of Experimental Psychology: General, 114*, 25–32.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processess: The principle of reversed association. *Psychological Review, 95*, 91–101.

Eriksen, C. W. (1962). Figments, fantasies, and follies: A search for the subconscious mind. In C. W. Eriksen (Ed.), *Behaviour and awareness* (pp. 3–26). Durham, NC: Duke University Press.

Hayes, N. A. (1987). *Systems of explicit and implicit learning.* Unpublished doctoral dissertation, University of Oxford, England.

Mathews, R. C. (1990). Abstractness of implicit grammar knowledge: Comments on Perruchet and Pacteau's analysis of synthetic grammar learning. *Journal of Experimental Psychology: General, 119*, 412–416.

Mathews, R. C., Buss, R. R., Chinn, R., & Stanley, W. B. (1988). The role of explicit and implicit learning processes in concept discovery. *Quarterly Journal of Experimental Psychology, 40*, 135–165.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). 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. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology* (pp. 95–100). New York: Free Press of Glencoe.

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. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 88–94.

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

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

Reber, A. S., Allen, R., & Regan, S. (1985). Syntactical learning and judgment: Still unconscious and still abstract: Comment on Dulany, Carlson, and Dewey. *Journal of Experimental Psychology: General, 114*, 17–24.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 492–502.

Reber, A. S., & Lewis, S. (1977). Implicit learning: An analysis of the

form and structure of a body of tacit knowledge. *Cognition, 5*, 333–361.

Schwartz, S. H. (1966). Trial-by-trial analysis of processes in simple and disjunctive concept-attainment tasks. *Journal of Experimental Psychology, 72*, 456–465.

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

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction, and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology, 41*, 553–577.

Truijens, C. L., Trumbo, D. A., & Wagenaar, W. A. (1976). Amphetamine and barbiturate effects on two tasks performed singly and in combination. *Acta Psychologica, 40*, 233–244.

Tulving, E. (1983). *Elements of episodic memory.* Oxford, England: Clarendon Press.

Wagenaar, W. A. (1970). Subjective randomness and the capacity to generate information. *Acta Psychologica, 33*, 233–242.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1047–1060.

Received September 24, 1990
Revision received January 3, 1991
Accepted January 14, 1991 ■