Memory & Cognition
2000, 28 (8), 1321-1332

Learning artificial grammars: No evidence for the acquisition of rules

ANNETTE KINDER and ANJA ASSMANN
Philipps University, Marburg, Germany

Two experiments investigated whether there is evidence for acquisition of rules in implicit artificial grammar learning (AGL). Two different methods were used in meeting this goal, multiple regression analysis and analysis of receiver-operating characteristics (ROCs). By means of multiple regression analysis, several types of knowledge were identified that were used in judgments of grammaticality, for example, about single letters and about larger stimulus fragments. There was no evidence for the contribution of rule knowledge. The ROCs were in accord with a similarity-based account of AGL and thus did not support the notion that rule knowledge is acquired in AGL either. Simulations with a connectionist model corroborated the conclusion that the results were in accord with a similarity-based, associative account.

An important question in implicit learning research has been whether implicit knowledge is stored in terms of the surface features of the stimulus environment or in terms of abstract, rule-like descriptions. A paradigm widely used to investigate this question is *artificial grammar learning* (AGL). In the experiments reported in this article, we applied two different methods to find out whether the knowledge acquired in AGL can be described solely in terms of surface stimulus features or whether an (additional) rule-based process has to be assumed. The first method, multiple regression analysis (e.g., Johnstone & Shanks, 1999), is an excellent tool to investigate the impact of rule adherence as well as several types of surface information on performance in AGL experiments. The second method, the analysis of receiver-operating characteristics (ROCs; e.g., Yonelinas, 1997), was applied to investigate the processes underlying performance in AGL.

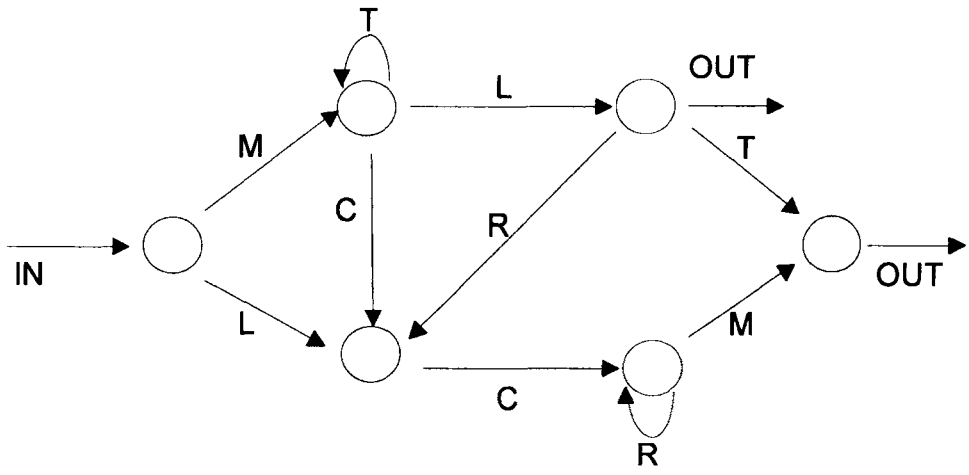
In AGL, strings are presented that were generated according to an artificial grammar (such as the one depicted in Figure 1). During training, a subset of all grammatical strings (i.e., the strings that can be generated by means of a particular grammar) are presented. Usually, participants are told that they are taking part in a simple short-term memory experiment and are instructed to memorize the strings. That way, incidental training conditions are provided. Only after the training stage is over are participants informed about the existence of a set of complex rules constraining letter order. Then they are presented with new strings that are either grammatical or nongrammatical. Nongrammatical strings violate at least one of the rules of the grammar. When participants are asked to categorize these strings, their performance is typically well above chance level.

The authors thank David Shanks and two anonymous reviewers for their helpful comments on a previous version of this article. Stimulus materials, simulation details, and simulation parameters used in this article will be made available on e-mail request. Correspondence should be addressed to A. Kinder, Fachbereich Psychologie, Gutenbergstr. 18, Philipps-Universität, D-35032 Marburg, Germany (e-mail: kinder@mailer.uni-marburg.de).

Models of AGL

Various models have been proposed that make entirely different assumptions about the knowledge acquired in AGL. We will use the terms *rule-based* and *similarity-based* to categorize these models (see Hahn & Chater, 1998). Rule-based models assume that knowledge is stored in collections of rules that are organized in theories. By contrast, similarity-based models assume that past situations or aspects of these situations are stored in memory. In the present article, similarity-based models refer to all AGL models that do assume that participants learn the training strings or the surface features of these strings rather than the rules of the grammar. In turn, we will give a selective overview of rule-based and similarity-based models of AGL and, subsequently, take a look at hybrid models that incorporate both types of processing.

Reber (1967, 1969, 1989) assumes that participants' knowledge in AGL is entirely rule-based. Participants are thought to acquire an abstract representation of the rules of the grammar and to apply this knowledge unconsciously. According to Reber, participants either know the grammatical status of an item or are guessing. Thus, he describes application of knowledge in AGL in terms of an all-or-none process or a threshold model. On the view that participants acquire rule knowledge, this is a plausible assumption: As Hahn and Chater (1998) note, the condition of a rule is either satisfied or not, while intermediate values are not allowed. In AGL, this means that a test string either adheres to the rules somebody has learned or not. Dienes, Kurz, Bernhaupt, and Perner (1997) argue that Reber's threshold model is a high-threshold model (Luce, 1963) because he assumes that participants are al-



Grammatical Strings	Nongrammatical Strings
MTTTLT	TTTLT
MLRCRRM	MLRRCM
LCRRM	LCLM

Figure 1. A typical finite-state grammar. Grammatical strings are generated by following the arrows starting at IN and continuing until an exiting path is taken (OUT). Each time an arrow is chosen, the letter associated to it is appended to the string.

ways correct in assessing grammaticality and make mistakes only due to guessing. In the present article, we refer to Reber's notion when we use the term *rule-based processing*. An alternative conception of rule-based processing will be described in the General Discussion.

According to similarity-based accounts of AGL, participants learn surface features of the training strings or entire strings rather than the rules of the grammar. Vokey and Brooks (1992), for example, proposed an account according to which participants store single training items but also build a pooled representation of multiple items. The probability of a test item to be endorsed depends on how similar it is to these representations. Thus, the model assumes that grammaticality judgments are made by assessing items on a continuous similarity dimension rather than on the basis of an all-or-none process. Unlike Vokey and Brooks's account, Servan-Schreiber and Anderson's (1990) competitive chunking model does not assume that information about entire training strings is preserved. According to this model, participants store knowledge about letter chunks, which can be bigrams, trigrams, or larger string fragments. In the core of Servan-Schreiber and Anderson's model is the concept of "familiarity", which they assume to be a continuous dimension. The familiarity of a test string depends on how many chunks it contains that have been stored during training and on how large these chunks are. The probability of an item to be endorsed is a positive function of its familiarity. Another similarity-based model of AGL is the *simple recurrent network* (SRN) model, which belongs to the class of connectionist models (Cleeremans, Servan-Schreiber, & McClelland, 1989; Kinder, 2000). The knowledge stored in an SRN could be described in a simplified way as knowledge about the absolute and relative frequencies of letters in the set of training strings (e.g., that "X" occurs very often or that "T" often occurs after "XS"). This information is stored in a distributed fashion in the network's connection weights. Whether or not a test string is endorsed depends on how closely it corresponds to the stored information. Thus, like in the similarity-based models described above, strings are assessed on a continuous dimension. Although it is sometimes argued that the information stored in an SRN is in some way abstract (Cleeremans et al., 1989), the SRN is similarity-based in that it involves an associative learning mechanism rather than assuming acquisition of rules.

Our overview has shown that similarity-based accounts of AGL differ in their assumptions about which kind of surface information is acquired in AGL and how it is acquired. However, all of them assume that grammaticality judgments are made by assessing items on a (hypothetical) continuous dimension rather than on the basis of an all-or-none process as in Reber's rule-based account of AGL. Whereas Vokey and Brooks (1992) term this dimension *similarity* and Servan-Schreiber and Anderson (1990) term it *familiarity*, there is no explicit name for it in the SRN model.

Knowlton and Squire (1996) and Meulemans and Van der Linden (1997) argued that neither an entirely rule-based account nor an entirely similarity-based account

is sufficient to explain test performance in AGL experiments. Therefore, they propose hybrid accounts of AGL that combine rule- and similarity-based processing. These accounts assume that participants learn both surface information and the rules of the grammar and apply both types of knowledge while making grammaticality judgments.

Experimental Designs for Investigating Rule-Based Processing in AGL

Two experimental designs were used to find evidence for rule-based processing in AGL. In the first design, rule adherence and surface information are manipulated in such a way that grammatical and nongrammatical stimuli are identical in terms of surface information (Knowlton & Squire, 1996, Experiment 1; Meulemans & Van der Linden, 1997). If an effect of grammaticality still occurs, it is argued that participants must have acquired rule knowledge beyond the surface features of the strings. However, the problem with this design is that it is impossible to balance the items in terms of all surface features that possibly influence grammaticality judgments. In Knowlton and Squire's stimulus materials, for example, grammatical items contained a higher number of entirely novel chunks than did nongrammatical items. Thus, participants simply might have tended to reject items containing novel chunks. Although Meulemans and Van der Linden's materials were controlled in terms of chunk novelty, Johnstone and Shanks (1999) found that grammatical strings contained more chunks in old positions—that is, in positions in which they had already occurred in the training strings. A multiple regression analysis showed that participants mainly used information about novel chunk positions for making judgments.

The problem of rule adherence being confounded with surface information was thought to be circumvented by investigating the transfer of grammar knowledge to a new letter set. In experiments using this design, participants are first trained with strings generated by a particular grammar using one letter set. Then, test strings generated by the same grammar but using a different letter set are presented. Thus, surface information is completely changed, whereas the rules according to which letter strings were generated remain constant. In several experiments using this design, participants' test performance was above chance despite the changed letter set (e.g., Shanks, Johnstone, & Staggs, 1997; Whittlesea & Dorken, 1993). The seemingly straightforward conclusion that can be drawn from this finding is that participants must have learned something about the deep structure of the training strings, independent of their specific physical appearance. However, Redington and Chater (1996) showed that transfer to a new letter set can be explained solely on the assumption that participants have stored fragments of the training strings and find analogies between these stored fragments and the fragments of each test string. To summarize, both designs described in this section are inappropriate to show that rule knowledge is acquired in AGL.

An Alternative Approach

We therefore chose an alternative, two-pronged approach in order to investigate whether test items are judged on the basis of surface information, on the basis of rule knowledge, or both. First, we used the multiple regression method that was first applied to AGL data by Johnstone and Shanks (1999). In this method, the predictor variables are the various features of the training strings that possibly have an impact on grammaticality judgments, such as associative chunk strength (which captures the number of familiar bigrams and trigrams a test string comprises), chunk novelty, and grammaticality (for a detailed description of these measures, see Johnstone & Shanks, 1999, and Kinder & Shanks, in press). The dependent variable codes the participants' responses to each test string. By means of this method, it is possible to assess which sources of information significantly influence grammaticality judgments. The clear advantage of the multiple regression method is that it allows the isolation of the influences of various properties of the test strings without manipulating them in an orthogonal fashion. Most important, it provides information about whether participants have learned something about the rules of the grammar: If all surface features one can think of are included as predictors, and the grammatical status of the items is still a significant predictor, this indicates that participants have acquired knowledge about the grammar that is independent of the surface features of the stimuli.

Like Johnstone and Shanks (1999), we used the individual regression equation method recommended by Lorch and Myers (1990) in order to test the effects of the predictors against the appropriate error term. In this method, a regression analysis is computed on the data of every participant, thus providing a set of regression weights for every data set. Subsequently, the regression weights are averaged across participants. We intended not only to replicate the results of Johnstone and Shanks but also to apply the method to different stimulus materials generated with a different grammar. By extending the set of predictors used by Johnstone and Shanks, we investigated additional sources of information that might underlie grammaticality judgments.

The second method we used was the analysis of ROCs, which was introduced by Yonelinas (1994, 1997) for investigating the processes mediating recognition memory. According to an influential theory, recognition involves two distinct processes: *familiarity* and *recollection* (e.g., Jacoby, Toth, & Yonelinas, 1993). Familiarity is thought to be a continuous variable, and test performance based on it should be in accord with signal detection theory (e.g., Yonelinas, 1994). By contrast, recollection is assumed to be a discrete retrieval process that can return an exact match to the test stimulus presented (e.g., Yonelinas, 1997). Mathematically, this process can be described in terms of a threshold model: The threshold is exceeded if an exact match in the memory store is found. Obviously, this can happen only with old items, which makes the process a high-threshold process (Luce, 1963). These two

processes have the same general characteristics as the two processes that possibly are involved in AGL: As noted above, rule-based models assume that assessment of grammaticality is a high-threshold process, whereas similarity-based models assume that items are assessed on a continuous dimension. Thus, if the analysis of ROCs is useful for investigating the processes involved in recognition memory, it should also be useful for studying the processes mediating grammaticality judgments.

Generation and Analysis of Receiver Operating Characteristics

How are ROCs generated from data obtained in recognition experiments? In these experiments, a list of study words is presented first. Subsequently, a test list is given comprising old items, which were in the study list, and new items, which were not. A recognition ROC is the function that relates the proportion of hits (i.e., old items that are called "old") to the proportion of false alarms (i.e., new items that are called "old") at different criteria. Typically, this function is obtained by asking the participants to give confidence judgments on a scale ranging from *sure the item is old* to *sure the item is new*. The points on the ROC are plotted separately for each participant as follows: The first point includes only the proportions of old and new items remembered most confidently. The second one additionally includes the next most confident responses, and so on. The last point includes all responses, except the "sure new" ones. Thus, if responses are made on a 6-point confidence scale, a 5-point ROC will result. The same procedure can be used to generate ROCs in AGL experiments if participants are asked to give grammaticality judgments on a confidence scale rather than giving binary responses.

Some of the ROCs that theoretically can be obtained by means of this method are shown in the right-hand panels of Figure 2. If the points of these ROCs are plotted in *z*-space, the curves in the left-hand panels of Figure 2 do emerge. What does the shape of the ROC tell us about the cognitive processes that are involved in responding? If a ROC like the one in Figure 2a is obtained in a recognition experiment, this is interpreted as evidence that old/new judgments are solely based on familiarity. In general, this kind of ROC can emerge only if responses are made by assessing items on a continuous dimension. Therefore, it is in support of similarity-based models of AGL, rather than rule-based models, if we observe this kind of ROC in AGL experiments.

By contrast, ROCs like the ones in Figures 2b and 2c emerge if a high-threshold process underlies responding. There are two possible thresholds: one for detecting targets and another for detecting nontargets. In a high-threshold model with a threshold for detecting targets, the percentage of hits is given by the equation $P(\text{hit}) = \pi + (1 - \pi)\, P(\text{false alarm})$, where π is the probability that the threshold is exceeded by target stimuli. Thus, the ROC relating hits and false alarms follows a straight line and has a slope less than 1 (Figure 2b, left-hand panel). Consequently, the *z*-ROC will be curvilinear and will also have a slope less than 1 (Figure 2b, right-hand panel). Whereas a "yes" response is given if a threshold for detecting targets is exceeded, a "no" response is given if a threshold for detecting nontargets is exceeded. In the latter high-threshold model, *hit* (items correctly categorized as targets) can be exchanged for *correct rejection* (items correctly categorized as nontargets), and *false alarm* (items incorrectly categorized as targets) can be exchanged for *omission* (items incorrectly categorized as nontargets). Thus, the following linear equation results: $P(\text{correct rejection}) = \pi + (1 - \pi)\, P(\text{omission})$, were π is the probability that the threshold is exceeded on the presentation of a nontarget. This function relating correct rejections and omissions can be easily transformed into a ROC relating hits and false alarms because the probability of a hit is 1 minus the probability of an omission and the probability of a false alarm is 1 minus the probability of a correct rejection. The ROC in a high-threshold model with a threshold for detecting nontargets also follows a straight line, but both the ROC and the *z*-ROC have a slope larger than 1 (see Figure 2c). In the present article, we have to consider both types of thresholds because it is unclear whether rule-based accounts of AGL assume that there is a threshold for detecting grammaticality (i.e., for detecting targets) or for detecting nongrammaticality (i.e., for detecting nontargets). However, in both cases, ROCs different from the one in Figure 2a should emerge. If rule-based and similarity-based processing were combined in AGL, as hybrid models assume, *z*-ROCs would be less curvilinear and have a slope less different from 1 than the ones shown in Figures 2b and 2c but would still be different from the *z*-ROC in Figure 2a.

In the experiments reported in this paper, we obtained ROCs for every participant by means of the confidence method. We then estimated the slope and the intercept separately for each ROC and subsequently averaged these values across participants (Yonelinas, 1994). By means of this procedure, we tried to find out whether grammaticality judgments are made by assessing test items on a continuous dimension, as similarity-based models assume, or whether there is evidence for an (additional) all-or-none-process, as rule-based models assume.

EXPERIMENT 1

In Experiment 1, we replicated Knowlton and Squire's (1996) Experiment 1. In this experiment, Knowlton and Squire varied chunk strength and grammaticality in an orthogonal fashion. Thus, they sought to investigate the influences of chunk strength and grammaticality independently of each other. Our training procedure and our stimulus materials were identical to those of Knowlton and Squire, and we changed only the testing procedure: Whereas Knowlton and Squire asked their participants to give a binary judgment about the grammatical status of each test item, we asked our participants to give a judgment on a 6-point confidence scale. The scale ranged from 1, (the item is) *surely correct*, to 6, (the item is) *surely incorrect*.

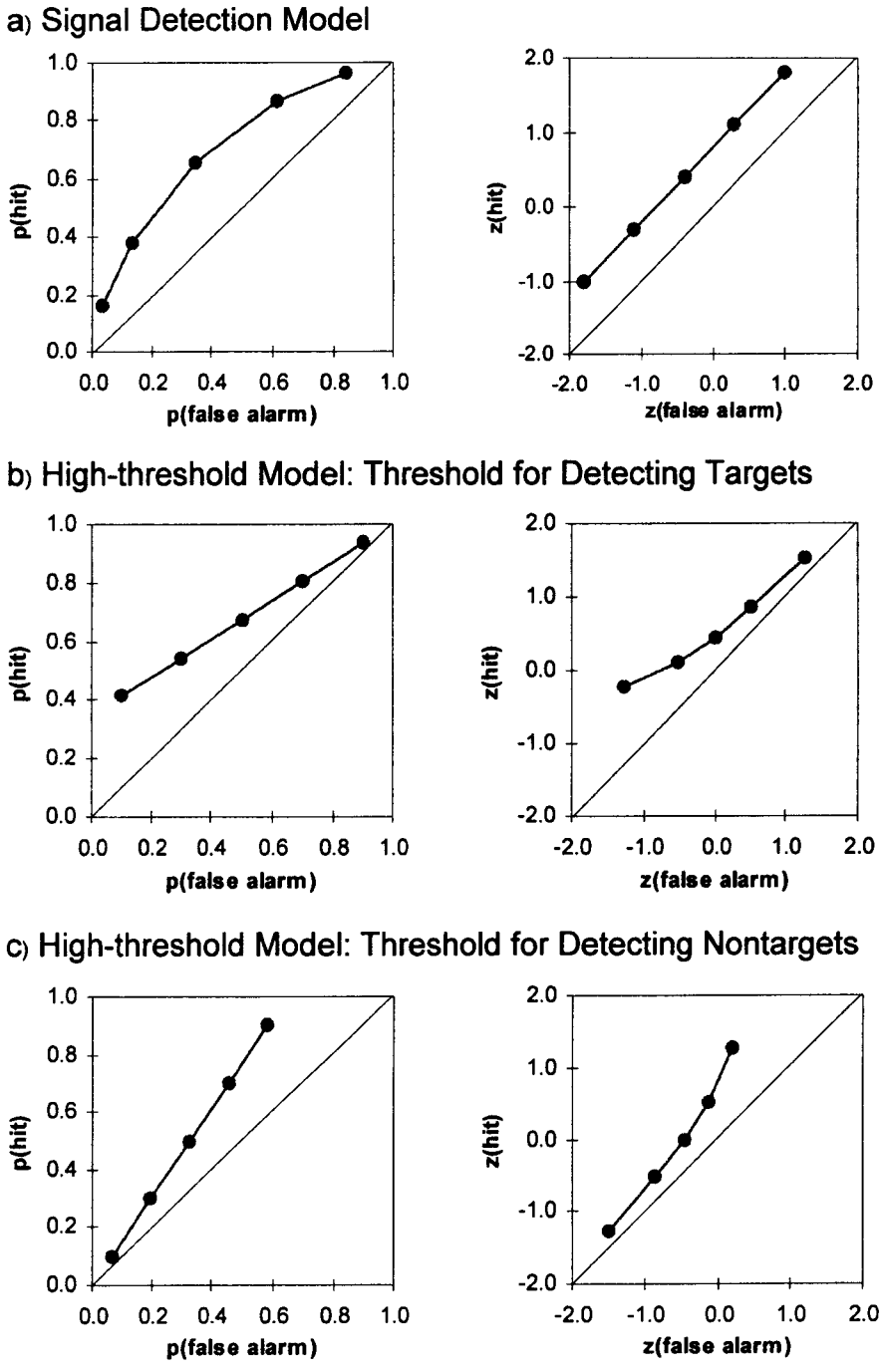


Figure 2. The ROCs predicted by signal detection theory and two types of high-threshold models. The left-hand panels show the original ROCs, and the right-hand panels show the ROCs plotted in *z*-space.

Method

Participants. The participants were 20 students from Philipps University, Marburg. They were from 19 to 27 years old (M = 22.35 years).

Stimuli. We used the same grammar as used by Knowlton and Squire (1996, Experiment 1). The 23 training stimuli and the 32 test stimuli were generated by means of this grammar and were identical to the ones used by Knowlton and Squire except for a single letter that was exchanged. We replaced the letter *T* with the letter *F* in all stimuli, because the original stimuli comprised some well-known (German) abbreviations. There were four types of test stimuli, which differed with respect to grammaticality and chunk strength: These were 8 grammatical strings with high chunk strength, 8 grammatical strings with low chunk strength, 8 non-

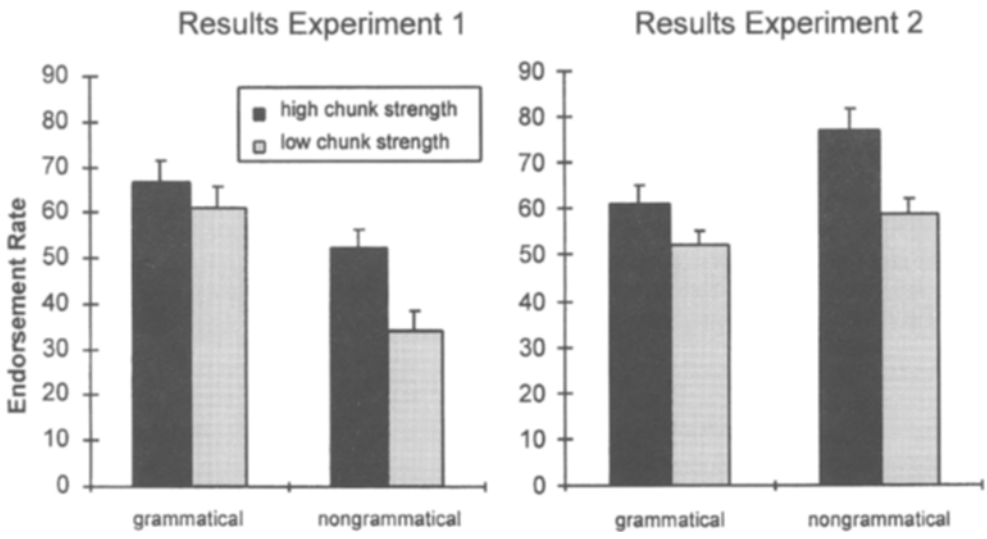


Figure 3. Mean endorsement rates in Experiments 1 and 2. The error bars indicate the standard errors.

grammatical strings with high chunk strength, and 8 nongrammatical strings with low chunk strength (see Knowlton & Squire, 1996, for the stimuli and their mean chunk strength values).

Procedure. *Training stage*: The participants were told that they were taking part in a short-term memory experiment. The 23 training stimuli were presented one at a time on the computer screen for 4 sec each. Four seconds after a training string had disappeared, the participant was told to type the string on the keyboard. If the participant did not reproduce the letter string correctly, it was shown again until it was reproduced without any mistakes. The training strings were presented in random order. The set of 23 training strings was presented twice, for a total of 46 items.

Test stage: After the training stage was over, the participants were informed for the first time that the letter strings they had just seen had been formed according to a complex set of rules constraining letter order. They were instructed to classify the test strings, which were presented one at a time on a computer screen, according to whether or not they followed these rules. Like Knowlton and Squire (1996), we encouraged the participants to "rely on their feeling" while making their judgment. Grammaticality judgments were given on a 6-point rating scale ranging from 1, (the item is) *surely correct*, to 6, (the item is) *surely incorrect*. This scale was presented on every test trial beneath the test string. The participants were asked to type the appropriate number on the keyboard.

Results

The level of significance was set to .05 in the regression analysis and the analysis of variance (ANOVA). To compare our results with the results of Knowlton and Squire (1996, Experiment 1), we counted responses from 1 to 3 as "grammatical" judgments and responses from 4 to 6 as "nongrammatical" judgments. Figure 3 (left-hand panel) shows the mean endorsement rates for grammatical and nongrammatical strings with high and low chunk strength in Experiment 1. As can be seen in this figure, both grammaticality and chunk strength influenced the endorsement rates. Like in Knowlton and Squire's experiment, the chunk strength effect was larger with nongrammatical items than with grammatical items. A 2 × 2 factorial ANOVA was computed with grammaticality and chunk strength as within-subjects variables. Like in Knowlton and Squire's experiment, there was a significant main effect of both grammaticality [$F(1,19) = 20.1$, $MS_e = 0.042$, $p < .001$], and chunk strength [$F(1,19) = 13.2$, $MS_e = 0.021$, $p < .002$]. Unlike Knowlton and Squire, we found no significant grammaticality × chunk strength interaction [$F(1,19) = 2.4$, $MS_e = 0.033$, $p > .14$].

In order to find out which kind of information the participants used to give grammaticality judgments, we performed a multiple regression analysis according to the method suggested by Lorch and Myers (1990; see Johnstone & Shanks, 1999, for details on this method). Nine predictor variables were defined, which already had been used by Johnstone and Shanks (1999) or by Kinder and Shanks (in press): (1) grammaticality (1, *grammatical*; 0, *nongrammatical*), (2) a variable indicating whether or not the first letter of the string had appeared in that position in the training strings, which was called "familiarity of the starting letter" (1, *familiar starting letter*; 0, *novel starting letter*), (3) anchor chunk strength, (4) global chunk strength, (5) length, (6) novel chunk positions, (7) chunk novelty, (8) a variable indicating whether the pattern of repetitions was a familiar one, and (9) specific item similarity. Global chunk strength, anchor chunk strength, chunk novelty, and novel chunk positions were computed as described by Johnstone and Shanks. The familiarity of the starting letter was coded because, with only two different initial letters (X and V), a new letter occurring at the first position of a test string would surely be highly salient. Specific item similarity of a string was defined as the number of positions in which it diverged from its most similar training string, which could be of identical or of different length. The test string VXJ, for example, was assigned a specific item similarity value of 1, because there was a training string VXJJ. The pat-

Table 1
Regression Weights Averaged Across Participants Found in the Initial Multiple Regression Analyses of Experiments 1 and 2

	Experiment 1				Experiment 2			
	β				β			
Predictor	*M*	*SE*	*t*	*p*	*M*	*SE*	*t*	*p*
Grammaticality	−.119	.056	−2.12	.047	.221	.048	4.59	.000
Familiarity of the starting letter	−.170	.038	−4.46	.000				
Anchor chunk strength	.084	.071	1.18	.252	−.055	.064	−0.85	.404
Global chunk strength	−.148	.059	−2.50	.022	−.098	.055	−1.77	.092
Length	.092	.082	1.13	.273	−.089	.066	−1.36	.191
Novel chunk positions	.007	.058	0.12	.903	.230	.049	4.73	.000
Novelty	.039	.069	0.57	.575				
Pattern of repetitions	−.119	.045	−2.64	.016				
Similarity	−.019	.050	−0.39	.702	−.015	.038	−0.38	.706

tern of repetitions was assessed in the following way: For each letter in a string, beginning from the second position, we determined whether or not it was identical to its predecessor. This was coded as either C for change or R for repetition. The pattern of repetitions of the string XXVJJJ, for example, was RCCRR. For each test string, we then assessed whether or not there was a training string with an identical pattern of repetitions (1, *familiar pattern of repetition*; 0, *new pattern of repetitions*).

A regression analysis was computed separately for each participant, and the regression coefficients (standardized beta weights) were then averaged across participants. For each predictor variable, a one-sample *t* test was computed to assess whether or not it differed reliably from zero. Table 1 shows the mean beta weights (with standard errors), the *t* values, and the probability values for each predictor. On average, the predictors accounted for 39.84% of the variance. A positive beta weight means that an item having a high value on the predictor variable has a high probability of being rejected; a negative beta weight means that it has a high probability of being accepted (because the scale ranged from 1, *surely correct*, to 6, *surely incorrect*). We found four significant predictors: grammaticality, familiarity of the starting letter, global chunk strength, and pattern of repetitions. Thus, items being grammatical, starting with a familiar letter, having high global chunk strength, or having a familiar pattern of repetitions were endorsed with a higher probability than were items being nongrammatical, starting with a novel letter, having low global chunk strength, or having an unfamiliar pattern of repetitions.

ROCs were computed as follows: For each participant, the cumulated relative frequencies for each response category were calculated separately for grammatical and nongrammatical strings beginning with Category 1 (*surely correct*). Subsequently, these cumulative frequencies were *z*-transformed. The mean *z*-ROC obtained with this method is shown in Figure 4 (left-hand panel). In order to find out whether *z*-ROCs had a slope different from 1 or whether they had a significant quadratic component indicating concavity, we computed a regression analysis including a quadratic component for every participant. We

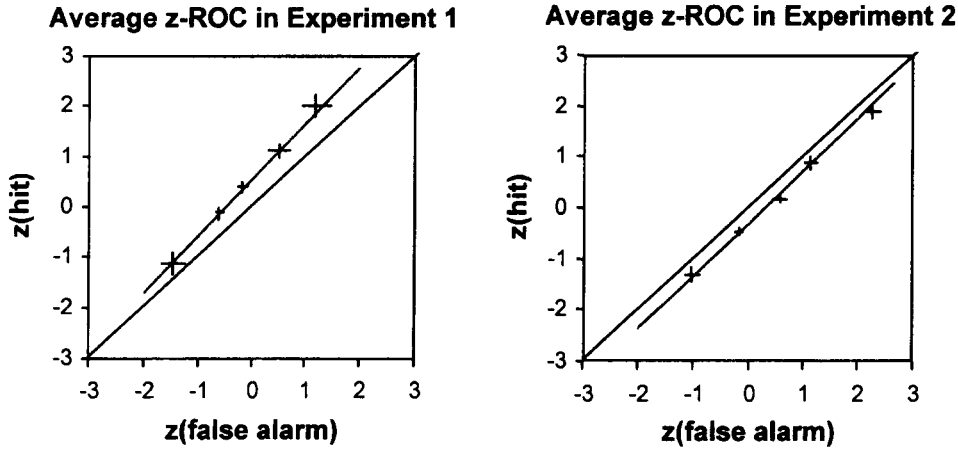


Figure 4. Mean *z*-ROCs obtained in Experiments 1 and 2. The horizontal error bars indicate the standard errors of the cumulative proportions of false alarms. The vertical error bars indicate the standard errors of the cumulative proportions of hits.

then computed one-sample t tests with the coefficients obtained in these analyses in order to find out whether possible deviations from the linear z-ROC having a slope of 1 were statistically significant. The level of significance was set to .10 in these t tests in order to enhance power, because we expected the slope not to differ from 1 and the quadratic component not to differ from 0. We used the regression equation $z(\text{hit}) = a + b * z(\text{false alarm}) + c * z(\text{false alarm})^2$. The average goodness of fit using this equation was $r^2 = .96$. The mean intercept (a) was .46, the mean slope (b) was 1.11, and the mean quadratic constant (c) was .02. The t tests showed that the slope did not differ significantly from 1 [$t(19) = 0.72, p > .47$] and that the quadratic constant did not differ significantly from 0 [$t(19) = 0.13, p > .89$].

Discussion

In Experiment 1, we basically replicated the results obtained by Knowlton and Squire (1996): Like these authors, we found significant effects of both grammaticality and chunk strength. Although the grammaticality × chunk strength interaction failed to reach statistical significance in our experiment, the overall pattern (a larger effect of chunk strength with nongrammatical items than with grammatical items) was very similar to the one reported by Knowlton and Squire. Thus, the fact that our participants had to give confidence judgments instead of yes/no responses seemed to have no remarkable impact on the results.

The results of the regression analysis showed that three different surface features of the test stimuli affected grammaticality judgments. As in the analysis of the mean endorsement rates, chunk strength significantly influenced the participants' responses. Like in Kinder and Shanks's (in press) experiment, the familiarity of the starting letter and the repetition pattern also turned out to be important. However, not only surface features of the strings but also the test items' grammatical status influenced the participants' judgments significantly. One possible interpretation of this result is that the participants indeed learned something about the rules of the grammar and applied that knowledge at test. Another possible interpretation is that the grammaticality effect occurred because of some other variable confounded with the grammatical status of the stimuli. It is in favor of the latter explanation that the mean amount of variance accounted for by the various predictor variables was rather low (39.84%). Thus, it is likely that there were other, still undiscovered factors influencing grammaticality judgments. Furthermore, the analysis of the z-ROCs did not support the notion that rule knowledge is applied in judgments of grammaticality. The slope of the mean 2-ROC did not differ significantly from 1, and the quadratic component did not differ significantly from 0. Thus, the analysis of the z-ROCs provided results in accord with an entirely similarity-based account of AGL.

EXPERIMENT 2

One reason for the absence of clear evidence for rule-based processing in Experiment 1 could be that the experimental conditions did not support acquisition of rules. According to Meulemans and Van der Linden (1997), participants need a high number of different training stimuli in order to become sensitive to the deep structure of the strings. Thus, in Experiment 1, the number of training trials might have been too low. In Experiment 2, we therefore decided to replicate Meulemans and Van der Linden's Experiment 2B, in which 125 different training items were presented.

Method

Participants. Twenty students from the Philipps University, Marburg, participated in the experiment. They were from 18 to 29 years old ($M = 22.2$ years).

Stimuli. We used the same grammar as used by Meulemans and Van der Linden (1997, Experiment 2B). The 125 training stimuli and the 32 test stimuli were generated by means of this grammar and were identical to the ones used by Meulemans and Van der Linden except for two letters: We replaced *T* with *F* and replaced *R* with *J* in all strings, because the original stimuli comprised some well-known (German) abbreviations. The test stimuli differed with respect to grammaticality and chunk strength: There were 8 grammatical strings with high chunk strength, 8 grammatical strings with low chunk strength, 8 nongrammatical strings with high chunk strength, and 8 nongrammatical strings with low chunk strength.

Procedure. Training and test stages were very similar to those in Experiment 1. The 125 training stimuli were presented one at a time for 4 sec on a computer screen. After a pause of 2 sec, the participants were asked to type the string on the keyboard. If the participant did not reproduce the letter string correctly, it was shown again until it was reproduced without any mistakes. The training strings were presented in the same random order to half of the participants and in inverse order to the other participants. Every string was presented twice. In all other respects the testing procedure was identical to that of Experiment 1.

Results

The level of significance was set to .05 in the regression analysis and the ANOVA. To compare our results with the results of Meulemans and Van der Linden (1997, Experiment 2B), we counted responses from 1 to 3 as "grammatical" judgments and responses from 4 to 6 as "nongrammatical" judgments. Figure 3 (right-hand panel) shows the mean endorsement rates for grammatical and nongrammatical strings with high and low chunk strength in Experiment 2. Our results differed considerably from the results of Meulemans and Van der Linden: We found a reverse effect of grammaticality (i.e., grammatical strings were endorsed less often than were nongrammatical strings). Whereas Meulemans and Van der Linden found no effect of chunk strength, we found that items with high chunk strength were endorsed more frequently than were items with low chunk strength. A 2 × 2 factorial ANOVA was computed with grammaticality and chunk strength as within-subjects variables. There was a significant main ef-

Table 2
Regression Weights (Averaged Across Participants) in Experiments 1 and 2 Including Single Letters as Predictors

	Experiment 1				Experiment 2			
	β				β			
Predictor	*M*	*SE*	*t*	*p*	*M*	*SE*	*t*	*p*
Grammaticality	−.102	.055	−1.86	.079	.112	.051	2.19	.042
Familiarity of the starting letter	−.117	.044	−2.66	.016				
Anchor chunk strength	.074	.079	0.93	.364	−.076	.066	−1.14	.267
Global chunk strength	−.211	.119	−1.77	.093	.020	.068	0.30	.767
Length	−.005	.091	−0.05	.961	−.096	.066	−1.49	.154
Novel chunk positions	−.079	.075	−1.06	.305	.175	.057	3.06	.006
Novelty	.092	.082	1.13	.274				
Pattern of repetitions	−.080	.044	−1.81	.084				
Similarity	−.019	.053	−0.35	.730	−.040	.042	−0.95	.352
F	−.032	.081	−0.39	.700	.275	.047	5.91	.000
J	.145	.039	3.69	.002				
M					−.074	.036	−2.04	.056
V	.023	.064	0.36	.727	.070	.042	1.67	.112
X	.090	.078	1.15	.263	.123	.044	2.77	.012

fect of grammaticality [$F(1,19) = 12.9$, $MS_e = 0.020$, $p < .002$], indicating that grammatical strings were endorsed less often than were nongrammatical strings; there was also a significant main effect of chunk strength [$F(1,19) = 11.0$, $MS_e = 0.034$, $p < .004$], indicating that strings with high chunk strength were endorsed more often than were strings with low chunk strength. There was no significant grammaticality × chunk strength interaction [$F(1,19) = 2.1$, $MS_e = 0.021$, $p > .16$].

As in Experiment 1, regression analyses were computed separately for each participant. Because every test item was presented twice in Experiment 2, the responses given on the two presentations of each test string were averaged. Because all strings began with familiar letters, we did not use the predictor familiarity of the starting letter. Furthermore, we did not include the predictor chunk novelty, because there were no novel chunks in the test strings. Also, we did not use the predictor pattern of repetitions because there were no immediate repetitions of letters in the test strings. The remaining six predictor variables were (1) grammaticality, (2) anchor chunk strength, (3) global chunk strength, (4) length, (5) novel chunk positions, and (6) specific item similarity. Table 1 shows the mean beta weights (with standard errors), the *t* values, and the probability values for each predictor. A positive beta weight means that an item having a high value on the predictor variable has a high probability of being rejected; a negative beta weight means that it has a high probability of being accepted (because the scale ranged from 1, *surely correct*, to 6, *surely incorrect*). Grammaticality was a significant positive predictor, indicating that grammatical stimuli were endorsed less frequently than were nongrammatical stimuli. The other significant predictor was novel chunk positions, indicating that the endorsement probability increased as the number of chunks at novel positions decreased. On average, 36.81% of the variance was accounted for.

The result that the predictor grammaticality was significant but in the reverse direction was rather puzzling and could not be explained on the basis of the analyses presented so far. However, we hypothesized that grammaticality might be confounded with some other variable that was responsible for this effect. We therefore looked for further predictors and noticed that there was a striking difference in the frequencies of single letters appearing in the training strings: The most frequent letter was V, which appeared 210 times. J appeared 179 times, M appeared 162 times, X appeared 122 times, and F appeared only 73 times. Interestingly, grammatical test strings contained a higher number of the low-frequency letters F and X than did nongrammatical strings (31 vs. 24). We therefore performed a second regression analysis with four additional predictors: For each of the letters F, M, V, and X, we computed a predictor that was assigned the value 1 if the test string contained the letter and the value 0 if the test string did not contain the letter. The letter J appeared in every test string, and so we did not use this letter as predictor. Table 2 (right-hand portion) shows the results of the second regression analysis. On average, 56.05% of the variance was explained. Thus, the amount of explained variance was considerably higher than in the first regression analysis. In this analysis, the predictor novel chunk positions was still significant. Additionally, F and X reached statistical significance, indicating that test strings comprising these letters were endorsed less often than were other test strings. F and X were the letters occurring least frequently in the training strings. The regression weight of the predictor grammaticality was much lower than in the previous analysis but was still statistically significant. As in the previous analysis, it was positive, indicating that grammatical items were rejected more often than were nongrammatical items. Of course, it does not make much sense to assume that the participants endorsed items when they detected some violation of grammaticality or rejected items when they detected no such violation. Thus, there still must be a confounding factor that is the true reason for this effect. Unless the effect occurred by chance, the reverse grammaticality effect in-

dicates that we did not discover all factors that influence grammaticality judgments.

The *z*-ROCs were computed separately for every participant as described in Experiment 1. The mean *z*-ROC is shown in Figure 4 (right-hand panel). The intercept is negative because nongrammatical strings were endorsed more often than were grammatical strings. In the ROC analyses, the level of significance again was set to .10 in order to enhance power, because we expected the slope not to differ from 1 and the quadratic component not to differ from 0. The average goodness of fit in the quadratic regression was $r^2 = .97$. The mean intercept (a) was $-.31$, the mean slope (b) was 1.06, and the mean quadratic constant (c) was .06. One-sample t tests showed that the slope did not differ significantly from 1 [$t(19) = 0.71, p > .48$], and that the quadratic constant did not differ significantly from 0 [$t(19) = 0.66, p > .51$].

Discussion

In Experiment 2, we failed to replicate the results of Meulemans and Van der Linden (1997). Whereas these authors found no effect of chunk strength, we obtained a reliable effect of chunk strength. We also found a grammaticality effect, but it was a negative one: Nongrammatical strings were endorsed with a higher probability than were grammatical strings. Of all our results, this is perhaps the most compelling evidence against rule-based processing. If the participants had learned something about the rules of the grammar, a reverse grammaticality effect would have been expected least of all.

The second regression analysis, including the absence or presence of all letters as separate predictors, provided at least a partial explanation of the reverse grammaticality effect. According to the regression analyses, the participants relied considerably on information about single letters. Grammatical items comprised a higher number of atypical letters than did nongrammatical items (i.e., letters that had appeared in the training strings not very frequently). Thus, using information about single letters led to higher endorsement rates for nongrammatical strings than for grammatical strings.

The regression analysis revealed that the participants were far from applying rule knowledge in judgments of grammaticality. Instead, they apparently relied on simple information about fragments and about single letters. Further evidence against the application of rule knowledge came from the ROC analysis. The *z*-ROCs were in accord with the assumption that the grammaticality judgments were made by assessing items on a continuous dimension: The slope of the *z*-ROC did not differ reliably from 1, and there was not a significant quadratic constant indicating concavity.

Reanalysis of Experiment 1: Regression Analysis Including Single Letters as Predictors

The results of the second regression analysis in Experiment 2 made us wonder whether or not the presence of single letters could also have been important in Experiment 1. Therefore, we computed another regression analysis on the data of Experiment 1 (with the same methodology as before) including the presence of all letters used in the grammar as predictors. The results, which can be seen in Table 2, indicate that, in Experiment 1, information about single letters was also used in judgments of grammaticality. Strings containing the letter J had a higher probability of being rejected than did strings not comprising this letter. This is a replication of a result reported by Kinder and Shanks (in press), who also found that strings including J were rejected more often. By including single letters as predictors, the amount of explained variance rose considerably from 39.84% to 50.69%. Most important, in this analysis grammaticality failed to reach statistical significance. This indicates that the effect in the first regression analysis was at least partially due to a confounding of grammaticality with single-letter information.

GENERAL DISCUSSION

In the two experiments reported in this paper, we found no evidence for rule-based processing in AGL. The only result that, at first sight, seemed to indicate that the participants were applying rule knowledge was that the predictor grammaticality reached statistical significance in the multiple regression analysis of Experiment 1. This effect, however, can be explained by grammaticality being confounded with single-letter information, as our second regression analysis of these data revealed. The conclusion that no rule knowledge was acquired was corroborated by the *z*-ROCs observed in the two experiments, which showed neither a slope significantly different from 1 nor a significant quadratic constant that would have indicated concavity. Thus, the ROCs were in favor of a similarity-based account of AGL rather than a rule-based account.

However, it could be argued that our ROCs do not contradict all kinds of rule-based models of AGL, but only Reber's (1967, 1969, 1989) account. Reber assumes that participants build a representation of the rules of the grammar, which is in some way similar to the grammar itself. Furthermore, he assumes that participants either know or do not know the grammatical status of a test item. Therefore, the application of rule knowledge is an all-or-none process and should lead to ROCs rather different from the ones observed in our experiments. An alternative characterization of the rules underlying grammaticality judgments is proposed by Mathews and Roussel (1997). These authors describe a classifier model called *THIYOS*, which assumes that participants acquire numerous fragmentary rules such as "endorse strings that begin with an X" or "endorse strings that end with several Vs," which differ in strength. In this kind of rule-based model, grammaticality judgments could be assumed to be made by assessing items on a continuous rule-adherence dimension: The (hypothetical) value of an item on this dimension could be assumed to be a function of both the number and the strength of rules applying

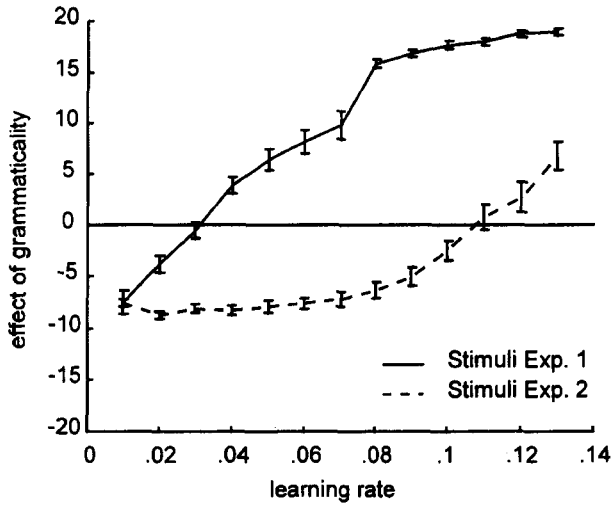


Figure 5. Grammaticality effects predicted by the SRN with the stimuli in Experiments 1 and 2 as a function of the learning rate. Each data point represents an average across 100 simulation runs. The error bars indicate the confidence intervals.

to this item. Thus, a variant of THIYOS might be capable of producing ROCs like the ones found in our experiments. However, even if such a model could account for the ROCs, it could hardly explain the reverse effect of grammaticality in Experiment 2. Actually, this effect is the strongest evidence against rule-based processing in our experiments. Any type of rule-based account would predict that a positive effect of grammaticality should occur particularly in this experiment: The participants had many opportunities to extract the rules of the grammar, because they were presented with so many different grammatical strings.

What did the participants learn if they did not learn the rules of the grammar? We could answer this question by simply describing the results of the regression analyses: The participants must have learned information about single letters, both unrelated and (as far as the first position is concerned) related to position. Furthermore, they must have stored chunks also both unrelated and related to positional information (global chunk strength in Experiment 1, and novel chunk positions in Experiment 2). However, by describing the types of information learned in the two experiments, we cannot explain why there was a positive effect of grammaticality in Experiment 1 but a negative effect of grammaticality in Experiment 2. Although it is rather clear that a rule-based model cannot explain the reversal of the grammaticality effect, it is not at all clear that a similarity-based model of AGL can. Apparently, neither the model by Vokey and Brooks (1992) nor the competitive chunking model by Servan-Schreiber and Anderson (1990) would have predicted this result. This is because these models are not sensitive for single-letter information. However, acquisition of single-letter information is the only explanation we found for the reverse grammaticality effect in Experiment 1.

A model that is sensitive to single-letter information is the SRN model of AGL (Cleeremans et al., 1989; Kinder, 2000). At present, the SRN model appears to be the most successful model of AGL. It has been shown to reproduce several effects found in AGL, such as the effects of similarity to training strings and grammaticality reported by Vokey and Brooks (1992; see Dienes, Altmann, & Gao, 1999), the effects of grammaticality and chunk strength found by Knowlton and Squire (1996; see Redington, 1998), and the transfer of grammar knowledge to a new letter set (e.g., Shanks, Johnstone, & Staggs, 1997; see Dienes et al., 1999). We ran simulations with the stimulus materials of both experiments in order to find out whether or not this model could produce a positive effect of grammaticality in Experiment 1 and a negative effect of grammaticality in Experiment 2 with an identical set of parameters (parameters and simulation details are given on e-mail request). It is important that the parameters are identical in both simulations, because they reflect psychological variables such as learning efficiency and attention, and there is no reason to assume differences between the two samples of participants in these variables. In our simulations, we systematically varied the learning rate parameter. Figure 5 shows the effect of grammaticality (P[yes|grammatical] − P[yes|nongrammatical]) observed in these simulations. With the stimuli of Experiment 1, there is a positive effect of grammaticality already with a learning rate of .04. With the stimuli of Experiment 2, the learning rate parameter has to be three times as large for a positive effect of grammaticality to occur. As a result, there is a range of learning rate parameters where we observe a positive effect of grammaticality with the stimuli of Experiment 1 while observing a negative effect of grammaticality with the stimuli of Experiment 2. This reversal occurs despite the fact that the parameters in the simulations of Experiments 1 and 2 are the same.

What is the reason for the reverse grammaticality effect in our simulations of Experiment 2 with learning rates smaller than .12? It is a result not of the particular grammar used in this experiment but of the stimuli selected for training and for test. Grammatical test stimuli comprise a larger number of letters that occurred in the set of training strings with relatively low frequency than nongrammatical stimuli. The model is particularly sensitive for frequencies of single letters if the learning rate is low: The higher the learning rate, the larger is the model's sensitivity to larger fragments, while its sensitivity to single letters decreases.

The simulation results indicate that the SRN model might be an adequate model to explain our data. Also, it has been shown to account for a variety of other effects observed in AGL experiments (see above). We therefore would like to describe in more detail how knowledge is stored according to this model: The same kind of mechanism might be responsible for acquisition of grammar knowledge in human participants. During training, strings are stored in the network by the gradual adjustment of its connection weights on every trial. If only a single string

was presented to the SRN several times, the network would learn this string perfectly. However, in an AGL experiment, many strings are presented, all of which are stored within the same associative structure. Each string can have individual features that do not occur in any other string. Also, each string has common features that occur in a few or many other strings. Connection weights that represent individual features of a string will be weakened if other strings are presented thereafter. In contrast, connection weights representing common features will be strengthened on the next trials. Thus, the network will store the common features of the strings rather than the individual ones. Typically, grammatical test strings are more similar to the training strings than are nongrammatical test strings in that they incorporate a higher number of common features. This is why the SRN can discriminate between grammatical and nongrammatical strings. It could be argued that, although the model does not assume acquisition of rule knowledge, it nevertheless assumes some kind of abstraction process. However, this abstraction is not the kind of abstraction that is involved in rule learning. It is a result of interference due to new information that is stored in the same memory structure and thus is entirely associative in nature.

In conclusion, we have obtained both empirical and computational evidence for the hypothesis that a similarity-based associative process can explain performance in AGL. Extending other approaches (e.g., Servan-Schreiber & Anderson, 1990), we found that not only fragment information is important in AGL but that information about single letters also influences grammaticality judgments. In our data, there was no evidence for acquisition of rule knowledge in AGL.

REFERENCES

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, **1**, 372-381.

Dienes, Z., Altmann, G. T. M., & Gao, S. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, **23**, 53-82.

Dienes, Z., Kurz, A., Bernhaupt, R., & Perner, J. (1997). Application of implicit knowledge: Deterministic or probabilistic? *Psychologica Belgica*, **37**, 89-113.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, **65**, 197-230.

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General*, **122**, 139-154.

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, & Cognition*, **25**, 524-531.

Kinder, A. (2000). The knowledge acquired during artificial grammar learning: Testing the predictions of two connectionist models. *Psychological Research*, **63**, 95-105.

Kinder, A., & Shanks, D. R. (in press). Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience*.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 169-181.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 149-157.

Luce, R. D. (1963). A threshold model for simple detection experiments. *Psychological Review*, **70**, 61-69.

Mathews, R. C., & Roussel, L. G. (1997). Abstractness of implicit knowledge: A cognitive evolutionary perspective. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 162-194). Oxford: Oxford University Press.

Meulemans, T., & Van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1007-1028.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, **6**, 855-863.

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology*, **81**, 115-119.

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

Redington, M. (1998). *On hybrid models of artificial grammar learning: Commentary on Knowlton and Squire (1996) and Meulemans and Van der Linden (1997).* Unpublished manuscript.

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

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

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

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

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, **25**, 747-763.

(Manuscript received September 20, 1999;
revision accepted for publication March 21, 2000.)