Implicit Learning of Conjunctive Rule Sets: An Alternative to Artificial Grammars

Greg J. Neil and Philip A. Higham

University of Southampton

Author Notes

Greg J. Neil, Psychology, University of Southampton; Philip A. Higham, Psychology, University of Southampton.

Portions of this research were presented at the Annual Meeting of the Psychonomic Society, 2009, Boston, MA. This research was supported in part by an ESRC studentship ES/F02200x/1 to the first author.

Correspondence concerning this article should be addressed to Greg J. Neil, Psychology, University of Southampton, Highfield, Southampton, UK, SO17 1BJ. E-mail: G.Neil@soton.ac.uk.

Abstract

A single experiment is reported that investigated implicit learning using a conjunctive rule set applied to natural words. Participants memorized a training list consisting of words that were either rare-concrete and common-abstract or common-concrete and rare-abstract. At test, they were told of the rule set, but not told what it was. Instead, they were shown all four word types and asked to classify words as rule-consistent words or not. Participants classified the items above chance, but were unable to verbalize the rules, even when shown a list that included the categories that made up the conjunctive rule and asked to select them. Most participants identified familiarity as the reason for classifying the items as they did. An analysis of the materials demonstrated that conscious micro rules (i.e., chunk knowledge) could not have driven performance. We propose that such materials offer an alternative to artificial grammar for studies of implicit learning.

Keywords

Implicit learning; Subjective measures; Artificial grammar; Classification; Chunks

1. Introduction

People can learn regularities in the world seemingly with no intention of doing so and without the ability to verbally describe those regularities, a phenomenon called implicit learning. By far the most common experimental paradigm used to investigate implicit learning is artificial grammar learning (AGL; e.g., Dienes & Scott, 2005; Higham, 1997a, 1997b; Reber, 1967, 1969; see Pothos, 2007 for a review). In typical AGL experiments, participants first observe or attempt to memorize a set of letter strings (e.g., MVXRT) that conform to an underlying rule set (finite-state grammar). After this training phase, participants are informed that all the strings they were just exposed to conformed to a rule set, but they are not informed about its nature. Instead, they are shown new strings during a testing phase that either conform to the rule set (grammatical strings) or not (nongrammatical strings) and asked to classify them accordingly. Many studies have shown that participants can make this discrimination at above-chance levels, but post-experimental interviews indicate that they are unable to verbalize the rule set used during training.

The mechanism(s) underlying AGL have been a subject of considerable debate. Reber (1967, 1969, 1989) has argued that learning occurs because participants abstract the underlying rule set and that this learning is not available to consciousness. Others have contested this assertion, suggesting instead that the learning that takes place is more superficial and/or more conscious in nature. For example, Johnstone and Shanks (2001) and Wright and Whittlesea (1998) argue that people do not distinguish grammatical from non-grammatical stimuli, but perform in ways that are consistent with the demands of the task. Where these demands happen to coincide with grammaticality, participants perform above chance, but without unconsciously abstracting the underlying rule set. Dulany, Carlson and Dewey (1984) argued that people learn simple conscious rules about permissible string fragments (*chunks;* e.g., bigrams or trigrams) within the strings. They demonstrated that although participants could not verbalize the complete rule set, they could identify those parts of the strings that made them grammatical. Furthermore, simulated rule sets from the underlined parts of the strings were enough to almost perfectly reconstruct participants' classification performance. Similarly, Perruchet and Pacteau (1990) argued that AGL is driven by conscious knowledge of sub-sequences of letters in the grammatical strings. They compared performance of a group of participants trained on typical strings used in AGL experiments with a group who only saw valid bigrams in training and found that test performance between the two groups was equivalent. Furthermore, items that were nongrammatical due to non-permissible bigrams were easier to reject than items that were nongrammatical due to permissible bigrams in the wrong place. Their conclusion was that learning was entirely due to conscious bigram knowledge.

Other researchers have also maintained that chunk knowledge is critical to AGL, but placed less emphasis on the acquired knowledge being conscious. For example, in Servan-Schreiber and Anderson's (1990) competitive chunking model, it is assumed that chunks are formed during training and are later used to judge test strings. Frequently occurring letter combinations allow for larger chunks to be formed, and if those larger chunks can be used to process test strings, larger familiarity values result. Because of the inherent structure of grammatical strings, larger chunks can be used to process them at test compared to nongrammatical strings, resulting in them having higher familiarity and more likely to be classified as grammatical. Other frameworks and models have been proposed to account for AGL that also fundamentally rely on superficial knowledge of chunks rather than rule abstraction (e.g., Dienes et al., 1991; Jamieson & Mewhort, 2009a; Knowlton & Squire, 1994; Redington & Chater, 1996; although see Higham, 1997a).

1.1 Conjunctive Rule Sets as an Alternative to Artificial Grammar Stimuli

The nature of AG stimuli is such that regularities in the letter positions are hard to avoid, and so explanations of classification performance in AGL experiments often rely on, or must account for, chunks. Historically, chunking has figured highly in cognitive psychology, with demonstrations of its importance dating at least as far back as Miller's (1956) critical work on short-term memory and Chase and Simon's (1973) classic experiments with chess players. However, in the context of AGL, chunk knowledge, and in particular *conscious* chunk knowledge, is typically seen as a nuisance variable. In short, if conscious chunk knowledge can account for all learning in AGL experiments, as researchers such as Dulany et al. (1984) have proposed, then there actually may be nothing implicit about so-called "implicit learning". This has resulted in attempts to either control chunks or else obscure their presence. For example, Higham (1997a) controlled for chunks in AGL and demonstrated that above-chance performance still occurred and could be affected by factors that did not influence chunks, such as the pronouncibility of the strings. Norman, Price and Jones (2011) obscured the nature of their artificial grammar strings by changing the font and color of the letters used across the test list, finding that even participants that claimed to be responding to these changes classified grammatical strings at above-chance accuracy. Other attempts to sidestep the issue of chunks have involved using non-local rules in music (Kuhn & Dienes, 2005) and even Tang poetry (Jiang et al., 2012). In our view, new materials and paradigms of this nature will enable us to further investigate the wide-ranging questions about human learning raised by AGL experiments whilst sidestepping issues related to the form of the stimuli.

In the current research, we abandoned AGL materials in favor of real English words incorporating a structure that is less likely to involve the influence of micro-rules. Higham and Brooks (1997; see also Higham, Bruno & Perfect, 2010) were the first to use a structure of this sort, so we will explain their methodology and results in some detail here. The training list of their second experiment consisted of natural words that conformed to the conjunction of two word categories. One category denoted the lexical frequency the words (i.e., rare versus common) and the other category denoted the grammatical class of the word (i.e., noun versus verb). Two training lists were constructed using these materials. The first list consisted of 50% rare-nouns (e.g., *hyacinth*) and 50% common-verbs (e.g., *destroy*), whereas the second consisted of 50% common-nouns (e.g., *carpet*) and 50% rare-verbs (e.g., *inculcate*). At test, participants either rated test items as consistent or inconsistent with the training list structure (classification) or as presented earlier in the training list or not (recognition). In both cases, test items were words representing all four conjunctions (i.e., common-nouns, common-verbs, rare-nouns and rare-verbs) and some had been presented earlier in the training phase (old words). For half the participants who were trained with the first training list, rare-nouns and common-verbs were consistent with the training-list structure, whereas rare-verbs and common-nouns were inconsistent. However, the opposite was true for the other half of participants trained on the second training list. For each training list, the data were collapsed across the two different consistent conjunctions and the two different inconsistent conjunctions to yield three stimulus types: old, new-consistent (NC) and new-inconsistent (NI).

Higham and Brooks (1997) found that participants were sensitive to the structure; that is, compared to NI words, NC words were given a higher *consistency* rating in classification and a higher *oldness* rating in recognition (i.e., higher false alarm rate), a difference they dubbed the *structural effect*. They also found that old words (which necessarily were consistent with the structure) were rated higher than NC stimuli, a difference they dubbed the *episodic effect*. However, not a single participant was able to verbalize the structure when asked about it in a post-experimental interview.

Higham and Brooks' (1997) design had a number of positive features for investigating implicit learning. First, because natural words were used instead of meaningless letter strings (as in typical AGL studies), the true nature of the training-list structure was made obscure (i.e., participants were "garden-pathed"). Indeed, the post-experimental questionnaire revealed that several participants were led to the erroneous conclusion that the training-list structure was semantic in nature (e.g., all training-list items belonged to the same semantic category). Second, the design had the advantage that the identifiable categories used to generate the structure (i.e., lexical frequency and grammatical class) were orthogonal to that structure. For example, suppose that during the classification task, a participant remembered that several of the training-list items were rare and so rated rare test items to be consistent with the structure. Regardless of whether participants were exposed to the first or second training list, only half the rare test items were consistent with the structure, so merely responding to the rarity of the test items would reveal no structural effect. For the same reason, if there were particular bigrams or trigrams that occurred more frequently in one category versus another (e.g., the bigram *ns* occurred more frequently in rare words than common words), such knowledge would not be helpful in accurately classifying the test stimuli. Instead, knowledge of the category *conjunctions* was needed to produce a structural effect, which participants showed sensitivity to in their classification and recognition ratings, but which they could not verbalize. Such lack of verbalization suggests that any knowledge of the rule set was held implicitly.

1.2 Experimental Overview

Given that conjunctive rule sets using words have the potential to further our understanding of implicit learning without the confounding of conscious micro rules, the following experiment was conducted in an effort to extend Higham and Brooks' (1997) original research. The first aspect of the extension was to test whether the structural effect occurs with a different conjunctive rule set. The training list in one condition of the current experiment consisted of common-concrete words (e.g., *hotel*) and rare-abstract words (e.g., *tidal*), whereas the second consisted of rare-concrete words (e.g., *kite*) and common-abstract words (e.g., *written*). Higham et al. (2010, Experiment 1) used these specific materials in an experiment on recognition memory and found higher false alarm rates for NC versus NI test items. However, the focus in the current research is on classification and whether or not the structure is available to consciousness, which Higham et al. did not address. Furthermore, we also include here an analysis of the frequency of chunks within the items, which, again, was missing from Higham et al.'s research.

A second extension was to test whether learning would be observed with a 2AFC task in which there was a choice between a NC and a NI word on each trial, without possible influences from old items. These changes were implemented partly to test the generality of learning across different task types (yes/no versus 2AFC), but also to determine if learning would occur in a paradigm more typical of those used in implicit learning research in which all test items were new. Additionally, the 2AFC task would eliminate any possible yes/no response bias that may have contaminated previous results. Higham and Brooks (1997) attempted to eliminate response bias contamination by computing the discrimination index *A'* (Grier, 1971) from signal detection theory. However, *A'* has been shown to have undesirable threshold characteristics under certain circumstances, which means that response bias can still affect it (e.g., Macmillan & Creelman, 2005). The 2AFC design adopted in the current research circumvents these issues because yes/no response bias does not apply.

However, perhaps the most important extension was that we examined the extent to which participants were aware of the rule set in more depth than in Higham and Brooks (1997). Since the publication of Higham and Brooks' experiments, Dienes and Scott (2005) introduced a distinction between *structural knowledge* (i.e., knowledge of the rule set itself) and *judgment knowledge* (i.e., knowledge about whether individual stimuli conform to the rule set). Both types of knowledge can be implicit or explicit. For instance, a person may have high confidence that an item is consistent with the rule set (explicit judgment knowledge) but not know why it is the case (implicit structural knowledge). To measure structural knowledge, we administered a more detailed post-experimental questionnaire than that originally used by Higham and Brooks. To measure judgment knowledge, confidence ratings were required at the time of each individual response (Dienes, Altmann, Kwan, & Goode, 1995). By using both a post-experimental questionnaire and confidence ratings, both the state of participants' judgment knowledge and the state of their structural knowledge could be investigated.

2. Method

2.1 Materials and Design

Four categories of words were drawn from the MRC psycholinguistic database (see Wilson, 1988) – common-concrete words (e.g. *hotel*), rare-abstract words (e.g. *tidal*), rare-concrete words (e.g. *kite*) and common-abstract words (e.g. *written*). Words were classified as either common (frequency of 80+ per million) or rare (frequency of 1 or less per million) by Kucera-Francis written-frequency norms (Kucera & Francis, 1967). Each word was also classified as concrete (rating of 520 or more) or abstract by the MRC concreteness rating which merges data from several sources (e.g., Coltheart, 1981). Due to a shortage of words with a low concreteness rating in the database, abstract words were identified by the experimenter from the set of unrated words in the database. Two training lists were created by randomly selecting sets of 40 words from each of the four categories. Training List A consisted of 40 common-concrete words and 40 rare-abstract words, whereas Training List B consisted of 40 rare-concrete words and 40 common-abstract words. Words were therefore assigned to each training list based upon a conjunctive rule set, which combined two factors.

Each word on a training list could be rule consistent in one of two ways (i.e. common-concrete or rare-abstract on Training List A and rare-concrete or common-abstract on Training List B).

Words used during testing were all NC or NI. A set of 160 word pairs (e.g. hotel/kite) was created that consisted of 40 common-concrete/rare-concrete pairs, 40 common-concrete/common-abstract pairs, 40 rare-abstract/common-abstract pairs and 40 rare-abstract/rare-concrete pairs. With these pairings, one word in each pair was always a NC word and the other word in the pair was always a NI word regardless of the Training List participants were exposed to. Each type of rule violation (frequency or concreteness) was equally represented across pairs because the words in each word pair were always matched on either frequency or concreteness but not both (e.g., if both of the words in a pair were rare, then one word was abstract and the other was concrete). From the set of 160 word pairs, two test lists of 80 word pairs were compiled by randomly assigning 20 word pairs of each type to each list.

A questionnaire was also used to assess verbalizable knowledge of the rule set. The questionnaire consisted of six questions. The first four questions asked participants what they believed the rule set was, and if there were any rules they considered but rejected. The fifth question gave participants a list of possible categories that may have been included in the rule set and allowed them to select as many or as few as they liked, provided that they offered further detail for any selected category. The following categories were the options: word length, number of syllables in the words, grammatical class of the words (noun/verb/adverb), the number of letters, the words' meaning, the familiarity of the words, the words' lexical frequency (e.g. rare or common), the words' association to other words, the words' likely position in a sentence, and the words' concreteness (e.g. concrete or abstract). The final question directly informed the participant that the rule set was a conjunctive rule set involving two of the categories, and asked what the participant thought the two categories were and how they were related.

2.2 Procedure

All participants signed consent forms before completing the experiment on an Apple Macintosh computer using a script implemented in Runtime Revolution. In the training phase, 33 participants were shown words in sets of eight from one of the two training lists and asked to rate each word for understanding on a scale of 1 (did not understand the meaning) to 4 (fully understood the meaning). Half of the participants were given Training List A and the other half were given Training List B, with presentation order being randomized separately for each participant.

Following the training phase, all participants were given the first test phase. Participants were informed of the existence of the rule set but not what it was, following which they were given the word pairs one at a time and were asked to complete a classification task in which they had to choose the NC word in each pair. They were then asked to make a confidence judgment on a scale of 50-100 about their classification, with 50 being described as a guess and 100 as complete certainty. All participants then completed a second test phase that was identical to the first, with the addition of an attribution question about whether the basis of their judgment was random chance, intuition, memory or rules. The test was self-paced – for each word all judgment prompts were present on screen simultaneously and participants initiated the next trial by clicking a button with the mouse. Participants were prevented from moving on to the next trial until all information had been entered. Assignment of test list to test phase was counterbalanced across participants. Presentation order of the word pairs at test was randomized anew for each participant.

After the two tests phases, all participants completed the questionnaire as described above, with questions administered one at a time.

3. Results

In the interests of brevity, we collapsed the two test blocks except where analyzing the attribution data.

3.1 Accuracy and Confidence

Means for accuracy by training list, overall accuracy, accuracy for responses given 50% confidence ratings, confidence for correct answers, and confidence for incorrect answers are presented in Table 1. Mean classification accuracy was higher than chance (50%), for List A, List B and for both lists combined, $t(15) = 7.00$, $p < .001$, $d = 1.75$; $t(16) = 5.29$, $p < .001$, $d = 1.28$; and $t(32) = 8.29$, $p < .001$, $d = 1.10$, respectively. Like Higham and Brooks' (1997) and Higham et al.'s (2010) participants, our participants acquired sensitivity to the studied conjunctive rule.

Mean confidence ratings for correct answers were compared to those for incorrect answers to examine the status of their judgment knowledge. Although the two means were very close, confidence for correct answers was significantly higher than that for incorrect answers, $t(32) = 3.24$, $p = .003$, $d = .15$, suggesting that judgment knowledge was explicit.

3.2 Chunk Frequency

To determine whether participants achieved above-chance classification accuracy on the basis of chunks, a chunk-strength analysis was conducted on the stimuli. First, the bigrams in each word for each training list were counted (e.g., the word "table" has four bigrams in it: "ta", "ab", "bl" and "le"). Each test stimulus could then be expressed in terms of its average bigram chunk strength. For instance, suppose that across the training-list items, "ta" had occurred 20 times in the training list, "al" had not appeared at all, and "le" had appeared 16 times. With these frequencies, the test word "tale" would have an average bigram chunk strength of (20+0+16)/3 = 12. This counting procedure was repeated for the trigrams in the stimuli. The mean of the bigram and trigram chunk strengths for each word was also calculated, a measure we refer to as *associative chunk strength*. The mean chunk strengths for bigrams, trigrams and associative chunk strength are presented in Table 2.

The bigram strengths were entered into a 2 x 2 between-subjects Analysis of Variance (ANOVA) with training list (A versus B) and word type (NC versus NI) as between-subject factors. There were no significant main effects and no interactions, highest $F < 1$. In fact, numerically the NI bigram strength was slightly higher than the NC bigram strength for each training list. An analogous ANOVA on the trigram chunk strengths yielded an effect of rule set, $F(1, 636) = 7.05$, $p = .008$, $\eta^2 = .01$, reflecting higher chunk strengths for Training List A than Training List B. There were no other effects, highest $F < 1$. An ANOVA on associative chunk strength yielded no effects at all, highest $F < 1$. Crucially, there were no main effects of word type for bigrams, trigrams or associative chunk strength. Together, these analyses suggest that the above-chance accuracy observed with classification was not supported by an unexpected confound between chunk strength and the conjunctive rule used to construct the study and test materials.

Although it would appear that responding to the overall chunk strength of the stimuli would not result in above-chance performance, it is possible that participants responded to the *difference* between the chunk strengths of each individual pair. If this were the case, then participants' endorsement rates would have been correlated with the difference in chunk strength between the NC and NI words across trials. For each participant, a point-biserial correlation between accuracy of response (correct or incorrect) and chunk strength differential (NC chunk strength – NI chunk strength) was computed. The mean correlation coefficient across participants was then compared to zero. For bigrams the mean correlation ($M = 0.02$, $SEM = 0.02$) was no different from zero, $t(32) = 1.13$, $p = .27$. The same was true of trigrams ($M = 0.01$, $SEM = 0.01$), $t(32) = .44$, $p = .66$ and of associative chunk strength (M

= 0.03, *SEM* = 0.02), $t(32) = 1.17$, $p = .25$. These analyses suggest that participants were not responding to the difference in chunk strength between NC and NI words.

3.3 Basis of Judgment

In Test Phase 2, participants attributed each decision to random chance, intuition, memory or rules. Mean accuracy, confidence and proportion of use for each attribution are presented in Table 3, with accuracy for each attribution type being proportion of correct classification responses that were assigned that attribution.

Accuracies for the attribution types were entered into a one-way, repeated-measure ANOVA. Only those participants who had used all four attributions were entered into the ANOVA. Dienes and Scott (2005) found higher accuracy for memory and rules attributions than for guess and intuition attributions. Here, there were no differences in accuracy by attribution at all, $F(3, 51) = 1.11$, $p = .35$. Accuracies for intuition and memory were above chance, $t(32) = 3.19$, $p = .003$, $d = 0.56$ and $t(23) = 2.28$, $p = .03$, $d = 0.46$, respectively. However, for random chance and rules the accuracy failed to reach above-chance levels, highest $t(32) = 1.88$, $p = .07$.

Confidence by attribution was also entered into a one-way repeated-measures ANOVA. There was a main effect of attribution, $F(3, 51) = 26.22$, $p < .001$, $\eta^2 = .61$. Pairwise comparisons showed that random chance confidence ratings were smaller than intuition, memory and rules confidence ratings, $F(1, 17) = 54.03$, $p < .001$, $\eta^2 = .76$, $F(1, 17) = 32.79$, $p < .001$, $\eta^2 = .70$ and $F(1, 17) = 44.56$, $p < .001$, $\eta^2 = .72$, respectively. Intuition confidence ratings were smaller than memory and rules confidence ratings, $F(1, 17) = 17.38$, $p = .001$, $\eta^2 = .51$ and $F(1, 17) = 23.87$, $p < .001$, $\eta^2 = .58$. Memory and rules confidence ratings were not different, $F < 1$.

3.4 Questionnaire Data

Several measures were examined to see if any participant developed verbalizable knowledge of the rule set. Above-chance performance when participants claim to be guessing indicates non-verbalizable knowledge by the guessing criterion (Dienes et al., 1995). The guessing criterion was tested in the same way as mean accuracy using accuracy for 50% confidence responses, with one participant removed who only made one 50% confidence response. Above-chance performance for 50% confidence responses was found, $t(31) = 4.00$, $p < .001$, $d = 0.70$. However, as can be seen in the analysis of attribution types, testing the guessing criterion using random chance attributions resulted in chance performance

Similar to Higham and Brooks (1997), the questionnaire indicated that no participant developed verbalizable knowledge of the rule set. The penultimate question of the questionnaire invited participants to select as many potential rule categories as they liked from the list of length, number of syllables, category of word (noun/verb/adverb), letters, meaning, familiarity, frequency (e.g. rare or common), association to other words, likely position in a sentence, and concreteness (e.g. concrete or abstract). The number of participants that selected each category can be seen in Table 4. Critically, only 10 participants (30%) selected frequency and only two participants (6%) selected concreteness, the two categories that made up the conjunctive rule set. However, different participants selected each category. Thus, even when given several options and allowed to select more than two, no participant selected both relevant categories.

In the final question, participants were directly told that the training words followed a conjunctive rule set and were asked to describe the rule set. Of the thirty-three participants, nineteen (58%) picked neither of the correct categories, thirteen (39%) selected one out of the two correct categories but incorrectly described its involvement, and one participant (3%) selected the correct two categories but failed to describe the link between them. Thus, no participant could verbalize the rule set even when directly told it was conjunction between two categories.

4. Discussion.

Consistent with Higham and Brooks (1997), participants' classification performance discriminated between NI and NC words at above-chance levels. Using identical materials to those used here, Higham et al. (2010) demonstrated that NC words had a higher false alarm rate than NI words in old/new recognition. However, ours is the first study to demonstrate a structural effect with these particular materials in classification. Furthermore, unlike both of Higham and colleagues' previous studies on word category conjunctions, yes/no response bias could not have created this pattern of performance because a 2AFC rather yes/no design was employed.

More importantly, there was good evidence that the structural effect observed in this research was implicit in nature. First, as explained in detail above, the design made it very unlikely that participants were merely responding to conscious rules about permissible bigrams and trigrams (e.g., Dulany et al, 1984) or other small sub-components of the literal stimuli. Higham (1997a) demonstrated that for pronounceable AG strings, participants relied less on chunk information and more on whole-item processing compared to unpronounceable AG strings, an effect that is likely to also hold for natural words used in our experiment. The lack of word type effects in the chunk analysis further demonstrated that participants could not perform the task on the basis of bigram and/or trigram strength. In fact, in all four comparisons across chunk size (bigram, trigram or associative chunk strength) and Training List (A or B), the chunk strength of NI items was either matched with (one comparison) or numerically larger than (five comparisons) that of NC items (Table 2). Thus if participants had merely responded on the basis of chunk strength, overall there would have been a *negative* structural effect, which did not occur. Furthermore, the correlational analysis demonstrated that participants did not use the difference in chunk strength between each item in a pair to select the items. Instead, to discriminate the NC and NI items, participants must have learned about the conjunction of concreteness and lexical frequency that composed the rule set. For this reason, materials of the sort used here are useful because the learning is unlikely to be based on conscious chunk knowledge.

Further evidence that the observed learning was implicit was derived from the confidence ratings and in-depth questionnaire data, neither of which was available in previous conjunctive rule-learning studies. First, the guessing criterion (Dienes et al., 1995) was met using one measure; participants performed above chance even when they claimed to be guessing (i.e., 50% confidence rating). If participants were conscious of the rule set then there would have been no need to guess, so the guessing-criterion result using 50% confidence ratings suggests that participants were using non-verbalizable knowledge to perform at least part of the classification task. In contrast, when using attributions of random chance the guessing criterion was not met. However, performance was above chance for intuition attributions. Intuition is defined as knowing that you are correct, but not knowing why you are correct (Dienes & Scott, 2005). Thus the attribution data also supports the use of some non-verbalisable knowledge. Second, the in-depth questionnaire data indicated that not a single participant was able to successfully verbalize the rule set on the questionnaire. Even on the final question, for which participants were directly told that the rule was conjunctive, and given a list of candidate categories to choose from, no participant successfully verbalized the rule set. Instead, as in Higham and Brooks (1997), many participants identified semantic attributes of the training-list words as providing the basis of the rule set, attributes such as word meaning (52%), semantic category (48%), and word association (48%; see Table 4). Indeed, most participants failed to identify even one part of the rule set correctly (only 30% for lexical frequency and 2% for word concreteness), and none identified both. Despite this, participants were able to classify stimuli at above-chance levels. This finding mirrors Higham and Brooks' (1997) findings despite our use of a questionnaire that specifically guided participants towards the correct answer.

Although structural knowledge was implicit, judgment knowledge was not. That is, participants' subjective confidence was somewhat sensitive to the accuracy of their classification responses. Participants' intuition and memory based performance was also above chance, both attributions being associated with explicit judgment knowledge. At first glance, explicit judgment knowledge may appear contradictory to the fact that participants could not verbalize the rule set. However, again the in-depth questionnaire provides an account of this finding. Sixty-one percent of participants identified *familiarity* as the basis of their classification decisions, the most frequently chosen category from the available list of choices. Similarly, Higham et al.'s (2010) participants endorsed NC words more than NI words in old/new recognition and explicitly attributed their "old" judgments to *familiarity* on a metacognitive rating task. Familiarity has been implicated as contributing to performance in AGL tasks as well (e.g., Scott & Dienes, 2010; Tunney, 2007). In recognition memory research, familiarity is typically considered to be a vague, automatic process, but despite this, it gives rise to a very conscious feeling that a stimulus was previously encountered (e.g., Yonelinas, 2002). Thus, the explicit judgment knowledge that participants demonstrated likely arose from a conscious subjective feeling of familiarity, the source of which could not be identified. Familiarity could be reflected in both intuition and memory attributions. Thus, the fact that the source of that familiarity was an item's consistency with the rule set meant that participants simultaneously possessed implicit structural knowledge and explicit judgment knowledge.

In conclusion, we believe the current data coupled with those from Higham and Brooks (1997) and Higham et al. (2010) provide a strong case for using conjunctive rule sets with natural words as stimuli in future implicit learning research. Robust structural effects were observed in all cases, and in no cases were participants able to verbalize the rule set even when presented in the current research with the categories making up the conjunction. Instead, responding was primarily attributed to vague feelings of familiarity. Critically, these effects could not have been caused by conscious knowledge of micro rules. Conjunctive-rule word materials are easy to construct, requiring only a pool of words that are selected for different attributes from resources readily available on the Internet. Consequently, conjunctive rule-sets instantiated in natural words offers a credible alternative route to investigate the questions raised by previous AGL experiments. The specific rule set used here is of additional interest because the underlying meaning of the words must be processed in order to learn the rule set, an aspect that is difficult to achieve with artificial stimuli. Using natural words as stimuli has the additional advantage that experimental results can potentially inform recognition memory research as well as implicit learning research. That such an endeavor may be fruitful is evidenced by Higham & Brooks' (1997) discussion of mirror effects and in Jamieson and Mewhort's (2009a, 2009b) more recent attempts to account for implicit-learning effects with recognition memory models (e.g., MINERVA2; Hintzman, 1988). The results presented here contribute to this on-going effort.

References

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81. doi:10.1016/0010-0285(73)90004-2

Coltheart, M. (1981). MRC Psycholinguistic database user manual: version 1.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1322-1338. doi: 10.1037/0278-7393.21.5.1322

Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 875-887. doi: 10.1037/0278-7393.17.5.875

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research/Psychologische Forschung, 69*, 338-351. doi: 10.1007/s00426-004-0208-3

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. doi: 10.1037/0096-3445.113.4.541

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424-429. doi: 10.1037/h0031246

Higham, P. A. (1997a). Chunks are not enough: The insufficiency of feature frequency-based explanations of artificial grammar learning. *Canadian Journal of Experimental Psychology, 51*, 126-138. doi: 10.1037/1196-1961.51.2.126

Higham, P.A. (1997b). Dissociations of grammaticality and specific similarity effects in artificial grammar learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23,* 1029-1045. doi: 10.1037/0278-7393.23.4.1029

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. The Quarterly of Experimental Psychology, 50A. doi: 10.1080/027249897392297

Higham, P. A., Bruno, D., & Perfect, T. J. (2010). Effects of study list composition on the word frequency effect and metacognitive attributions in recognition memory. *Memory*, *18*, 883-899. doi: 10.1080/09658211.2010.517757

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *The Quarterly Journal of Experimental Psychology, 62*, 550-575. doi: 10.1080/17470210802055749

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *The Quarterly Journal of Experimental Psychology, 62*, 1757-1783. doi: 10.1080/17470210802557637

Jiang, S., Zhu, L., Guo, X., Ma, W., Yang, Z., & Dienes, Z. (2012). Unconscious structural knowledge of tonal symmetry: Tang poetry redefines limits of implicit learning. *Consciousness and Cognition: An International Journal, 21*, 476-486. doi: 10.1016/j.concog.2011.12.009

Johnstone, T., & Shanks, D. R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology, 42*, 61-112. doi: 10.1006/cogp.2000.0743

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, and Cognition, 22*, 169-181. doi: 10.1037/0278-7393.22.1.169

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present day american english*. Providence: Brown university press.

Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*, 1417-1432. doi: 10.1037/0278-7393.31.6.1417

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. doi:10.1037/h0043158

Norman, E., Price, M. C., & Jones, E. (2011). Measuring strategic control in artificial grammar learning. *Consciousness and Cognition: An International Journal, 20*, 1920-1929. doi: 10.1016/j.concog.2011.07.008

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119*, 264-275. doi: 10.1037/0096-3445.119.3.264

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133*, 227-244. doi: 10.1037/0033-2909.133.2.227

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior, 6*, 855-863. doi: 10.1016/S0022-5371(67)80149-X

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81*, 115-119. doi: 10.1037/h0027454

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General, 125*, 123-138. doi: 10.1037/0096-3445.125.2.123

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 592-608. doi: 10.1037/0278-7393.16.4.592

Scott, R. B., & Dienes, Z. (2010). Knowledge applied to new domains: The unconscious succeeds where the conscious fails. *Consciousness and Cognition: An International Journal, 19*, 391-398. doi: 10.1016/j.concog.2009.11.009

Tunney, R. J. (2007). The subjective experience of remembering in artificial grammar learning. *European Journal of Cognitive Psychology, 19*, 934-952. doi: 10.1080/09541440600778083

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, Version 2.00. *Behavior Research Methods, Instruments & Computers, 20*, 6-10.

Wright, R. L., & Whittlesea, B. W. A. (1998). Implicit learning of complex structures: Active adaptation and selective processing in acquisition and application. *Memory & Cognition, 26*, 402-420. doi: 10.3758/BF03201149

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517. doi: 10.1006/jmla.2002.2864

Table 1

Mean Accuracy and Confidence in the 2AFC Classification Task. Standard Error of the Mean is Also Shown.

	Training List A Accuracy (0-100%)	Training List B Accuracy (0-100%)	Overall Mean Accuracy (0-100%)	Accuracy for 50% Confidence (0-100%)	Correct Answer Confidence (50-100%)	Incorrect Answer Confidence (50-100%)
M	59.78	55.55	57.31	55.80	60.30	59.17
SEM	1.31	1.05	0.88	1.45	1.37	1.30

Table 2

Mean Bigram, Trigram and Associative Chunk Strength by Training List and Word Type. Standard Error of the Mean is Also Shown.

Training list and word type	Bigram		Trigram		Associative	
	M	SEM	M	SEM	M	SEM
Training list A						
NC	1.47	0.08	0.34	0.03	0.87	0.05
NI	1.49	0.08	0.34	0.03	0.89	0.05
Total	1.48	0.05	0.34	0.02	0.88	0.03
Training list B						
NC	1.50	0.08	0.25	0.03	0.91	0.05
NI	1.51	0.08	0.26	0.03	0.92	0.05
Total	1.51	0.05	0.25	0.02	0.91	0.04

Note. NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3

Mean Accuracy, Confidence and Proportion of Use for Attribution Choices. Standard Error of the Mean is Also Shown.

Attribution	Accuracy (0-100%)	Confidence (50-100%)	Proportion (0-1)
Random Chance	53.98 (2.12)	52.42 (0.79)	0.46 (0.04)
Intuition	58.89 (2.78)	61.62 (1.24)	0.30 (0.03)
Memory	62.40 (5.43)	72.97 (2.93)	0.09 (0.02)
Rules	58.29 (4.51)	71.37 (2.39)	0.15 (0.03)

Table 4

Mean Number and Mean Percentage of Participants Indicating that a Given Category was Used to Guide Classification (Maximum Possible of 33)

	Measure	
Category Type	Mean Number of "Yes" Responses	Mean Percentage of "Yes" Responses
Word Length	7	21%
Number of Syllables	2	6%
Semantic Category	16	48%
Number of Letters	6	18%
Word Meaning	17	52%
Word Familiarity	20	61%
Lexical Frequency	10	30%
Word Association	16	48%
Word Position	0	0%
Word Concreteness	2	6%