Psychological Research (2016) 80:195–211
DOI 10.1007/s00426-015-0660-2

ORIGINAL ARTICLE

Information theory and artificial grammar learning: inferring grammaticality from redundancy

Randall K. Jamieson[1] · Uliana Nevzorova[1] · Graham Lee[2] · D. J. K. Mewhort[2]

Received: 21 May 2014 / Accepted: 9 March 2015 / Published online: 1 April 2015

Abstract In artificial grammar learning experiments, participants study strings of letters constructed using a grammar and then sort novel grammatical test exemplars from novel ungrammatical ones. The ability to distinguish grammatical from ungrammatical strings is often taken as evidence that the participants have induced the rules of the grammar. We show that judgements of grammaticality are predicted by the local redundancy of the test strings, not by grammaticality itself. The prediction holds in a transfer test in which test strings involve different letters than the training strings. Local redundancy is usually confounded with grammaticality in stimuli widely used in the literature. The confounding explains why the ability to distinguish grammatical from ungrammatical strings has popularized the idea that participants have induced the rules of the grammar, when they have not. We discuss the judgement of grammaticality task in terms of attribute substitution and pattern goodness. When asked to judge grammaticality (an inaccessible attribute), participants answer an easier question about pattern goodness (an accessible attribute).

Introduction

In an artificial grammar task, participants study stimuli constructed according to the rules of a finite-state grammar (e.g., Fig. 1). Following study, the participants are asked to sort novel grammatical from ungrammatical test items. Typically, they can sort the grammatical from the ungrammatical items, but they cannot articulate the rules.

Some theorists have taken the ability to sort grammatical from ungrammatical test items as evidence for implicit rule induction (e.g., Reber, 1967, 1989; Knowlton & Squire, 1994, 1996). Others argue that participants judge a test string's grammaticality on the basis of its similarity to the studied list (e.g., Brooks, 1978; Jamieson & Mewhort, 2009a, 2010; Perruchet & Pacteau, 1990; Pothos & Bailey, 2000; Vokey & Brooks, 1992).

Although the similarity-based position accommodates a majority of the data in the implicit-learning database, it cannot handle the transfer version of the standard task (e.g., Reber, 1969; Manza & Reber, 1997). In the transfer task, participants study grammatical training strings composed of one set of letters but are tested on strings composed of different letters. For example, they might study strings such as *MTVXRM* and *TXRMMV* and, then, judge the grammatical status of test strings such as *BQHZ* and *PKQZZB*. Changing the letters obscures the similarity of the training and test items; hence, successful transfer challenges the similarity-based accounts. As a result, success in the transfer task reinvigorates the case for implicit rule induction (e.g., Knowlton & Squire, 1996; Mathews et al., 1989).

Brooks and Vokey (1991) rebutted the reinvigorated claim for rule induction by arguing that participants use both analogical and literal similarity to judge test strings. For example, participants might endorse the test string *BCCCCD* as grammatical because its pattern of unique and repeated symbols is similar to the pattern in the training string *VMMMMR* (see also Vokey & Higham, 2005; Lotz & Kinder, 2006). We take Brooks and Vokey's point, but we note that specific pattern similarity is a special case of a more general stimulus property, local redundancy (see Garner, 1962, 1974; Jamieson & Mewhort, 2005).

✉ Randall K. Jamieson
randy.jamieson@umanitoba.ca

[1] Department of Psychology, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

[2] Department of Psychology, Queen's University at Kingston, Kingston, Canada

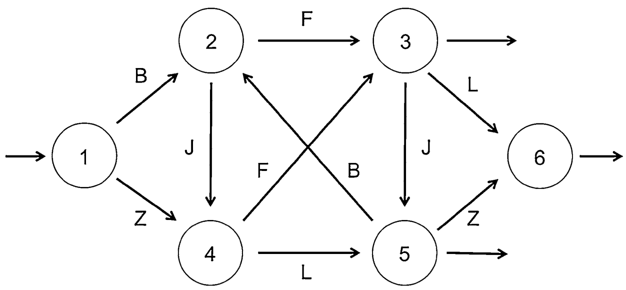

Fig. 1 The finite-state grammar used in Experiments 3 and 4. A grammatical string is constructed by entering the grammar at the leftmost node marked *1* and following the paths (indicated by *arrows*) until reaching an exit path from nodes *3*, *5*, or *6*. When a path is taken, the associated letter is added to the end of the string. For example, moving from nodes *1* to *2*, *2* to *3*, *3* to *5*, *5* to *2*, and *2* to *3* produces the string *BFJBF*

Information theory has been used to examine and predict behavior in a number of research domains. Garner and Clement (1963) used information theory to predict peoples' preference for visual patterns. Hick (1952) and Hyman (1953) used it to predict performance in speeded choice. Miller (1956) used information theory to quantify the processing limits of immediate memory (see also Miller, Bruner & Postman, 1954). Berlyne (1971) used it to relate stimulus complexity to arousal, and Meyer (1956) used it to relate musical structure to emotion. More recently, Jamieson and Mewhort (2005, 2009b; see also Pothos, 2010) used information theory to predict serial recall and judgement of grammaticality in the artificial grammar task.

Informally, redundancy refers to the stability of relations between elements within a sequence. For example, the relations between the letters *T*, *V*, and *X* are more regular and thus more redundant in the sequence *TVXTVX* than in the sequence *XTVTXV*. In general, redundant sequences have the quality of being well patterned, whereas non-redundant sequences do not.

Formally, redundancy is a statistic that quantifies the degree of internal constraint in a sequence (e.g., Attneave, 1959; Garner, 1962; Shannon & Weaver, 1949). Redundancy ranges from 0 to 1. If redundancy equals 0, the sequence is perfectly unpredictable (i.e., random). If it equals 1, the sequence is perfectly predictable. If redundancy is intermediate between 0 and 1, the sequence is partially predictable (i.e., structured but not determined).

In the context of the artificial grammar task, zero-order local redundancy, R_0, indexes the predictability of individual letters in a string,

$$R_0 = 1 - \frac{-\sum_{i=1}^{m} p_i \log_2(p_i)}{\log_2(n)}, \tag{1}$$

where m is the number of symbols in the string, n is the string length, and p_i is the probability of symbol i in the string. To illustrate the measure, consider the three strings *ABABABAB*, *ABCDABCD*, and *ABCDEFGH*. *ABABABAB* has two unique symbols $\{A, B\}$ both of which occur with probability 0.5. Therefore, *ABABABAB* has a zero-order redundancy equal to 0.67. *ABCDABCD* has four unique symbols $\{A, B, C, D\}$ all of which occur with probability 0.25. Therefore, *ABCDABCD* has a zero-order redundancy equal to 0.33. Finally, *ABCDEFGH* has eight unique symbols $\{A, B, C, D, E, F, G, H\}$ all of which occur with probability 0.125. Therefore, *ABCDEFGH* has a zero-order local redundancy equal to 0. Note that if the number of unique letters in a string, m, is equal to the number of letters in the string, n, the numerator in Eq. 1 (i.e., the Shannon entropy), $-\sum_{i=1}^{m} p_i \log_2(p_i)$, is equal to the denominator, $\log_2(n)$.

Strings can also be measured at the first order of local redundancy, R_1, which indexes the predictability of bigrams rather than letters within a string,

$$R_1 = 1 - \frac{-\sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij} \log_2(p_{ij})}{\log_2(n-1)}, \tag{2}$$

where m is the number of unique symbols in the string, n is the string length, and p_{ij} is the probability that letter i follows letter j in the string. Like R_0, R_1 ranges between 0 and 1 with larger values indicating greater sequential predictability. Note that maximal uncertainty is now equal to $\log_2(n-1)$ because there are only $n-1$ bigrams in a string of length n. In summary, local redundancy provides a defined and formal measure of stimulus structure, a variable that is known to predict judgements of well formedness.

In this paper, we investigate the role of local redundancy in judgments of grammaticality. Specifically, we assess the proposition that people judge the grammaticality of a string by its redundancy, not by its grammaticality. Because local redundancy does not require a comparison of the test string to the training list, and because redundancy is independent of the letters used to construct an item, we expect that participants will use redundancy similarly in both the standard and transfer versions of the grammaticality task. Finally, assuming that grammaticality and redundancy are confounded, judgement of grammaticality by redundancy will yield an illusion that participants know the grammar when, in fact, they may not.

Experiment 1

Experiment 1 asks if local redundancy predicts participants' ratings of grammaticality in the transfer task. The experiment included a training phase followed by a test phase. In the training phase, participants studied grammatical training strings instantiated using one set of letters. In the test phase, participants judged the grammatical status of novel test strings constructed using a different set of letters. We allowed redundancy to vary freely, with the intent to examine the relationship between judgements of grammaticality and local redundancy after the fact.

Method

Participants

Eight undergraduate students from Queen's University participated in the experiment. All reported normal or corrected-to-normal vision.

Apparatus

The experiment was administered on a desktop computer equipped with a 17-inch. CRT monitor, a standard keyboard, and a standard mouse.

Stimuli

Materials were constructed using the grammar in Table 1.

The grammar in Table 1 presents the probabilities with which letters *B*, *C*, *D*, *F*, *G*, *H*, *J,* and *K* could follow one another in successive serial positions of a string. For example, the letter *B* could be followed by the letters *F* and *K* but could not be followed by itself or the letters *C*, *D*, *G*, *H,* or *J*.

We used the grammar to generate 60 unique grammatical and 20 unique ungrammatical strings. All strings were eight letters in length. To construct a grammatical string, we selected the first letter at random and then selected letters to the subsequent positions according to the transition probabilities in the grammar. If the sequence was already included in the stimulus list, it was discarded and replaced. The process was repeated until we had constructed 60 unique grammatical sequences.

Table 1 The grammar used to construct stimuli in Experiment 1

Letter in position n	Letter in position $n + 1$							
	B	*C*	*D*	*F*	*G*	*H*	*J*	*K*
B	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5
C	0.0	0.5	0.0	0.0	0.0	0.0	0.5	0.0
D	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5
F	0.5	0.0	0.0	0.5	0.0	0.0	0.0	0.0
G	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.5
H	0.0	0.0	0.0	0.5	0.0	0.0	0.5	0.0
J	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.5
K	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.5

The grammar shows transition probabilities for letters *B*, *C*, *D*, *F*, *G*, *H*, *J*, and *K* in position n to position $n + 1$ of a sequence

The ungrammatical sequences were constructed by selecting letters at random and with replacement from the set $\{B, C, D, F, G, H, J, K\}$ to each of the eight positions in the string. Sequences that did not include at least one illegal transition or that were already included in the list were discarded and replaced.

Forty of the grammatical strings were assigned to the training list. The remaining 20 grammatical strings and the 20 ungrammatical strings were assigned to the test list. The test items were rewritten by randomly mapping the letters *L, N, P, R*, *S*, *T*, *Y,* and *Z* to the letters *B*, *C*, *D*, *F*, *G*, *H*, *J*, and *K*, and rewriting the test strings accordingly. The full list of materials is presented in "Appendix 1".

Procedure

The participant was seated at a computer terminal. The word "Study" was printed at the center of the screen. The participant was told that letter strings would be presented on the screen and that it would be his or her job to read each string silently.

When the participant clicked on the word "Study", the screen was cleared. Three-quarters of a second later, the first training string was presented for 6 s after which the screen was cleared; 750 ms later, the next training string was presented. The cycle continued until all 40 of the training strings had been shown.

After presenting the training items, instructions for the test phase were provided. The instructions indicated (a) that the study materials had been generated using rules, (b) that the task was to rate the rule compliance of novel test strings, (c) that half of the test strings would follow the rules and half would not, and (d) that the test strings would be constructed using different letters than the training strings. The participant clicked on a word "Begin" to start the test phase.

On each test trial, a string was presented at the center of the computer screen. A response tool was presented below it. The response tool was a line approximately 4 cm in length. Tick marks visibly dissected the line into four equally spaced regions. However, the line was invisibly dissected into 100 points corresponding to numbers -100 to -1 to the left of the midpoint (i.e., 0) and 100 points corresponding to numbers 1–100 to the right of the midpoint. A slider was positioned at the center of the line (i.e., the position corresponding to zero). The phrases "Does not conform to the rules" and "Conforms to the rules" were displayed to the left and right of the line, respectively. The word "OK" was displayed below the line.

To submit a response, the participant moved the slider to a position along the line and then clicked on the word "OK". When the participant clicked on "OK", their rating between -100 and $+100$ was recorded to a response file, the screen was cleared, and after a 1 s pause, the next test string was presented. If the participant clicked "OK" without moving the slider, a message was presented instructing the participant to move the slider. The cycle continued until all of the test strings had been presented, and the participant had provided a response to each one.

Following the series of test trials, a text editor was presented along with a message that invited the participant to describe the rules of the grammar. The participant used the computer keyboard to type the rules.

Results and discussion

The mean rating for grammatical strings ($M = 12.95$, $SD = 4.68$) was higher than the mean rating for ungrammatical strings ($M = -4.34$, $SD = 8.81$), $t(7) = 7.67$, $p < 0.001$. Participants discriminated grammatical from ungrammatical test strings.

We computed the correlation between each participant's item ratings against calculations of both zero- and first-order local redundancy. The mean correlation between ratings and zero-order local redundancy was 0.27 ($SD = 0.21$), $t(7) = 3.67$, $p < 0.01$; the mean correlation between ratings and first-order local redundancy was 0.25 ($SD = 0.21$), $t(7) = 3.33$, $p < 0.02$.[1]

Finally, we measured the confounding between grammatical status and local redundancy. As expected (see Jamieson & Mewhort, 2005), grammatical test strings were more redundant than ungrammatical test strings at both the zero- ($M = 0.38$ versus $M = 0.24$), $t(38) = 3.29$, $p < 0.01$, and first- ($M = 0.20$ versus $M = 0.02$), $t(38) = 4.18$, $p < 0.001$, orders of redundancy.

In summary, participants' judgements are consistent with discrimination based on both grammatical status and local redundancy. However, because the two factors are confounded, it remains unclear which factor participants used. To disentangle the two factors, we must separate grammaticality from redundancy. But, another issue must be addressed first.

All of our participants identified single-letter runs as a salient characteristic of the materials (e.g., *S<u>LLLLLL</u>R*). Importantly, several participants indicated that they used single-letter runs as a basis for judgement (Kinder & Assmann, 2000; Redington & Chater, 1996).

A post hoc analysis confirmed that participants did rate strings with single-letter repetitions as more grammatical than strings without. However, it also revealed that our materials not only confounded grammaticality and redundancy but also confounded both of those factors with single-letter runs: 90 % of the grammatical strings included a single letter but only 45 % of the ungrammatical strings did. Because redundancy measures predictability, and because single-letter runs introduce predictability, local redundancy is necessarily confounded with their occurrence. Such multiple confounding not only illustrates the difficulty of identifying the basis of peoples' decisions, but also points to the need to bring stimulus properties under experimental control.

Experiment 2

In Experiment 2, we assess the role of single-letter runs in the results from Experiment 1. To do so, we repeated Experiment 1, but excluded single-letter runs from both training and test strings. If participants relied on single-letter runs but not redundancy to judge the grammaticality of test strings, their judgements should no longer correlate with redundancy (or perhaps grammaticality).

Method

Participants

Eight undergraduate students from Queen's University participated in the experiment. All reported normal or corrected-to-normal vision.

Apparatus

We used the same apparatus as in Experiment 1.

Stimuli

Materials were constructed in the same way as in Experiment 1. However, grammaticality was defined by the grammar in Table 2. As shown, each letter could be followed by two others in the set; hence, the two grammars have equal grammatical redundancy. However, the grammar differed in two ways. Firstly, no letter could be followed by itself, thereby eliminating single-letter runs. Secondly, the letter *H* was exchanged for the letter *N* in the training list and the letter *N* was exchanged for letter *X* when rewriting strings in the test list.[2] The materials are presented in full in "Appendix 2".

[1] Collapsing ratings over participants, the correlation between mean ratings and zero-order redundancy was $r(38) = 0.47$, $p < 0.01$, whereas the correlation between mean ratings and first-order redundancy was $r(38) = 0.45$, $p < 0.01$. Although the correlations computed in this way are more impressive, they are subject to an overestimation bias (see Lorch & Myers, 1990).

[2] We made the change because letters *H* and *N* looked similar to one another.

Table 2 The grammar used to construct stimuli in Experiment 2

Letter in position n	Letter in position $n+1$							
	B	*C*	*D*	*F*	*G*	*N*	*J*	*K*
B	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5
C	0.0	0.0	0.5	0.5	0.0	0.0	0.0	0.0
D	0.0	0.5	0.0	0.0	0.5	0.0	0.0	0.0
F	0.0	0.0	0.5	0.0	0.5	0.0	0.0	0.0
G	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.5
N	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.5
J	0.0	0.5	0.0	0.0	0.5	0.0	0.0	0.0
K	0.0	0.0	0.5	0.5	0.0	0.0	0.0	0.0

The grammar shows transition probabilities for letters *B*, *C*, *D*, *F*, *G*, *H*, *N*, and *K* in position n to position $n + 1$ of a sequence

Procedure

The procedure was the same as Experiment 1.

Results and discussion

Results were consistent with Experiment 1. The mean rating for grammatical strings ($M = 6.81$, $SD = 7.41$) was higher than the mean rating for ungrammatical strings ($M = -2.27$, $SD = 9.16$), $t(7) = 2.45$, $p < 0.05$. The mean correlation between ratings and zero-order local redundancy was equal to 0.43 ($SD = 0.15$), $t(7) = 8.31$, $p < 0.0001$, and the correlation between ratings and first-order local redundancy was equal to 0.38, ($SD = 0.21$), $t(7) = 5.20$, $p < 0.01$, $p < 0.01$.[3]

Although the data resolve one uncertainty (i.e., the correlation between ratings and redundancy was not based solely on the confounding with single-letter runs), they still do not distinguish whether performance reflects knowledge of grammaticality or redundancy. To assess the distinction, we tested discrimination in a third experiment where the materials unconfounded grammaticality and redundancy.

Experiment 3

As before, participants studied grammatical strings and then judged the grammatical status of test strings. Participants assigned to a *standard condition* were shown training and test strings constructed with the same letters. Participants assigned to a *transfer condition* were shown training strings constructed with one set of letters and test strings constructed with a different set of letters. Participants assigned to no-study *control condition* rated the grammaticality of test strings without the benefit of study.

We designed the test items so that half were grammatical and half were ungrammatical. We also balanced redundancy across the factor of grammaticality: within each class, one-third of the strings had low first-order redundancy, one-third had medium first-order redundancy, and one-third had high first-order redundancy.

If participants' discrimination of grammaticality in Experiments 1 and 2 reflects a confounding of redundancy and grammaticality, the participants should not discriminate grammatical from ungrammatical strings in the current experiment. Instead, we anticipate that participants' ratings will track redundancy in all three of the study conditions (standard, transfer, and control).

Method

Participants

Forty-eight students from the University of Manitoba participated in the study. Sixteen participants were assigned to the standard condition, 16 to the transfer condition, and 16 to the no-study control condition. All participants reported normal or corrected-to-normal vision.

Apparatus

The experiment was administered on desktop computers. Each computer was equipped with a 22-inch. wide- and flat-screen monitor, a standard keyboard, and a standard mouse.

Materials

The stimuli were consonant strings constructed using the grammar in Fig. 1. To select the materials, we generated all 72 strings that included at least four and no more than eight letters. Next, we selected 24 of the 72 grammatical items to the test list: Eight of the test items had low redundancy ($M[R_0] = 0.20$, $M[R_1] = 0.00$), eight had medium redundancy ($M[R_0] = 0.32$, $M[R_1] = 0.14$), and eight had high redundancy ($M[R_0] = 0.33$, $M[R_1] = 0.25$). Then, we rearranged the letters in each of the grammatical test strings to generate a corresponding ungrammatical test string with the same (or nearly the same) first-order local redundancy. Of the 24 ungrammatical test strings, eight were low redundancy ($M[R_0] = 0.20$, $M[R_I] = 0.00$), eight were medium redundancy ($M[R_0] = 0.31$, $M[R_I] = 0.15$), and eight were high redundancy ($M[R_0] = 0.33$, $M[R_I] = 0.24$).

After we had constructed the test list, we selected 30 of the remaining 48 grammatical strings to serve as training

[3] Collapsing over participants, the correlation between ratings and zero-order redundancy was $r(38) = 0.77$, $p < 0.01$, whereas the correlation with first-order redundancy was $r(38) = 0.68$, $p < 0.01$. Although the correlations computed in this way are more impressive, they are subject to an overestimation bias (see Lorch & Myers, 1990).

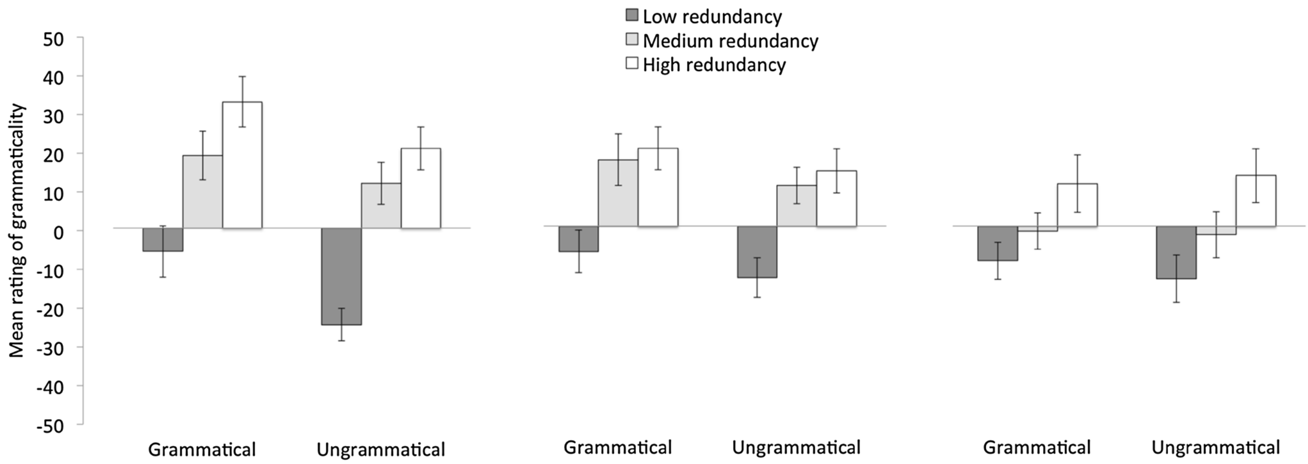

Fig. 2 Mean grammaticality ratings as a function of redundancy and grammatical status. Performance by the standard group is shown in the *left panel*. Performance by the transfer group is shown in the *middle panel*. Performance by the no-study control group is shown in the *right panel*. *Whiskers* indicate one standard error above and below each mean

items. The training list had a mean zero-order local redundancy equal to 0.27 (SD = 0.11) and a mean first-order local redundancy equal to 0.10 (SD = 0.13). The training list in the transfer condition was structurally identical to the training list in the standard condition, except that the letters *X*, *Q*, *H*, *M*, and *D,* were substituted for the letters *B*, *F*, *J*, *L*, and *Z*, respectively. The full set of materials is presented in "Appendix 3".

Procedure

The procedure was nearly identical to procedure from Experiments 1 and 2 with four exceptions: (a) Participants were tested in groups with each participant seated at a different computer terminal; (b) each training string was presented for 8 rather than 6 s; (c) the training list included 30 rather than 40 grammatical strings; and, (d) the phrases "Conforms to the rules" and "Does not conform to the rules" were replaced with the phrases "Grammatical" and "Ungrammatical" in the response tool.

Results and discussion

Figure 2 shows the mean grammaticality rating (slider position) for each class of test item as a function of both grammaticality (grammatical and ungrammatical) and local redundancy (low, medium, and high). Whiskers indicate one standard error above and below each mean. Performance of the standard group is shown on the left. Performance of the transfer group is shown in the center. Performance by the no-study control group is shown on the right.

We analyzed the data using a three-factor ANOVA model with grammaticality as a two-level within-subjects factor, redundancy as a three-level within-subjects factor, and test condition as a three-level between-subjects factor. As shown in Fig. 2, participants in all three conditions rated high-redundancy strings ($M = 18.55$, SD $= 25.69$) as more grammatical than medium-redundancy strings ($M = 9.05$, SD $= 23.57$) and rated medium-redundancy strings as more grammatical than low-redundancy strings ($M = -12.25$, SD $= 22.25$), $F(2, 90) = 33.08$, $p < 0.0001$. In contrast, only participants in the standard study condition discriminated grammatical from ungrammatical test items: whereas participants in the standard study condition discriminated grammatical from ungrammatical test strings better than participants in the no-study control condition, $F(1, 45) = 4.55$, $p < 0.05$, participants in the transfer condition did not, $F(1, 45) = 0.94$, $p > 0.30$.

Unlike the data from Experiments 1 and 2, performance in the transfer condition provided no evidence that participants knew the grammar, implicitly or otherwise. Critically, the null result was a possibility anticipated in our experimental design. In summary, after grammatical and ungrammatical strings are balanced for redundancy, participants' use of redundancy to infer grammaticality ceased to distinguish grammatical from ungrammatical test strings. However, there is a potential weakness in our analysis.

Under usual conditions, the amount of structure in a grammatical stimulus derives from the structure in the underlying grammar (Jamieson & Mewhort, 2005). For example, in Experiments 1 and 2, training and test strings were sampled from the grammar at random and, accordingly, grammatical strings were, on average, more redundant than ungrammatical strings. By contrast, in Experiment 3, we were able to separate the two factors, but we had to work hard to do so.

Unfortunately, the strategy of selecting materials to unconfound redundancy and grammaticality may have forced us to distort the participants' view of the underlying grammar. If so, one could argue that participants' failure to discriminate grammatical from ungrammatical test strings in the transfer condition reflects an unfair test rather than a demonstration that participants failed to learn.

To resolve the problem, we used Poletiek and van Schijndel's (2009) *statistical coverage* to measure how well the strings in our training list covered the grammar. Like redundancy, statistical coverage ranges between 0 and 1, with larger values signifying better coverage.

Statistical coverage, SC, of a string is equal to the product of its corresponding first-order transition probabilities in the underlying grammar,

$$\mathrm{SC} = \prod_{i=1}^{n-1} p_{i,(i+1)}, \tag{3}$$

where n is the length of the string, i indexes the serial position in the string, and $p_{i,(i+1)}$ is the transition probability associated with the grammatical transition between letters at serial positions i and $(i+1)$ in the string. For example, using the grammar in Fig. 1, the string *ZFJBFL* has a statistical coverage equal to $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = 0.0046$, whereas the string *ZLBF* has a better statistical coverage equal to $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} = 0.139$. Formally, a string's statistical coverage represents the probability that it is generated by the grammar. Psychologically, a string's statistical coverage represents how well the string represents the grammar.

The measure of string-specific statistical coverage can be extended to quantify the statistical coverage offered by a list of items. List wise statistical coverage, $\mathrm{SC_L}$, is computed as,

$$\mathrm{SC}_L = \sum_{i=1}^{m} \mathrm{SC}_i, \tag{4}$$

where m is the number of strings in the training list and i indexes the $1 \ldots m$ strings. To illustrate, a training list that includes the two strings discussed above would have a list wise statistical coverage equal to $0.0046 + 0.0139 = 0.0185$. The greater the list wise statistical coverage, the better the list covers the grammar.

Disappointingly, the statistical coverage offered by the 30 training strings in Experiment 3 ($\mathrm{SC_L} = 0.0855$) was much smaller than the statistical coverage expected from a random sample of 30 strings from the full list 72 grammatical strings between 4 and 8 characters in length ($M[\mathrm{SC_L}] = 0.1401$). In fact, a Monte Carlo analysis confirmed that the training list's $\mathrm{SC_L}$ of 0.0855 was statistically unlikely relative to a random sampling distribution, $p < 0.07$.

In summary, the training list used in Experiment 3 offered a poor view of the grammar in Fig. 1. Consequently, it may be premature to conclude that participants in the transfer condition cannot learn to discriminate the test strings on the basis of grammaticality.

To resolve the ambiguity, we conducted a fourth experiment. In the experiment, we presented participants with the same test strings. However, we presented participants with a different list of training strings selected to offer good statistical coverage of the underlying grammar, $\mathrm{SC_L} = 0.1955$. If the failure to discriminate grammatical from ungrammatical test strings in Experiment 3 reflects a consequence of poor grammatical coverage, participants in the transfer condition should now be able to discriminate grammatical from ungrammatical test strings.

Experiment 4

We re-ran both the standard and transfer study conditions from Experiment 3. The only difference was that the training list offered two times the grammatical coverage of the training list from Experiment 3.

Method

Participants

Thirty-two students from the University of Manitoba undergraduate participant pool took part in the study. Half of the participants were assigned to a standard condition; the remaining participants were assigned to the corresponding transfer condition. All participants reported normal or corrected-to-normal vision.

Apparatus

See Experiment 3.

Materials

The full set of materials is shown in "Appendix 4".

As shown, the test strings in the current experiment are identical to the test strings presented in Experiment 3. However, the training strings differed. As in Experiment 3, training strings in the transfer condition were rewritten such that the letters *X*, *Q*, *H*, *M*, and *D* replaced the letters *B*, *F*, *J*, *L*, and *Z*, respectively. The list had $\mathrm{SC_L} = 0.1953$, double that of the training list from Experiment 3 and statistically better than the coverage offered by random samples of strings, $p < 0.05$. The strings had a mean zero-order local redundancy equal to 0.65 (SD = 0.31) and a mean first-order local redundancy equal to 0.24 (SD = 0.32).

Procedure

See Experiment 3.

Results and discussion

Figure 3 shows the mean grammaticality rating (slider position) for each class of stimulus as a function of both grammaticality (grammatical and ungrammatical) and local redundancy (low, medium, and high). Performance of the standard study group is shown on the left. Performance of the transfer group is presented on the right.

As shown, participants in both conditions rated high-redundancy strings ($M = 17.96$, $SD = 30.46$) as more grammatical than medium-redundancy strings ($M = 9.95$, $SD = 25.53$) and rated medium-redundancy strings as more grammatical than low-redundancy strings ($M = -3.55$, $SD = 28.69$), $F(2, 60) = 14.82$, $p < 0.0001$. However, discrimination of grammatical from ungrammatical strings depended on the condition. Whereas participants in the standard study condition discriminated grammatical from ungrammatical strings, $F(1, 15) = 25.16$, $p < 0.001$, participants in the transfer condition did not, $F(1, 15) = 3.23$, $p > 0.09$; the test for the two-way interaction confirmed the contrast, $F(1, 30) = 11.51$, $p < 0.01$.

For completeness, we also conducted a series of cross experiment comparison. Firstly, whereas participants in the standard study condition discriminated grammaticality better than participants in the no-study control group from Experiment 3, $F(1, 30) = 14.75$, $p < 0.001$, participants in the transfer condition did not, $F(1, 30) = 0.89$, $p > 0.37$. Secondly, whereas increasing statistical coverage in the training list improved discrimination of grammaticality by participants in the standard condition, $F(1, 30) = 4.65$, $p < 0.04$, it did not affect discrimination by participants in the transfer condition, $F(1, 30) = 0.02$, $p > 0.85$. In fact, a visual comparison of results in Experiments 3 and 4 shows a strong impact of statistical coverage on performance in the standard condition with almost no impact whatsoever on performance in the transfer condition.

Finally, we examined judgements of grammaticality as a function of statistical coverage. Judgements were uncorrelated with the statistical coverage of individual test strings, $r(46) = -0.14$, $p > 0.35$.

Increasing the grammatical coverage of the training list improved discrimination of grammatical from ungrammatical test strings, but the improvement was constrained to performance in the standard test condition. The difference suggests that statistical coverage may only be relevant to an understanding of performance in the standard test. We conclude that participants' failure to discriminate grammatical from ungrammatical strings in the transfer condition of Experiment 3 represents ignorance of the grammar, not an artifact of poor statistical coverage following from our stimulus-selection strategies.

General discussion

We conducted four experiments to examine judgement of grammaticality as a function of local redundancy. In Experiment 1, we showed that judgements in a transfer test are correlated with both grammatical status and local redundancy. In Experiment 2, we showed that the correlation is not constrained to endorsement of strings with single-letter repetitions. In Experiment 3, we unconfounded redundancy from grammaticality and tested discrimination of grammaticality in both the standard and transfer tests. Although participants in both conditions discriminated strings consistent with redundancy, only participants in the standard test condition discriminated strings consistent with grammaticality; performance in no-study group was

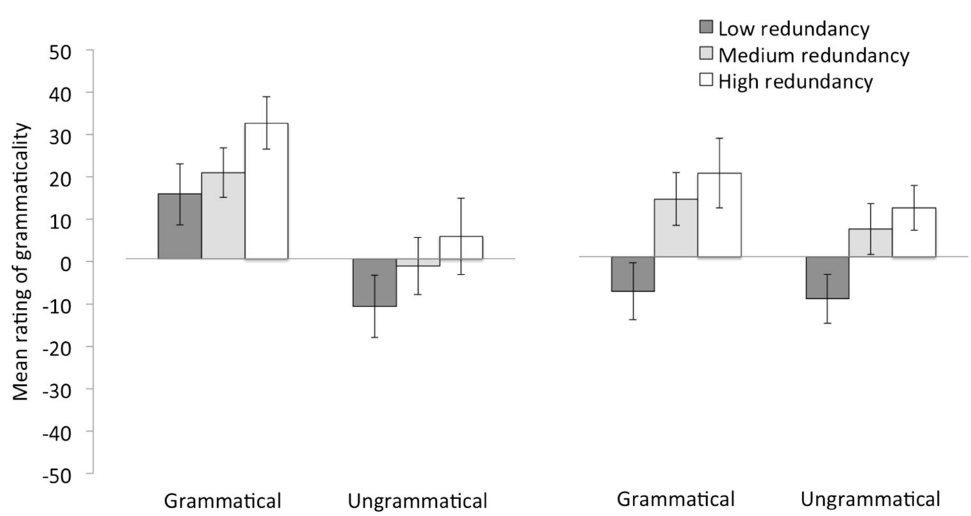

Fig. 3 Mean grammaticality rating as a function of grammatical status and local redundancy. Performance by the standard group is shown on the *left*. Performance by the transfer group is shown in the *right*. *Whiskers* indicate one standard error above and below the mean

also correlated with redundancy. In a final experiment, we increased the statistical coverage of the training list and repeated Experiment 3. Despite the change in materials, results were stable: participants in both conditions discriminated strings consistent with local redundancy, but only participants in the standard condition discriminated strings consistent with grammaticality. Our data show that people use local redundancy to infer grammaticality when test strings are constructed with the same or with different letters than the training strings, and even when the training phase is excluded altogether.

Our examination joins with a history of work focused on identifying the basis for judgements of grammaticality: rules (Reber, 1967), micro rules (Dulany, Carlson & Dewey, 1984), specific similarity (Vokey & Brooks, 1992), fragment similarity (Perruchet & Pacteau, 1990), analogical similarity (Brooks & Vokey, 1991), global similarity (Jamieson & Mewhort, 2009a, 2010), information (Pothos, 2010), local transition probabilities (Poletiek & Wolters, 2009), associative chunk strength (Knowlton & Squire, 1994), statistical coverage (Poletiek & van Schijndel, 2009), single-letter runs (Redington & Chater, 1996), and repetition patterns (Lotz & Kinder, 2006).

The work presented here adds local redundancy to the list. From an empirical standpoint, the addition is useful: a good predictor is valuable, especially if it can be used to uncover and resolve potential confounding. However, it is pertinent to ask why participants judged grammaticality on the basis of redundancy.

Beginning in the 1950s, researchers used redundancy to quantify and predict peoples' perception of *figural goodness*. Garner and Clement (1963), for example, reported a correlation of 0.85 between the two factors. Work that followed extended the analysis to show that people also perceive, remember, classify, and learn redundant patterns both faster and more easily than non-redundant ones (e.g., Clement & Vernadoe, 1967; Garner & Degerman, 1967; Garner & Whitman, 1965; Jamieson & Mewhort, 2005; Mewhort, 1972; Miller, 1958; Miller, Bruner & Postman, 1954; Reber, 1967; Royer & Garner, 1966; Schnore & Partington, 1967; see Chater, 1996, Pothos & Ward, 2000, and van der Helm & Leeuwenberg, 1996, for revisions).

We argue that participants used pattern goodness as a proxy for grammaticality. The explanation fits within a framework of bounded rationality (Simon, 1957) and heuristic decision (Gigerenzer & Brighton, 2009; Kahneman & Tversky, 1974). In particular, it fits with the theory of attribute substitution (Kahneman & Frederick, 2002). According to attribute substitution, when people are asked a hard question, they substitute an easier one. People are particularly prone to engage in attribute substitution when (a) the target attribute is inaccessible, (b) an associated attribute is accessible, and (c) the substitution is applied without feedback. Of course, the judgement of grammaticality task meets all three criteria. When asked to judge the inaccessible attribute of grammaticality, participants substitute the available and intuitive property: pattern goodness. Because participants receive no feedback at test, they have no reason to revise or change their decision strategy.

Explaining performance in the standard test

Participants in the standard conditions discriminated strings by redundancy and grammaticality. One conclusion is that participants learned some aspect of the grammar. Another is that participants judged test strings by their similarity—a factor that we did not control in our materials but that is also naturally correlated with grammatical redundancy (Jamieson & Mewhort, 2005). To evaluate the similarity hypothesis, we applied the holographic exemplar model (HEM) to our materials and procedures.

The HEM is a computational model of memory that combines the representation scheme from Jones and Mewhort's (2007) BEAGLE model of semantic memory with the account of storage and retrieval from Hintzman's (1986) MINERVA 2 model of episodic memory. The theory explains a number of results from the database on artificial grammar learning (Chubala & Jamieson, 2013; Jamieson & Hauri, 2012; Jamieson & Mewhort, 2011) as well as learning in related tasks: grammatical string completion (Jamieson & Mewhort, 2010) and serial reaction time (Jamieson & Mewhort, 2009b).[4]

In the HEM, a letter is represented as an n-dimensional vector. The value of each dimension is sampled at random from a normal distribution with mean zero and variance $1/n$. Letter subsequences (e.g., bigrams and trigrams) are represented by binding the constituent letter vectors. For example, the subsequence AB is encoded by binding the corresponding letter vectors: $AB = \mathbf{a} \otimes \mathbf{b}$, where $\otimes$ denotes non-commutative circular convolution (see Jamieson & Mewhort, 2011, pp. 210–212). Letter strings are represented by summing the subsequences. For example, consider the item $ABCD$. First, we generate a vector for each letter: $A = \mathbf{a}$, $B = \mathbf{b}$, $C = \mathbf{c}$, and $D = \mathbf{d}$. Then, we encode the string by summing all of its subsequences to a single vector: $ABCD = \mathbf{a} + \mathbf{b} + \mathbf{c} + \mathbf{d} + (\mathbf{a} \otimes \mathbf{b}) + (\mathbf{b} \otimes \mathbf{c}) + (\mathbf{c} \otimes \mathbf{d}) + (\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}) + (\mathbf{b} \otimes \mathbf{c} \otimes \mathbf{d}) + (\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} \otimes \mathbf{d})$.

Of course, people do not encode all information in a string nor do they encode the same information about the

[4] Demonstrations of grammatical string completion and serial reaction time learning are reported using a standard version of the MINERVA 2 model. However, those demonstrations are reproducible using the holographic representation scheme in the HEM.

string on different encounters with it (e.g., Wright & Whittlesea, 1998). To acknowledge those facts, we encoded each string in a simulation as a random sample of four subsequences ranging from length one to three. Thus, in our simulations, a string such as *ABCD* might be represented as $\mathbf{a} + \mathbf{b} + (\mathbf{a} \otimes \mathbf{b}) + (\mathbf{b} \otimes \mathbf{c} \otimes \mathbf{d})$ in one simulation but $\mathbf{a} + \mathbf{b} + (\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}) + (\mathbf{b} \otimes \mathbf{c} \otimes \mathbf{d})$ in another.

Memory in the HEM is an m by n matrix, $\mathbf{M}$, where m is the number of independent traces stored in the matrix and n is the number of features in each trace. Memory storage is represented by copying each training item to a row in $\mathbf{M}$. Imperfect encoding is simulated by resetting a proportion of elements in the trace to zero (indicating data loss). The amount of data loss is controlled by a parameter L that specifies the probability of storing a feature correctly; thus, each element in $\mathbf{M}$ has a probability $1 - L$ of reverting to zero.

Retrieval follows a resonance metaphor. Presenting a probe vector, $\mathbf{p}$, to memory causes all traces in memory to activate in parallel. Each trace's activation is a non-linear function of its match to the probe. In the model, the activation of trace i, a_i, is computed as,

$$a_i = \left(\frac{\sum_{j=1}^{n} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{n} p_j^2} \sqrt{\sum_{j=1}^{n} M_{ij}^2}} \right)^3, \tag{5}$$

where $\mathbf{p}$ is the probe, $\mathbf{M}$ is the memory, i indexes the $1\ldots m$ traces in memory, and j indexes the $1\ldots n$ columns in the probe and memory matrix. Non-linearity is introduced in retrieval by raising the similarity metric (the term inside brackets in Eq. 5) to an odd-numbered exponent. The transformation ensures that retrieval selects the traces that match the probe most closely.

The information that is retrieved from memory is a vector called the echo, $\mathbf{c}$. The echo is a weighted sum of all traces in memory, where each trace's contribution to the sum is weighted in proportion to its activation by the probe. It is computed as,

$$c_j = \sum_{i=1}^{m} a_i \times M_{ij} \{\text{for } j = 1\ldots n\}, \tag{6}$$

where a_i is the activation of trace i, $\mathbf{M}$ is the memory, i indexes the $1\ldots m$ rows (i.e., traces) in memory, and j indexes the $1\ldots n$ columns (i.e., stimulus features) in both the echo and memory matrix.

Judgment of grammaticality is predicted by echo intensity, I, which is computed as

$$I = \frac{\sum_{j=1}^{n} p_j \times c_j}{\sqrt{\sum_{j=1}^{n} p_j^2} \sqrt{\sum_{j=1}^{n} c_j^2}}, \tag{7}$$

where $\mathbf{p}$ is the probe, $\mathbf{c}$ is the echo, and j indexes the $1\ldots n$ columns (i.e., stimulus features) in both the probe and the echo. I ranges between -1 and $+1$. The greater the I is, the higher the rating of grammaticality.

We applied the model to the materials from Experiments 3 and 4. We conducted 100 independent simulations with the materials from Experiment 3 and 100 independent simulations with the materials from Experiment 4.

Each simulation had four steps. First, random vectors were generated to represent the letters in the training and test sets. Second, a representation was developed for each training and test item. Third, the representation of each training string was stored to memory. Fourth, the echo intensity was computed and recorded for each test string.

The simulation results are presented in Fig. 4. Results of simulations using the materials from Experiment 3 are

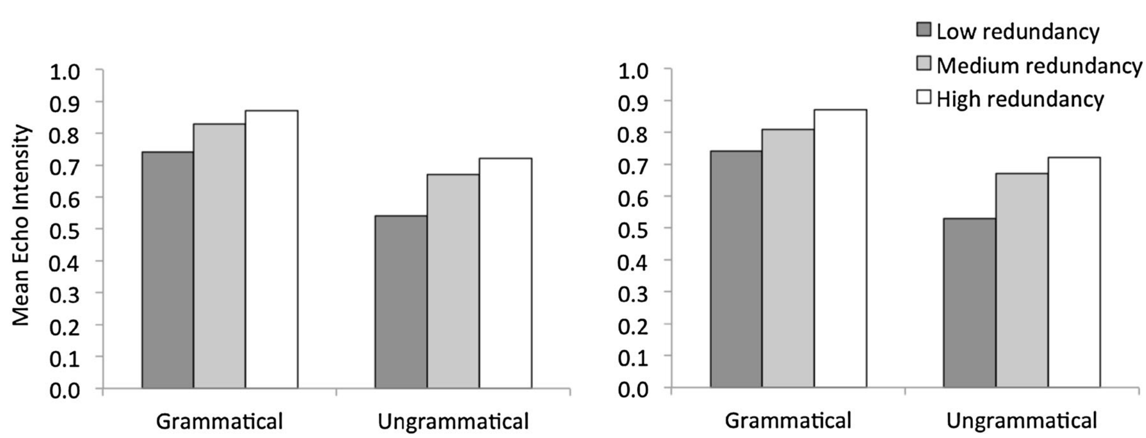

Fig. 4 Mean echo intensity (i.e., global similarity) for test items in the standard test condition in Experiment 3 (*left*) and Experiment 4 (*right*). Predictions are presented as a function of both grammatical status and local redundancy

shown on the left; results of simulations using the materials from Experiment 4 are shown on the right.

As shown, the model captures the key features from our experiments. Firstly, the mean echo intensity for grammatical strings is higher than the mean echo intensity for ungrammatical strings. Secondly, the mean echo intensity for high-redundancy strings is higher than the mean echo intensity for low-redundancy strings. Thirdly, there is a slight advantage in the discrimination of grammaticality in the results of Experiment 4 relative to Experiment 3. We conclude that participants' judgements in the standard condition in Experiments 3 and 4 are consistent with a similarity-based decision strategy.

But, the model has no direct appreciation of redundancy or grammaticality. How, then, did it capture our data? The model's success follows from the confounding of grammaticality with similarity and of similarity with redundancy. Because of the chain of confounding, responding to similarity gives an impression that the model is responding to the other factors as well when, in fact, it is not. The scenario illustrates a complication with drawing inferences from theory (as well as from human performance) in experiments with ill-controlled materials. Unless great care is taken, a model may succeed in predicting peoples' performance for reasons that the model's authors do not fully appreciate.

Statistical coverage

We used Poletiek and van Schijndel's (2009) formulas to improve the statistical coverage of the training list. Although increasing statistical coverage benefitted discrimination of grammaticality in the standard condition, it did not benefit discrimination in the transfer condition. One possible conclusion is that statistical coverage is relevant to an understanding of performance in the standard but not the transfer test. A second possibility is that statistical coverage benefitted participants' discrimination in the transfer test, but that the benefit was so small that we did not detect it. A third possibility is that the measure needs revision.

An analysis of statistical coverage suggests that it is flawed in three ways. Firstly, statistical coverage asserts that shorter strings provide better statistical coverage than longer strings. In some cases, the difference can be quite large: the string *BF* (SC = 0.0833) provides 36 times better statistical coverage than *BFJBJL* (SC = 0.0023), even though *BFJBJL* includes the string *BF*. As Poletiek and Lai (2012, p. 2053) suggest, "we do not know, for example, what the statistical coverage measure corresponds to from a cognitive point of view. The proportion of the full output of a grammar displayed in a sample might correlate with proficiency, but this relation remains to be specified." Secondly, the product rule in the calculation of statistical coverage asserts that each letter added to a grammatical string subtracts from rather than adds to its coverage of the grammar. For example, the string *BF* provides more statistical coverage than the string *BFJ*, even though *BFJ* traverses more paths, contacts more nodes in the grammar, and includes *BF*. Finally, the product rule also asserts that every ungrammatical string is equally unrepresentative of the grammar. But, we suspect that does not hold. For example, *BJFBJ* has a statistical coverage of zero but it is grammatical right up to the last letter. In contrast, *JBLZB* more completely contradicts the grammar. Although we did not test the idea, we suspect that participants trained on slightly ungrammatical strings would discriminate grammatical from ungrammatical test strings better participants trained on very ungrammatical strings.

We conclude that statistical coverage is a useful applied statistic that can be used to analyze materials and predict performance (e.g., Lai & Poletiek, 2011; Poletiek & Lai, 2012; Poletiek & van Schijndel, 2009; Poletiek & Wolters, 2009; Pothos, 2010). However, the concept should be examined and the arithmetic refined.

Calculating local redundancy

Previously, we calculated local redundancy using different formulas (Jamieson & Mewhort, 2005). For example, zero-order redundancy, L_0, was calculated as,

$$L_0 = 1 - \frac{1}{\prod_{i \in A} k_i!}, \quad (8)$$

where A is the set of all possible elements (i.e., letters in the alphabet) and k_i is the number of times that element i appeared in the string. First-order redundancy is calculated by substituting bigrams for letters as the unit of analysis. Like the measures used in this paper, L ranges between 0 and 1 with increasing values representing increasing redundancy. And, in a parallel analysis, we confirmed that our predictions and conclusions were unaffected by the measure we used. Why, then, did not we use our original formulas?

Pothos (2010) noted that our old measure represents a special case of the more general formulation for redundancy. More importantly, there were some inconsistencies in how the concept of an inferred subset maps into the formulas (see Jamieson & Mewhort, 2005). Given Pothos' insight and the fact that adopting the standard formulas had almost no impact on our predictions or outcomes, we elected to use the general formulas.

The confounding is pervasive

We identified a confounding between redundancy and grammaticality in our materials. But, is the confounding active outside of our own experiments?

A brief inspection of materials from published papers suggests the problem we have identified is widespread. Knowlton and Squire (1994) published materials with first-order redundancy greater than zero for 37.5 % of the grammatical but only 12.5 % of the ungrammatical test items. Brooks and Vokey (1991) published materials with first-order redundancy greater than zero for 43.75 % of the grammatical but only 25 % of the ungrammatical test items. Reber and Allen (1978) published materials with first-order local redundancy greater than zero for 24 % of grammatical but only 8 % of ungrammatical test items. In all three examples, discrimination of redundancy would masquerade as discrimination by grammaticality.

Unfortunately, researchers have re-used the confounded materials. Knowlton and Squire's (1994) materials were re-used by Meulemans and Van der Linden (1997). Brooks and Vokey's (1991) materials were re-used by Vokey and Brooks (1992), Vokey and Higham (2005), Lotz and Kinder (2006), Tunney & Shanks, (2003), and Meulemans and Van der Linden (1997). Reber and Allen's (1978) materials were re-used by Dulany, Carlson and Dewey (1984), Altmann, Dienes and Goode (1995), Dienes, Broadbent and Berry (1991), Dienes (1992), Jamieson and Mewhort (2010), and Perruchet and Pacteau (1990). In light of the confounding throughout published work, a re-assessment is needed to examine the influence it has had on the quality and precision of published experimental conclusions.

The relation to statistical learning

In a statistical learning experiment, participants listen to a stream of phonemes ordered according to rules (e.g., *ba*, *di*, *ku*, *pa*, *da*, *bu*, *bi*, *da*, *ku*…). Following training, they attempt to discriminate grammatical from ungrammatical subsequences. As in the artificial grammar test, participants can distinguish valid from invalid subsequences, but they cannot articulate the grammar. The result has been interpreted as evidence of "…a powerful mechanism for the computation of statistical properties of the language input" (Saffran, Aslin & Newport, 1996, p. 1926).

At first blush, the statistical learning procedure appears identical to the artificial grammar task—right down to the conclusion drawn from the results. However, there are important differences. For example, the artificial grammar task evaluates peoples' ability to discriminate grammatical from ungrammatical test items based on memory of studied grammatical exemplars. The statistical learning task, by contrast, measures peoples' ability to parse regular from irregular units within the input stream. Secondly, the scope of local redundancy is defined in the artificial grammar test (i.e., the letter string). In contrast, the input stream is continuous in statistical learning and, therefore, local redundancy will depend upon a subjective decision about where the relevant subsequence begins and ends.

Although we did not apply the measure to an analysis of statistical learning, we see no reason that local redundancy could not also be included as a conceptual and mathematical tool in the analysis of statistical learning. In fact, Garner (1970, 1974) provides a discussion and map that could be used to conduct an analysis of redundancy and the role it plays in the judgement of auditory patterns.

No evidence of successful amateur cryptography

Several participants in our transfer conditions reported a belief that we had mapped letters from the training list to letters in the test list. Although several tried to guess the mapping, only a few were successful. Of those who were successful, no one named more than three of five rewrite rules.

By one analysis, the problem is a serious issue—knowledge of even a single rewrite rule could tip the scales to produce discrimination of grammatical from ungrammatical test strings (Redington & Chater, 1996). However, it is a practical non-issue in our experiments because participants in the transfer conditions did not discriminate test strings by grammaticality.

Nevertheless, we repeated the transfer conditions from Experiments 3 and 4 using item-level rather than list wide rewrite rules (see Vokey & Higham, 2005); a condition that renders cryptographic discrimination strategies inefficient or at least unreliable.

Despite the change in materials, performance was stable: participants discriminated test strings consistent with redundancy but not grammaticality. The result is entirely anticipated: although item-level rewrite rules might obscure grammaticality and similarity, they have no influence on local redundancy and thus no influence on peoples' judgements based on local redundancy.

On the relationship between grammatical and local redundancy

In previous work, we demonstrated that the local redundancy in individual strings is correlated with redundancy of the underlying grammar (Jamieson & Mewhort, 2005). Given the confounding, one might be tempted to argue that the use of redundancy to infer grammaticality represents an ecologically valid decision strategy (Gigerenzer & Todd, 1999). Although we see nothing wrong with the argument, the conclusion is complicated by the fact that the training items in our experiments tended to be low rather than high redundancy.

One way to test the hypothesis is to present participants with a training list of low-redundancy grammatical strings and, then, to present them with low-redundancy grammatical test strings versus high-redundancy ungrammatical test strings. If participants rate high-redundancy ungrammatical test strings as grammatical and low-redundancy grammatical strings as ungrammatical, their behavior would be consistent with the assumptions about redundancy and in contradiction to the information about redundancy in the training list (i.e., the learning environment).

Regarding inferential limits

Participants in our transfer tests discriminated grammatical from ungrammatical strings. However, they only succeeded when grammaticality was confounded with redundancy (i.e., Experiments 1 and 2). Based on that fact, we suggested that our participants behaved as if they knew the grammar when, in fact, they did not. However, our data cannot force that conclusion.

Even in a carefully designed experiment, there is always a possibility that participants learned the grammar yet failed to exhibit that learning at test. It is also possible that measurements are insensitive to learning. It is also possible that some aspect of the experimental procedure prevented participants from learning. As long as these possibilities remain tenable (which they always are), there is no way to discount the claim that participants can learn the grammar (even if they do not).

The complication points to a limitation of the traditional inferential model: one cannot prove a point from a null difference, even if the null difference appears to be informative. In future work, we will use methods to develop evidence for both the null and the alternative hypotheses (Dienes, 2014; Kruschke, 2011).

Starting small

Lai and Poletiek (2011; Poletiek & Lai, 2012) argue that artificial grammar learning is a scaffolded process, where knowledge of basic constraints must be learned first to support learning of more complex constraints. Their thinking echoes that of Elman (1990) as well as Kinder and her colleagues (see Kinder, 2000; Kinder & Lotz, 2009). Others have argued a simpler case: that rule induction takes time (Mathews et al., 1989; Meulemans and van der Linden, 1997).

From that perspective, the evidence that participants' use pattern goodness to judge grammaticality following a short training session might be re-cast as evidence for learning basic elements in the materials (i.e., stimulus cues) necessary to learn grammatical rules with extended practice. Although our data provide no concrete empirical insight on the issue, it certainly merits examination in future work. Can teaching cues correlated with grammaticality bootstrap learning about untutored and more complex aspects of grammatical structure? Or, is cue learning independent from or parallel to a process of implicit rule induction? Reflecting on the issue, it seems odd that there are so few examinations of prolonged training in artificial grammar learning.

In summary

After studying grammatical exemplars, participants can discriminate novel grammatical from ungrammatical test strings. Yet, they cannot articulate the grammar. Several researchers have taken the discrepancy as evidence for implicit rule induction. Of all the evidence for the rule induction position, performance in the transfer test is the strongest. In four experiments, we unconfounded grammaticality from redundancy (i.e., pattern goodness). When the two features were unconfounded, participants failed to discriminate grammatical from ungrammatical test strings.

Acknowledgments The research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to RKJ and DJKM.

Appendix 1: materials in Experiment 1

Training strings			
BFBFFBKK	KGDKGDFF	CCJHJHFF	GDFFBFFB
DKKGKGKG	DFBKGDKG	JKKGDFFB	KGKKGDKK
KKKGDKKG	DFFFFBKG	HFBKKKKK	BFFBFBKG
JKKGDFBK	BKKKGDFF	BFBKGKKK	GDFBFFFB
CCJKGKKG	BKGKKGDK	DKGKKKKG	CCCCJHJH
FFFFBKKG	BKKKGKKK	CCCJHJKG	HFBKGDKK
HFBFFBKK	KKGKGDKG	CCCCJHJK	GDFBKKGK
BFBKKKKK	CCJKKGDK	HFFFFFFF	GKGKKKKG
GDKKGDFB	KKGDFFFB	DFBKGDKK	KGDKKKGK
JHFBFBFB	HFBKKKKG	DFBFBKGK	BKKKGDKK
Test strings			
Grammatical			
TTTLLLRN	TLRTLLLL	SLRTLLRT	SZTTTLLL
ZTLLRTTL	SLLLLLLR	LLRNRNZT	RNRTTTLL
SLRNRNZT	YRNZTTLL	YRTTTTTT	LLLRNRTT
RNZPNZTL	TLLLLLRT	NZTLLLRN	LLRNRNRN
ZTTLLRTL	RTTLRNZT	YRNRNRTT	SZTTTLLR
Ungrammatical			
SZPNRNZP	YRSTZYSN	TYLLNRYR	LPNPPYPP
PRSRNYST	NSZPNTPT	PZLRZTTY	NRZNTYYS
NYNLTSNP	ZZYYZZNY	SLTSZPRS	SLYPLSTZ
ZLRPRRLL	YTSTPNRR	PRZLSTRR	TLNLPTYS
YNZPSRLP	PTLTYZYT	PRLTPRPS	PSZSYYNZ

Appendix 2: materials in Experiment 2

Training strings			
BFGCFGKD	DGCDGCFD	GCDGCFDG	HKFDCDGK
BKDGKFGK	DGCFDGKD	GKDCDCFD	HKFDGKFD
CDCFDGCD	DGCFGKFG	GKDCDGCF	HKFGCDCF
CDCFGKDG	FDCFDCDC	GKDCDGKD	KFDGCFDG
CFDCDGCF	FDCFGCDC	GKDCFDGC	KFDGKFGK
CFGKDGCF	FGCDGKDC	GKFDGCDC	KFGCDGKD
DCDCDGCF	FGCDGKDG	GKFGKFDG	NCDCDGCD
DCDCFGKD	FGKFDGCF	HDCDCFDG	NCDCFGCF
DCDGCDCD	FGKFGCDC	HDCFDGKF	NCDCFGKD
DCFGCDGC	GCDGCDGK	HKDCFGKF	NGKFDGKD
Test strings			
Grammatical			
JXRXSPSP	LPSPSTRX	JXSPSTXS	PJTRWLJT
JXRXSTXR	RWJXSTXS	SPSTRWLP	SPJTXSTX
LPSTXRWL	STRWJXST	STRWJXST	TRXRWLJX
PSPSTRXS	WLJTXRWL	XRXSPSPJ	RWLPSPJX
STRXRXRW	XSPSTRXS	LPJXRWJT	LJTXRWJT
Ungrammatical			
TLRJTSRJ	RXTJWJRL	WLPRPXWS	JRLXWSXS
XLPWXJXR	XPWRSJSP	XJXJWXTX	PJTXSJXP
WSTLJLRP	PRLPJPWL	XSRXLRXJ	RSRSRJLP
SPXWJPXW	WPRSJWXL	PWPWPTXW	JWRPTXTR
SPSXLPLS	TWSPSRSR	XWLSTSXJ	RWJXLWXS

Appendix 3: materials in Experiment 3

Training strings					
Standard condition					
ZLBF	ZLBFJBJF	BJFJZ	BFJBF	ZLBFJBFL	ZFJBFJZ
ZLBJL	BFJBJL	ZFJBF	ZLBJLBF	ZLBJLBFL	BJLBJLBF
BJLBFJZ	ZLBJLZ	ZFJBJL	ZLBJLBFJ	BJLBJFJZ	ZLBJLBJL
ZLBFJBJL	ZFJBFL	BFJBJF	BJFJBJLZ	BJFJBFJZ	BFJBFJBF
BFJBJFL	BFJBJLZ	ZFJBFJ	BFJBJLBF	BJLBJL	BJFJBJFJ
Transfer condition					
DMXQ	DMXQHXHQ	XHQHD	XQHXQ	DMXQHXQM	DQHXQHD
DMXHM	XQHXHM	DQHXQ	DMXHMXQ	DMXHMXQM	XHMXHMXQ
XHMXQHD	DMXHMD	DQHXHM	DMXHMXQH	XHMXHQHD	DMXHMXHM
DMXQHXHM	DQHXQM	XQHXHQ	XHQHXHMD	XHQHXQHD	XQHXQHXQ
XQHXHQM	XQHXHMD	DQHXQH	XQHXHMXQ	XHMXHM	XHQHXHQH
Test strings					
Grammatical					
Low	Medium	High			
BJLZ	BFJBFL	BJLBJLZ			

ZLBJF	BJLBJF	ZLBJLBJF
ZLBFL	ZLBFJBF	BFJBFJ
ZFJBJLZ	BJLBJFL	ZLBFJBFJ
BJLBFL	ZFJBJFJ	BFJBFJZ
ZFJBJLBF	BJLBJFJ	BJFJBJF
BJFJBFL	BJFJBFJ	BJFJBJFL
BJFJBF	ZFJBJFJZ	ZFJBFJBF
Ungrammatical		
Low	Medium	High
JLBZ	BFLBFJ	BJLZBJL
BJFLZ	LBJFBJ	ZFLBJLBJ
ZBLFL	ZBFLJBF	JBFJBF
ZJBLJFZ	BJFLBJL	BFJLZBFJ
JBLFLB	BFJZJFJ	ZBFJBFJ
FBLJBJFZ	FLBJFBJ	FJBJFJB
BLFJFBJ	FJBFJBJ	JBJFLBJF
JFBFJB	ZBFJZJFJ	FBJFZBJF

Appendix 4: materials in Experiment 4

Training strings					
Standard condition					
BJFL	ZLBJLBFJ	BFJBFJBF	BFJBJFJZ	ZLBFJ	ZLBJLBJL
ZFJBFJZ	ZLBFJBJF	BFJBF	BJFJBJFJ	BJFJBJL	ZFJBJL
ZLBF	BJFJ	ZLBJFL	BFJBJFJ	ZFJZ	ZLBJFJZ
ZLBJL	ZFJBFJ	BJLBF	BJFJBJLZ	ZLBJFJBF	BFJZ
BFJBJL	ZLBFJBFL	ZLBFJZ	BJFJZ	BJLBFJ	BFJBJLZ
Transfer condition					
XHQM	DMXHMXQH	XQHXQHXQ	XQHXHQHD	DMXQH	DMXHMXHM
DQHXQHD	DMXQHXHQ	XQHXQ	XHQHXHQH	XHQHXHM	DQHXHM
DMXQ	XHQH	DMXHQM	XQHXHQH	DQHD	DMXHQHD
DMXHM	DQHXQH	XHMXQ	XHQHXHMD	DMXHQHXQ	XQHD
XQHXHM	DMXQHXQM	DMXQHD	XHQHD	XHMXQH	XQHXHMD
Test strings					
Grammatical					
Low	Medium	High			
BJLZ	BFJBFL	BJLBJLZ			
ZLBJF	BJLBJF	ZLBJLBJF			
ZLBFL	ZLBFJBF	BFJBFJ			
ZFJBJLZ	BJLBJFL	ZLBFJBFJ			
BJLBFL	ZFJBJFJ	BFJBFJZ			
ZFJBJLBF	BJLBJFJ	BJFJBJF			
BJFJBFL	BJFJBFJ	BJFJBJFL			
BJFJBF	ZFJBJFJZ	ZFJBFJBF			
Ungrammatical					
Low	Medium	High			
JLBZ	BFLBFJ	BJLZBJL			
BJFLZ	LBJFBJ	ZFLBJLBJ			

ZBLFL	ZBFLJBF	JBFJBF
ZJBLJFZ	BJFLBJL	BFJLZBFJ
JBLFLB	BFJZJFJ	ZBFJBFJ
FBLJBJFZ	FLBJFBJ	FJBJFJB
BLFJFBJ	FJBFJBJ	JBJFLBJF
JFBFJB	ZBFJZJFJ	FBJFZBJF

References

Altmann, G. T. M., Dienes, Z., & Goode, A. (1995). Modality independence of implicit learned grammatical knowledge. *Journal of Experimental Psychology. Learning, Memory and Cognition, 21*, 899–912.

Attneave, F. (1959). *Applications of information theory to psychology: a summary of basic concepts, methods, and results*. USA: Holt, Rinehart, and Winston.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

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

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

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103*, 566–591.

Chubala, C. M., & Jamieson, R. K. (2013). Recoding and representation in artificial grammar learning. *Behavior Research Methods, 45*, 470–479.

Clement, D. E., & Vernadoe, K. W. (1967). Pattern uncertainty and the discrimination of visual patterns. *Perception and Psychophysics, 2*(427), 431.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science, 16*, 41–79.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

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

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. Oxford: Wiley.

Garner, W. R. (1970). Good patterns have few alternatives: information theory's concept of redundancy helps in understanding the gestalt concept of goodness. *American Scientist, 58*, 34–42.

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

Garner, W. R., & Clement, D. E. (1963). Goodness of pattern and pattern uncertainty. *Journal of Verbal Learning and Verbal Behaviour, 2*, 446–452.

Garner, W. R., & Degerman, R. L. (1967). Transfer in free-recall learning of overlapping lists of nonsense words. *Journal of Verbal Learning and Verbal Behavior, 6*, 922–927.

Garner, W. R., & Whitman, J. R. (1965). Form and amount of internal structure as factors in free-recall learning of nonsense words. *Journal of Verbal Learning and Verbal Behavior, 4*, 257–266.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*, 1126.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*, 411428.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45*, 188196.

Jamieson, R. K., & Hauri, B. (2012). An exemplar model of performance in the artificial grammar task: holographic representation. *Canadian Journal of Experimental Psychology, 66*, 98–105.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: an analysis using information theory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 9–23.

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

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction time task: anticipating from experience. *Quarterly Journal of Experimental Psychology, 62*, 1757–1783.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: string completion and performance on individual items. *Quarterly Journal of Experimental Psychology., 63*, 1014–1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: a reply to Kinder (2010). *Quarterly Journal of Experimental Psychology, 64*, 209–216.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 137.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124–1131.

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

Kinder, A., & Assmann, A. (2000). Learning artificial grammars: no evidence for the acquisition of rules. *Memory and Cognition, 28*, 1321–1332.

Kinder, A., & Lotz, A. (2009). Connectionist models of artificial grammar learning: what type of knowledge is acquired? *Psychological Research, 73*, 659–673.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20*, 79–91.

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.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*, 299312.

Lai, J., & Poletiek, F. J. (2011). The impact of adjacent-dependences and staged-input on the learnability of center-embedded hierarchical structures. *Cognition, 118*, 265–273.

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

Lotz, A., & Kinder, A. (2006). Transfer in artificial grammar learning: the role of repetition information. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32*, 707–715.

Manza, L., & Reber, A. S. (1997). Representing artificial grammars: Transfer across stimulus forms and modalities. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 73–106). Oxford: Oxford University Press.

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

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

Mewhort, D. J. K. (1972). Scanning, chunking, and the familiarity effect in tachistoscopic recognition. *Journal of Experimental Psychology, 93*, 69–71.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 101*, 343–352.

Miller, G. A. (1958). Free recall of redundant letter strings. *Journal of Experimental Psychology, 56*, 433–491.

Miller, G. A., Bruner, J. S., & Postman, L. (1954). Familiarity of letter sequences and tachistoscopic identification. *Journal of General Psychology, 50*, 129–139.

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

Poletiek, F. H., & Lai, J. (2012). How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: a statistical account. *Philosophical transactions of the Royal Society B-Biological Sciences, 367*, 2046–2054.

Poletiek, F. H., & van Schijndel, T. J. P. (2009). Stimulus set size and grammar coverage in artificial grammar learning. *Psychonomic Bulletin and Review, 16*, 10581064.

Poletiek, F. H., & Wolters, G. (2009). What is learned about fragments in artificial grammar learning: a transitional probabilities approach. *Quarterly Journal of Experimental Psychology, 62*, 868–876.

Pothos, E. M. (2010). An entropy model for artificial grammar learning. *Frontiers in Psychology, 1*, 16.

Pothos, E. M., & Bailey, T. M. (2000). The importance of similarity in artificial grammar learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 26*, 847–862.

Pothos, E. M., & Ward, R. (2000). Symmetry, repetition, and figural goodness: an investigation of the Weight of Evidence Theory. *Cognition, 75*, B65–B78.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and 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.

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

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

Royer, F. L., & Garner, W. R. (1966). Response uncertainty and perceptual difficulty of auditory temporal patterns. *Perception and Psychophysics, 1*, 41–47.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Schnore, M. M., & Partington, J. T. (1967). Immediate memory for visual patterns: symmetry and amount of information. *Psychonomic Science, 8*(421), 422.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Simon, H. A. (1957). *Models of man*. New York: Wiley.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition, 31*, 1060–1071.

van der Helm, P. A., & Leeuwenberg, E. L. J. (1996). Goodness of visual regularities: a non-transformational approach. *Psychological Review, 103*, 429–456.

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

Vokey, J. R., & Higham, P. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology, 59*, 54–61.

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