When Global Structure “Explains Away” Local Grammar: A Bayesian Account of Rule-Induction in Tone Sequences

Colin Dawson and LouAnn Gerken

Department of Psychology

The University of Arizona

Submitted for the *Cognition* Special Issue:

“Probabilistic Models of Cognitive Development”

Send correspondence to:

Colin Dawson

`cdawson@email.arizona.edu`

Approximate Word Count: 5971

Abstract

While many constraints on learning must be relatively experience-independent, past experience provides a rich source of guidance for subsequent learning. Discovering structure in some domain can inform a learner's future hypotheses about that domain. If a general property accounts for particular sub-patterns, a rational learner should not stipulate separate explanations for each detail without additional evidence, as the general structure has "explained away" the original evidence. In a grammar-learning experiment using tone sequences, manipulating learners' prior exposure to a tone environment affects their sensitivity to the grammar-defining feature, in this case consecutive repeated tones. Grammar-learning performance is worse if context melodies are "smooth" — when small intervals occur more than large ones — as Smoothness is a general property accounting for a high rate of repetition. We present an idealized Bayesian model as a "best case" benchmark for learning repetition grammars. When context melodies are Smooth, the model places greater weight on the small-interval constraint, and does not learn the repetition rule as well as when context melodies are not Smooth, paralleling the human learners. These findings support an account of abstract grammar-induction in which learners rationally assess the statistical evidence for underlying structure based on a generative model of the environment.

Keywords: statistical learning, rule-learning, Bayesian modeling, language acquisition, music cognition

1 Introduction

In traditional learning theories, the relationship between knowledge and learning is static: Learning builds knowledge subject to *a priori*, biological constraints. A seldom-explored area is the dynamic interplay between learning and knowledge: (how) can previous learning change subsequent learning? Ignoring this feedback can lead to incorrect attribution of observed constraints to biologically provided "knowledge" instead of to previous learning.

Untangling the contributions of experience-independent biology and prior learning has been particularly important in studying infant cognition: if an infant learns one pattern and not another in the absence of a priori differences in difficulty, it is tempting to attribute the discrepancy to biology. This conclusion would be premature without further examination, however.

Previous research suggests that infants reorganize their domain knowledge in the first year, and even in the laboratory. Infants reorganize their phonetic categories (Werker & Tees, 1984; Bosch & Sebastián-Gallés, 2003; Maye, Werker, & Gerken, 2002) and exhibit shifts in what features they will consider for linguistic stress rules (Gerken & Bollt, 2008). In music, relative pitch takes over from absolute pitch as the dominant cue for organizing melodies (Saffran & Griepentrog, 2001; Saffran, 2003), and infants tonal and rhythmic categories change with cultural context (Hannon & Trehub, 2005; Lynch & Eilers, 1992).

Marcus, Vijayan, Rao, and Vishton (1999) and Marcus, Fernandes, and Johnson (2007) found that 7-month-old infants learn an AAB or ABB pattern over sequences of syllables, but infants the same age fail when the elements are non-linguistic events such as musical tones or animal noises. It was suggested that the child's innate endowment might attribute abstract, relational properties to speech, but not other auditory stimuli. While this is possible, infants can learn AAB-style structure with pictures of dogs (Saffran, Pollack, Seibel, & Shkolnik, 2007) and simple shapes (Johnson et al., 2009), and rats can learn such generalizations from tones (Murphy, Mondragon, & Murphy, 2008), casting doubt on the notion that language is intrinsically privileged for rule-learning.

Dawson and Gerken (2009) found that while 7-month-olds fail at learning AAB and ABA patterns with tones or chords, 4-month-olds succeed. They suggested that 7-month-olds' failure may stem from their having learned certain general properties about music: If they have learned, for example, that the intervals from one pitch to the next tend to be small in magnitude (Ortmann, 1926; Dowling, 1967; Dawson, 2007; Temperley, 2008), a high repetition rate would become much less surprising, and hence less informative about the abstract AAB-style structure. This change in information value is an example of "explaining away", a phenomenon central to cognitive models in a variety of areas including visual inference (Kersten, Mamassian, & Yuille, 2004), linguistic processing (Ciaramita & Johnson, 2000), and causal reasoning (Xu & Garcia, 2008; Gergely & Csibra, 2003).

The basic idea is as follows: When an observed pattern could arise from multiple causes, the causes "compete" over the evidence in the data, even when they do not conflict with each other *a priori*. A classic example comes from Pearl (1988): Both rain and a sprinkler can cause my lawn to be wet. By itself, observing the wet ground increases the plausibility of both causes. If I discover that the ground is also wet next door, this provides additional evidence for the rain hypothesis, but is not directly relevant to the sprinkler hypothesis: my sprinkler has no effect on the neighbor's lawn, and so, by itself, the state of the latter is irrelevant to inferences about the former. However, by increasing the likelihood of rain, the neighbor's wet lawn helps to *explain away* the evidence for the sprinkler. Although the sprinkler may well have run, my wet lawn constitutes weaker evidence than before.

In music, repetition is an ambiguous event. On the one hand, it constitutes a "sameness" relation between two tones. It is also an interval of magnitude zero. If one assumes that any tone is equally likely at any point (the tone distribution is uniform), hearing every melody begin with two repeated notes would be quite surprising, and evidence for a "sameness" interpretation would be strong. If, however, one knows that tones nearby in time also tend to be nearby in pitch (melodies are usually "Smooth), repetition should be more common (*qua* interval of distance zero), and it should take more evidence to conclude that repetition is special.

2 Human Experiment

We examine whether the presence of a "Smoothness Constraint" on melodies in the broader environment will lead adult learners to discount evidence for a repetition rule by reducing its surprise value. Since small intervals are not surprising in a Smooth environment, a learner modeling this tendency should not treat frequent repetitions as evidence for additional structure: Learners previously exposed to the Smooth environment should less readily infer the existence of a repetition rule than those familiarized with non-Smooth melodies. On this account, it is not merely the high rate of repetition in the Smooth environment that leads to discounting of the evidence for a repetition rule; rather, it is that a high rate of repetition is an incidental consequence of the Smoothness property. If the environment contains a large number of repetitions in the *absence* of a Smoothness Constraint, no discounting should occur.

The central prediction is that a repetition rule will be more difficult to learn in a Smooth melodic environment. To test this, participants are familiarized with one of three melodic contexts. In the Uniform condition, each of a set of tones is equally likely at any point. In the Smooth condition, small intervals are more common than large intervals. In the Repetition condition, repetition alone is more frequent than other intervals. The latter two groups are subdivided into high repetition (Low Variance) and low repetition (High Variance) conditions, with the absolute repetition rate matched between each Smooth group and its Repetition counterpart.

Following exposure to these contexts, participants perform a grammar-induction task where half of participants learn a repetition-initial (AABCD) grammar, and half learn a repetition-final (DCBAA) grammar. If learners model the overall interval distribution, the Smooth context should portray repetition more as a zero-magnitude interval expected to be frequent, and less as a specific grammatical feature. Learners in this context should exhibit decreased sensitivity to positional repetition, as well as decreased grammar-learning performance. No explaining-away should occur in the Repetition groups, since no broader property accounts for the high rate of repetition.

2.1 Methods

2.1.1 Participants

One hundred and thirty-eight University of Arizona undergraduates participated in the study for course credit.

2.1.2 Materials and Procedures

The experiment consists of a “Context” phase and a Grammar-Learning phase. The latter contains four blocks, each with training and test components. “Sentences” consist of five tones generated using the FM Synthesizer in the MIDI Toolbox for MATLAB (Eerola & Toiviainen, 2004). The first four are 250 msec, followed by 50 msec gaps. The last tone is 500 msec. Musically speaking, melodies contain four eighth notes followed by a quarter note, played at 200 beats per minute.

2.1.3 Procedures: Context Phase

The Context phase contains two randomized blocks of 100 sentences. Ten are “probe” sentences, after which either the same sentence is repeated or one of the other ten probe sentences is played. On probe trials, participants have 3 seconds to register “same” or “different” pairs via a button-press. Failure to respond is considered incorrect. Each block lasts about five minutes. Discrimination scores did not significantly differ across context conditions ($F(4, 133) = 1.60$, $p = 0.18$, $MSE = 6.95$). Data from participants who did not perform above chance on this discrimination task (15 or more out of 20 correct) was discarded, as these participants ($n = 18$) presumably either could not distinguish differences among melodies, or were not at-tempting to succeed. The proportion of discards did not significantly differ across groups ($p = 0.19$, Fisher's exact test).

During context exposure, all participants see a group of eight "aliens" (Folstein, Van Petten, & Rose, 2007), half "star-chested" and half "brick-chested".

2.1.4 Materials: Context Phase

Participants are assigned to the Uniform ($n = 24$), Smooth ($n = 48$) or Repetition ($n = 48$) contexts. The Smooth and Repetition conditions are divided into High Variance (HV) and Low Variance (LV) sub-conditions. All context melodies comprise a "vocabulary" of six tones (MIDI values: 57, 58, 61, 64, 67 and 68). The restricted vocabulary, with uneven steps between tones, parallels natural musical scales, which partially divorce "diatonic" and acoustic interval measures.

In the Uniform condition, each tone is equiprobable and independent of the last. The probability of repetition at any given point is 1/6. The empirical distribution is shown in Fig. 1a.

In the Smooth condition, the first tone is chosen uniformly from the six possible. For subsequent tones, a sample is generated from a truncated normal distribution on the interval $[0.5, 6.5]$. The mean is an integer corresponding to the previous tone (the lowest tone is 1; the highest tone 6). The variance is 4.00 in the HV condition and 1.44 in the LV condition. The tone generated is the nearest integer to the sample. The resulting distribution reflects the bias toward small intervals in Western folk music. The empirical interval distributions are shown in Fig. 1b-c.

The Repetition conditions control for the absolute rate of repetition, removing

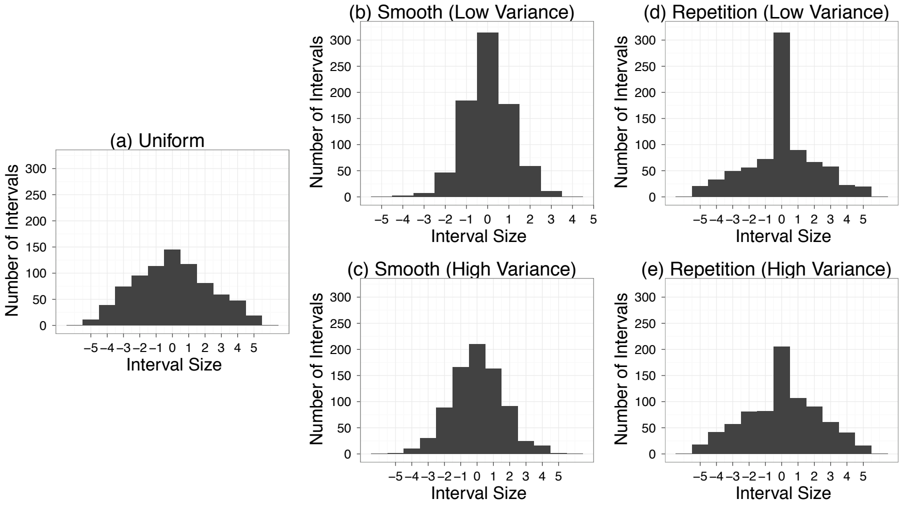

Figure 1: Interval Distributions by Context, collapsed over preceding pitch. Zero is a repetition, positive values are ascending, 1 represents a step to the next note, etc. More small intervals occur across all conditions due to the bounded pitch range.

the overall "Smoothness" constraint. Here, the LV and HV conditions (Fig. 1d-e) are matched to their Smooth counterparts in number of repetitions, but now the remaining notes are equiprobable. Here, the high repetition rate cannot be explained by a general bias for small intervals; instead, a learner modeling the tone distribution must encode repetitions separately.

Empirical repetition rates by position and Context are shown in Fig. 2.

2.1.5 Procedures: Grammar-Learning Phase

After the context phase, participants move on to the grammar-learning phase. They are asked to detect alien "spies" by identifying ungrammatical sentences.

In training blocks, participants hear thirty grammatical sentences in random

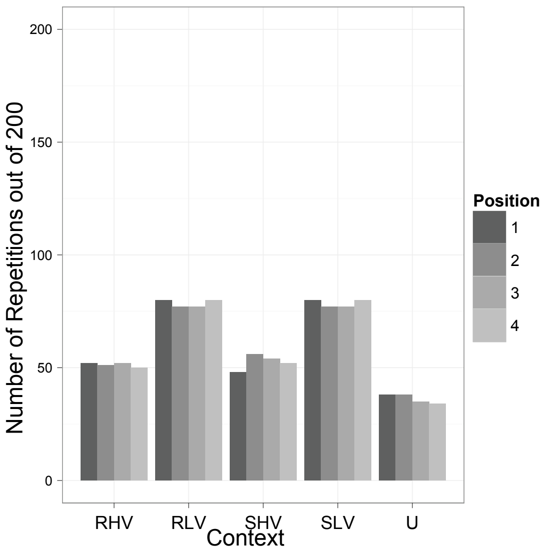

Figure 2: Repetition Rates by Context and Position

order while viewing an image of four star-chested aliens.

After each training block, participants hear twenty-four test sentences, half grammatical. After each sentence, participants make a grammaticality judgment by a mouse-click on a line whose left pole represents “definitely grammatical”, whose right pole represents “definitely ungrammatical”, and which allows every gradient response in between. There is no time limit. The computer records a binary response, corresponding to left or right of center, and a continuous “discrimination score” calculated by subtracting from 100 the percentage of the line lying between the response and the correct pole. Participants experience four training-test cycles on the same grammar.

2.1.6 Materials: Grammar-Learning Phase

The “Qixian” and “spy” sentences are again five tones long. Each participant is trained using one of two five-tone vocabularies. The first contains MIDI values 57, 60, 63, 66 and 67; the second contains MIDI values 58, 59, 62, 65 and 68. Each set shares two tones with the context vocabulary.

For half of participants, grammatical sentences follow an AABCD pattern (a repetition at the beginning and nowhere else), while the ungrammatical sentences have a DCBAA pattern. For the other half, the labels are reversed.

Of 120 possible sentences in each grammar, 60 are used as training items, and 24 as test items. The chosen items were balanced for pitch contour, with falling and rising intervals equally frequent at each position. Each sentence in one grammar has a sequential mirror-image in the other.

Thirty training sentences are used in the first two learning blocks; the other thirty in the last two blocks. On odd-numbered blocks, participants are tested with items from the training vocabulary; on even blocks they hear items from the opposite vocabulary. Both vocabularies were used to test whether the context manipulation has an effect on the level of abstraction at which participants learn the grammar.

2.2 Results

The central question is whether prior exposure to the Smooth distribution will impair detection of the repetition rule. If so, this will suggest that learners model the full interval distribution, which (partially) explains away the training repetitions. The key comparison is therefore between the Smooth groups and the non-Smooth groups.

A secondary question is what effect the overall rate of repetitions, independent of the Smoothness constraint, has on learning the rule. If broader structure is irrelevant and learners are influenced only by the amount of repetition they are exposed to, there are two possibilities: if background repetition has a desensitizing effect, then the Uniform group should outperform the High Variance Repetition group, which in turn should outperform the Low Variance Repetition group. Similarly, the High Variance Smooth group should outperform the Low Variance Smooth group. If background repetitions highlight identity relationships, the reverse rankings should obtain. If qualitative structure is primary, however, any effect of Repetition Rate and/or Variance should be subordinate to the shape of the interval distribution.

In order to examine the effects of Context condition on participants' ability to distinguish grammatical from ungrammatical test items, we analyzed both binary and confidence-weighted responses using a general linear mixed model. The two yielded qualitatively identical results, and so for concision and ease of interpretation, we report only the binary results.

Fixed effects of Context (five levels: RepetitionHV, RepetitionLV, SmoothHV, SmoothLV and Uniform), Block (four levels: 1 through 4), and a covariate based on melodic discrimination prior to grammar-learning, were included. Test grammar (AABCD vs. DCBAA) was examined initially, but did not have an effect, and hence was dropped from further analyses. Three orthogonal planned comparisons among the five Context conditions were of interest. The comparison of greatest interest (the "Smoothness contrast") is between the two Smooth groups, on the one hand, and the other three groups, on the other. Its coefficient, β_S, represents the

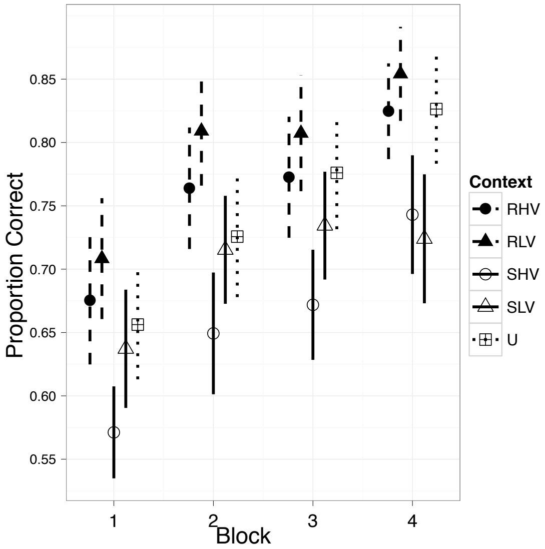

Figure 3: Mean Performance by Context and Block. Error bars are ±1 SEM

difference in mean percent of correct responses between the two Smooth conditions and the other three groups. The "Repetition Rate" contrast examines the effect of repetition rate among the non-Smooth contexts, assigning numerical scores to these three conditions in proportion to the rate of repetition in each. Its coefficient, β_{RR}, represents the increase in percentage of correct responses for each 5% increase in the rate of repetitions in the context. The "Variance contrast" compares the two Smooth groups. A positive value of its coefficient, β_V, denotes an advantage in percent of correct responses for the High Variance group.

Orthonormal polynomial contrasts were used for the Block factor. A positive linear coefficient (β_L) captures improvement over Blocks, while a negative quadratic coefficient (β_Q) captures leveling off of performance. The cubic coefficient (β_C) captures additional discrepancies between the same-vocabulary and transfer-vocabulary blocks, with positive values denoting a transfer advantage.

The "Baseline Discrimination" covariate was based on the number of correct responses in the 20 probe trials during the Context phase. Scores were standardized within each Context condition. A positive coefficient for β_{BD} would validate the intuition that some general tone processing ability is needed for both tasks.

Finally, four random effects, corresponding to between-subjects variability in overall performance as well as the polynomial Block coefficients, were estimated along with their correlations.

The first model included all interaction terms among Context, Block and Baseline Discrimination. Maximum-likelihood parameter estimates were computed using the `lme` function from the `nlme` package in `R` (Pinheiro, Bates, DebRoy, Sarkar, & R Core team, 2009). A second model contained a reduced set of random effects: only a random intercept and slope. The higher-order random effects provided a significantly better fit as determined by a Likelihood Ratio Test ($\chi^2(7) = 32.36$, $p < 0.0001$), and hence were retained. Next, the full model was compared to one without interactions involving the covariate. This time, the simpler model did not exhibit a significantly worse fit ($\chi^2(19) = 20.81$, $p = 0.34$). Moreover, removing the Context × Block interactions did not significantly worsen the fit ($\chi^2(12) = 7.00$, $p = 0.86$), and so the pure main effects model was retained.

Critically, the Smoothness contrast was significant, with fewer correct responses in the Smooth groups ($\beta_S = -8.7\%$, $SE = 3.2\%$, $t(114) = 2.75$, $p < 0.01$). Repetition Rate was not significant ($\beta_{RR} = 1.1\%$, $SE = 1.2\%$, $t(114) = 0.96$, $p = 0.34$), nor was Variance ($\beta_V = -3.3\%$, $SE = 4.9\%$, $t(114) = 0.66$, $p = 0.51$). The linear Block effect was significant ($\beta_L = 10.2$, $SE = 1.5$, $t(357) = 6.86$, $p < 0.001$), reflecting improvement over blocks. The quadratic effect was marginally significant ($\beta_Q = -2.0$, $SE = 1.19$, $t(357) = 1.72$, $p = 0.09$), reflecting a leveling off in performance. The cubic term was significant as well ($\beta_C = 1.9$, $SE = 0.9$, $t(357) = 2.08$, $p < 0.05$), with better performance in transfer blocks. A significant positive relationship obtained between Baseline Discrimination and grammar-learning performance ($\beta_{BD} = 5.9$, $SE = 1.6$, $t(114) = 3.80$, $p < 0.001$). Mean percent correct by Context and Block is depicted in Fig. 3.

2.3 Discussion

The primary prediction in this experiment was that participants in the Smooth environment would discount the evidence for the repetition pattern, exhibiting decreased grammar-learning performance, since a high rate of repetition is produced by Smoothness alone. This prediction was supported. This effect cannot be due to desensitization to repetition, as greater repetition did not significantly impact learning with or without the Smoothness constraint, and in fact, the numerical difference favored greater repetition. Though a significant positive effect of repetition rate might appear in a larger or less variable sample, what is clear is that the strongest predictor of performance is the qualitative shape of the context distribution.

These findings suggest that learners in this experiment are modeling the alien environment, and forming hypotheses about the input-generation process. They seem to use this model to guide subsequent learning in the environment, by assessing the evidentiary value of a cue to new potential underlying structure. Here, in the Smooth environment, repetitions do not appear to be an essential component of the environment at all, whereas without Smoothness it is necessary to represent them in order to understand the distribution of intervals. Greater improvement during transfer-vocabulary blocks suggests that learners may be entertaining multiple grammar hypotheses, with the vocabulary change serving as a hint that the relevant rule is vocabulary-independent.

We must acknowledge two alternative interpretations of the Smooth disadvantage. The first is that learners are collapsing across rising and falling intervals and encoding only absolute interval magnitude. In that case, listeners could be focusing on the most common absolute interval, which is a repetition in the Repetition conditions, and a step of one in the Smooth conditions. Though we cannot firmly rule out this possibility, it would seem to predict that the Uniform group should also suffer, since there, single steps are also the most frequent (and in fact, the ratio between single step and repetition frequencies is greater than in either Smooth group). Instead, the Uniform group was closer to the Repetition groups.

A second alternative is that, since melodies in participants' natural environment are Smooth, participants in the Smooth groups more readily engaged their prior preconceptions about the structure contained in music. While this would not contradict our account, attributing these prior preconceptions to Smoothness in the environment would be begging the question (though this account still requires that people encode the Smoothness in melodies at some level). Further research, perhaps with less natural properties, is needed to disentangle prior biases from in-laboratory learning.

3 Bayesian Model

In order to quantify "best-case" behavior predicted for a learner with access to the true causal structure of the environment, an idealized generative statistical learning model was constructed. It "observed" an abstraction of the melodies that the human participants heard. The use of such a model is not a claim about psychological mechanisms involved in grammar-learning; it is merely an attempt to make explicit the inferences that result from the rational use of melodic data, given a particular set of prior beliefs about how melodies are generated.

3.1 Model Definition

A generative probabilistic model has two components: a probability distribution (likelihood) over possible data points given a set of unknown parameters, and a prior belief distribution over possible values of those parameters. For simplicity, the likelihood component of our model entertains only the two qualitative processes used to generate melodies in the experiment. In the first process, repetition has a particular probability p, and the other $V-1$ tones are equiprobable (where a uniform distribution is the case with $p = {}^1\!/\!_V$). In the second, pitches are normally distributed around the value of the preceding tone. The model contains four free parameters, p_1 to p_4, determining the probability of repetition at each of the four sequential positions in the absence of Smoothness. A fifth parameter h governs the variance of the normal

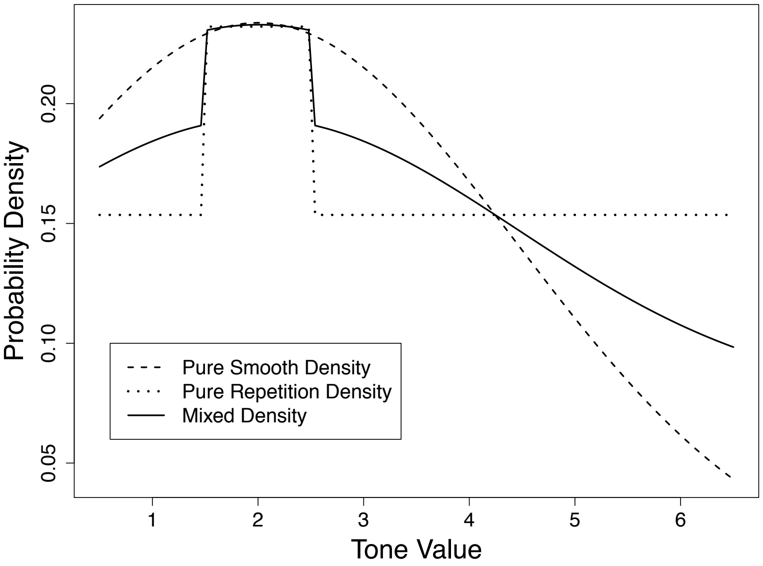

Figure 4: Probability density for a tone at position j after tone 2 occurs at position $j-1$, with $p_{j-1} = 0.23$, variance 6, and $\pi_s = 0.5$.

distribution produced by the Smoothness constraint. A final parameter π_s governs the extent to which the distribution is determined by the Smoothness constraint. Finally, melodies are assumed to be either grammatical or ungrammatical. The two types are allowed to have different repetition probabilities at each position, but are assumed to be subject to the same smoothness constraint, with the same mixing weight.

We use *conjugate priors* for all parameters, which have an “equivalent data” interpretation (Box & Tiao, 1973), i.e., their parameters represent a summary of previous imaginary data. The prior expected interval variance was set to 6, according to the interval distribution in children’s folk music (Dawson, 2007), and the prior

Parameter	Description	Learned or set?
$\{p_{1,j}\}_{j=1}^{4}$ and $\{p_{0,j}\}_{j=1}^{4}$	"Pure" repetition probabilities in the absence of smoothness. Separate values for grammatical and ungrammatical items, and for each sequential position.	Learned
h	Precision of smoothness constraint (inverse of the variance). Higher values reflect a greater tendency for small intervals.	Learned
π_s	Mixing weight of smooth process. Zero represents no smoothness constraint; one represents a distribution completely determined by smoothness.	Learned
N	Prior "Equivalent Sample Size". Higher values reduce the influence of the data relative to the prior.	Set
S	Prior estimate for π_s	Set

Table 1: Descriptions of the model parameters

expected repetition probabilities were set to capture a uniform interval distribution. Two free parameters were varied across simulation runs. The first, N, determined the strength of the prior (in "equivalent melodies")[1]. The second, S, determined the expected weight of the Smoothness constraint.

For details on the prior and likelihood functions, see the Appendix. Example likelihood functions are shown in Fig. 4, and a summary of the model parameters is

[1] Human learners presumably do not attach the same weight to generic prior experience that they do to specific training items in context. Hence, N should be set to a considerably smaller value than the actual number of melodies encountered in life.

given in Table 1.

3.2 Simulations and Evaluation

The model "heard" 200 Context melodies, abstractly identical to those encountered by the human participants, as well as grammatical training melodies. The model's grammatical discrimination performance was evaluated after 30, 60, 90 and 120 training melodies, corresponding to the four test blocks in the experiment.

The bulk of the simulation process consisted of estimating a joint posterior distribution over the ten data-generation parameters, $\{p_{1,j}\}_{j=1}^{4}$, $\{p_{0,j}\}_{j=1}^{4}$, h and π_s. This estimation was accomplished using Markov Chain Monte Carlo sampling (Gilks, Richardson, & Spiegelhalter, 1996) (see the Appendix for details). Simulations were run for each of the five Context conditions, with each of the four levels of training, and at three values each of the prior parameters N and S. N was set to 1, 20 or 200, and S was set to 0.1, 0.5 or 0.9, where higher values reflected greater prior weight for the smoothness constraint.

Of central interest was not what parameter values the model would infer, but how accurately it could infer the grammaticality of novel melodies. Since the sampling procedure produces a distribution of thousands of values, we can assign grammaticality probabilities to each test sentence for each set of parameter values in the sample. For each of the 24 test melodies, at each set of parameter values, the model can make a probabilistic binary decision. The mean proportions of correct responses for each simulation run are plotted in Fig. 5.

In this first set of simulations the model made no assumption about the grammaticality of the Context melodies, and hence they informed both $\{p_{1,j}\}_{j=1}^4$ and $\{p_{0,j}\}_{j=1}^4$ equally (as well as the general parameters h and π_s). Training melodies, which were known to be grammatical, only influenced parameters relevant for grammatical melodies (i.e., all but $\{p_{0,j}\}_{j=1}^4$). To assess whether this "agnosticism" about the Context melodies was critical, we ran four additional simulations in which we varied the probability of grammaticality assigned to the Context melodies, representing an additional parameter, G (previously fixed to 0.5). When G is 0, grammatical parameters are set using only the prior and the training melodies; when G is 1, Context and training are both used for the grammatical model, while the parameters for ungrammatical melodies come entirely from the prior. In these runs, the N and S parameters were set to their intermediate values of 20 and 0.5, respectively. Results are displayed in Fig. 6.

3.3 Discussion

For all parameters, the model performs more poorly in both Smooth conditions than in the other conditions, like the human participants. When too high a mixture weight is assigned to the Smoothness constraint ($S = 0.9$), the data conveys little information about repetition probabilities, since most of the information in the interval distribution is assumed to reflect Smoothness. This makes it difficult to learn the rule in any condition, except when the prior is very weak ($N = 1$). The smaller the value of S, the better the model performs across conditions, as a greater proportion of the interval evidence is taken to reflect particular repetition patterns. Performance degrades with very large values of N, as an overly strong prior makes

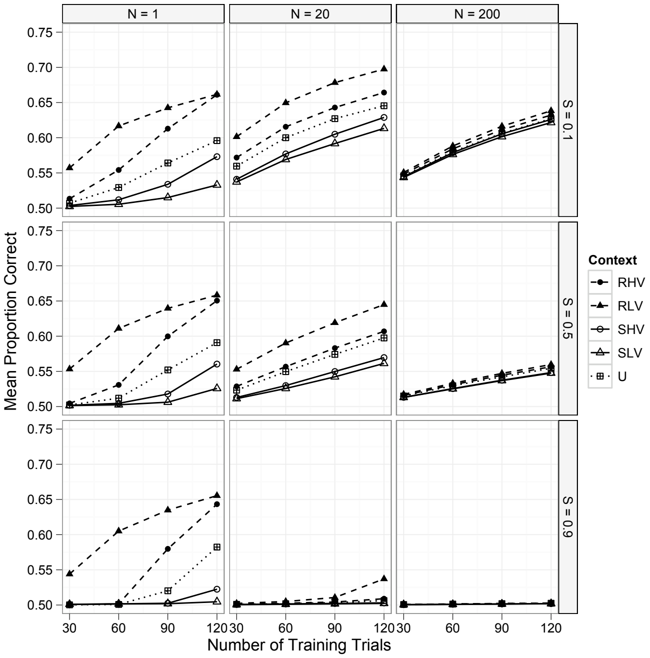

Figure 5: Performance of the Model by Block, Condition and Parameter Settings

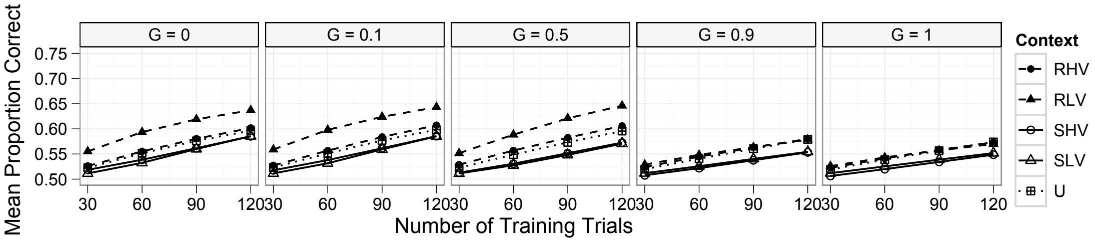

Figure 6: Performance Under Varying Assumptions About Grammaticality of Context Items. G is the prior probability of assigning a Context melody to be grammatical.

the model unresponsive to the data, sticking too strongly to its "preconceptions".

The model exhibits a positive effect of repetition rate in the non-Smooth groups, with fairly large differences between the two Repetition groups in most cases. The ranks of these three conditions was the same for the humans, though the differences were not statistically significant. Human participants were not told that a grammar existed during the Context phase, and hence might be expected to learn only general properties at that stage, in which case quantitative repetition rates may have been less influential for them than for the model which encoded all data in fine detail. When the model treated the Context melodies as completely grammatical, additional repetitions at both the initial and final position made both the grammatical and ungrammatical test melodies more likely under the model's grammar, and hence neither helps nor hurts discrimination. It may be that a subset of humans failed to make the distinction between Context and training, behaving as though everything they heard was grammatical.

For the two Smooth conditions, the model predicted slightly poorer performance in the Low Variance case. The humans exhibited the opposite pattern until the last block, though again the difference was not significant. Even if human learners are sensitive to the variance of the Smooth distribution, the very small effect exhibited by the model suggests that a large sample may be needed to detect this difference.

The present model is useful for its quantitative realization of the idea that a general environmental feature can explain away evidence for a particular underlying structure in the input. The model captures a difference in human performance between two context conditions (Repetition and Smooth) that would be difficult to explain without supposing that human learners generatively model their input. If they were merely trying to learn dependencies between features and categories, the frequency of repetition alone, and not its relationship to the broader environmental structure, should be dominant in driving performance.

Although the present model captures the qualitative performance of the human participants, it should not be taken literally as an account of psychological mechanisms. For one, the mathematical forms of the likelihood and prior distributions were chosen largely for computational convenience. Aside from the qualitative shapes of these distributions (i.e. unimodal and smooth), their particular choices should not be taken as psychological hypotheses.

More importantly, the hypothesis space with which the model is endowed is extremely narrow, and is suited particularly to this task. While differences between conditions are still meaningful, this model is freed from some difficulties facing humans; at the same time, it could not learn a range of other grammatical regularities that humans could likely learn. For example, the model only has access to abstract pitch information, and does not have to learn how to measure intervals, nor does it need to learn how to transfer its knowledge across vocabularies. In ongoing research we are examining the role of the tone vocabulary. Will different results obtain if participants are not required to switch between two tone sets? What effect does the use of uneven step sizes have on inference?

On the flip side, the model cannot represent pitch contour, special roles for particular pitches, or dependencies between non-consecutive tones, all of which are potentially meaningful features of a melodic environment; nor does it attribute special salience to repetition at edges (Endress, Scholl, & Mehler, 2005). We are currently investigating whether Smoothness might in fact have a facilitative effect if the grammar being learned is based on contour, rather than repetition, for example.

4 General Discussion

We have used both behavioral and computational methods to investigate the contribution of rational, generative “explaining away” to induction of an abstract rule for tone sequences. In sequences of musical tones, repetition has a dual nature, first as an identity relation between two consecutive events, and second as an interval of magnitude zero between two tones on a continuum. When a repetition occurs, it is ambiguous which of these two descriptions should be attached to it. Our central finding was that adult humans appear to take into account a global “Smoothness” constraint on melodies, which have a statistical tendency to move in small intervals, to set a baseline expectation for the rate of repetitions. This reduces the informational value of a repeated tone as a cue to an abstract rule. A Bayesian model that entertains a smoothness constraint and also allows for special “sameness” relations confirms the intuition that this pattern of results is to be expected under a rational hypothesis-testing account of rule-induction.

These findings are of substantial relevance to the rule-learning literature following Marcus et al. (1999), and are particularly supportive of our earlier conjecture (Dawson & Gerken, 2009) that 7.5-month-olds may have “learned to fail” at learning AAB rules by acquiring knowledge about tonality and the smoothness of natural melodies. A similar account applies to ABA rules, since Smoothness makes non-adjacent repetitions more frequent as well (and, indeed, by Dawson (2007), a normal distribution fits intervals with lag 2 as well). It will be revealing to see whether the present model behaves like infants in earlier studies after familiarization with natural musical contexts. We are also adapting the present experiment to infants to determine whether the explaining away process observed in adults occurs in the laboratory with infants as well. If these extensions bear out as predicted, we will have converging evidence that "metalearning" plays an important role in the formation of apparently domain-specific biases and constraints.

In order to explain away, learners must be explaining. The present findings add to a growing literature (Gopnik, 1998; Schulz & Bonawitz, 2007; Xu & Garcia, 2008; Gerken, 2010) suggesting that learning is like science: in addition to making specific predictions, an important role of cognition is to build explanatory models of the environment, and to construct and test hypotheses about why the world works as it does.

A Appendix: Modeling Details

A.1 Model Definition

Each melody i is represented as a sequence of five integers, $t_{i,1}$ through $t_{i,5}$, ranging from 1 to V in ascending order of pitch, where V is the number of tones in the vocabulary. For mathematical convenience, $t_{i,j}$ is determined by a probability distribution $f(t^*_{i,j})$ over the interval $[0, V]$, and a deterministic function rounding $t^*_{i,j}$ up to the nearest integer. When $j = 1$, $f(t^*_{i,j})$ is always uniform. For $j > 1$, $f(t^*_{i,j}|t_{i,j-1})$ is given by:

$$f(t^*_{i,j}|t_{i,j-1}) = \pi_s q_s(t^*_{i,j}|t_{i,j-1}) + (1 - \pi_s) q_r(t^*_{i,j}|t_{i,j-1}) \tag{1}$$

The functions $q_r(t^*_{i,j}|t_{i,j-1})$ and $q_s(t^*_{i,j}|t_{i,j-1})$, are the "repetition" and "smooth" densities over tones, given by:

$$\begin{aligned} q_r(t^*_{i,j}|t_{i,j-1}) &\propto \mathbf{1}_{[t_{i,j-1}-1,t_{i,j-1}]}(t^*_{i,j}) p_{g,j-1} + (1 - \mathbf{1}_{[t_{i,j-1}-1,t_{i,j-1}]}(t^*_{i,j})) \frac{1 - p_{g,j-1}}{V-1} && (2) \\ q_s(t^*_{i,j}|t_{i,j-1}) &\propto \mathbf{1}_{[0,V]}(t^*_{i,j}) h^{1/2} \exp(-\frac{1}{2} h (t^*_{i,j} - t_{i,j-1})^2) && (3) \end{aligned}$$

where $\mathbf{1}_{[a,b]}(x)$ is the indicator function which is 1 when $x \in [a, b]$ and 0 otherwise.

The prior distributions of π_s, h and $\{p_{g,j}\}$ are given by:

$$\begin{aligned} f_{\pi_s}(\pi_s) &\propto {\pi_s}^{4NS} (1 - \pi_s)^{4N(1-S)} && (4) \\ f_p(p_{g,j}) &\propto {p_{g,j}}^{\frac{N}{V}} (1 - p_{g,j})^{\frac{(V-1)N}{V}} && (5) \\ f_h(h) &\propto h^{\frac{4N}{2}-1} \exp(-\frac{6 \times 4N}{2} h) && (6) \end{aligned}$$

Here, N represents the prior "equivalent sample size" (in melodies), and S represents the prior strength of the smoothness constraint. The π_s and h parameters apply to 4 intervals per melody, for a total of $4N$, whereas each $p_{g,j}$ applies to only one interval per melody, for a total of N. The 6 in the $f_h(h)$ expression represents the prior variance in the smooth distribution, fixed by the empirical distribution in Dawson (2007).

A.2 Sampling Procedure

In Gibbs sampling, the full parameter set is partitioned, and at each step a sample is taken from the conditional distribution of one block given all the others. A sample is taken for each block at each iteration, conditioning on the most recent values for the other blocks. We used separate blocks for (1) π_s, (2) h and (3) $\{p_{g,j}\}$. In a final block, unlabeled melodies were assigned a grammaticality label at each step and individual intervals were hard-assigned to either the repetition or smooth distributions, both according to their conditional posterior probabilities. The hard-assignment is performed to simplify sampling; each interval should be thought of as coming from a weighted mixture of the two distributions.

Due to the truncation of the smooth distribution, the conditional posterior for h is nonstandard (it would be Gamma otherwise). Therefore, a Metropolis-Hastings step (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970) was incorporated, using a Gamma as the proposal distribution.

The sampler was run for 100,000 iterations at each parameter combination. The first 50,000 iterations were discarded as "burn-in". The remaining 50,000 samples were used to assign grammaticality probabilities to the test melodies.

References

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, *46*, 217-243.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Ciaramita, M., & Johnson, M. (2000). Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th international conference on computational linguistics* (p. 187-193). Saarbrucken, Germany.

Dawson, C. (2007). *Infants learn to attend to different relations when forming generalizations in different domains.* Unpublished master's thesis, University of Arizona.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378-382.

Dowling, W. J. (1967). *Rhythmic fission and the perceptual organization of tone sequences.* doctoral dissertation, Harvard University, Cambridge, MA.

Eerola, T., & Toiviainen, P. (2004). *MIDI toolbox: MATLAB tools for music research.* Unpublished master's thesis, University of Jyväskylä, Kopijvä, Jyväskylä, Finland.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*(3), 406-419.

Folstein, J. R., Van Petten, C., & Rose, S. A. (2007). Novelty and conflict in the categorization of complex stimuli. *Psychophysiology*, *45*, 467-479.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*, 287-292.

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362-266.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms. *Language Learning and Development*, *4*(3), 228-248.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Suffolk: Chapman and Hall.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*(1), 101-118.

Hannon, E. E., & Trehub, S. E. (2005). Tuning into musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences*, *102*(35), 12639-12643.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N. Z., Marcus, G. F., Rabagliatti, H., et al. (2009). Abstract rule learning for visual sequences in 8- and 11-month-olds. *Infancy*, *14*(1), 2-18.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review Psychology*, *55*, 271-304.

Lynch, M. P., & Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Perception and Psychophysics*, *52*(6), 599-608.

Marcus, G. F., Fernandes, K., & Johnson, S. (2007). Infant rule-learning facilitated by speech. *Psychological Science*, *18*, 387-391.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77-80.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 101-111.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule learning by rats. *Science*, *319*, 1849-1851.

Ortmann, O. (1926). On the melodic relativity of tones. *Psychological Monographs*, *35*, 1-35.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core team the. (2009). nlme: Linear and nonlinear mixed effects models [Computer software manual]. (R package version 3.1-96)

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Directions in Psychological Science*, *12*, 110-114.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*, 74-85.

Saffran, J. R., Pollack, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*, 669-680.

Schulz, L., & Bonawitz, E. B. (2007). Serious fun: Preschoolers play more when evidence is confounded. *Developmental Psychology*, *43*(4), 145-150.

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science: A Multidisciplinary Journal*, *32*(2), 418-444.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization in the first year of life. *Infant Behavior and Development*, *7*, 49-63.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.