What exactly is learned in visual statistical learning?
Insights from Bayesian modeling

Noam Siegelman[1,2], Louisa Bogaerts[2], Blair C. Armstrong[3,4], and Ram Frost[1,2,4]

[1] Haskins Laboratories, New Haven, CT, USA

[2] The Hebrew University of Jerusalem, Israel

[3] University of Toronto, ON, Canada

[4] BCBL, Basque center of Cognition, Brain and Language, San Sebastian, Spain

Corresponding author:
Noam Siegelman
Haskins Laboratories,
06511, New Haven, CT, USA
Email: noam.siegelman@yale.edu

Abstract

It is well documented that humans can extract patterns from continuous input through Statistical Learning (SL) mechanisms. The exact computations underlying this ability, however, remain unclear. One outstanding controversy is whether learners extract global clusters from the continuous input, or whether they are tuned to local co-occurrences of pairs of elements. Here we adopt a novel framework to address this issue, applying a generative latent-mixture Bayesian model to data tracking SL as it unfolds online using a self-paced learning paradigm. This framework not only speaks to whether SL proceeds through computations of global patterns versus local co-occurrences, but also reveals the extent to which specific individuals employ these computations. Our results provide evidence for inter-individual mixture, with different reliance on the two types of computations across individuals. We discuss the implications of these findings for understanding the nature of SL and individual-differences in this ability.

Keywords: *Statistical learning; Bayesian modeling; Online measures; Individual differences.*

It is well documented that humans are highly sensitive to the statistical structure of their surrounding input. Since the seminal investigation by Saffran and her colleagues (Saffran, Aslin, & Newport, 1996), a large number of studies demonstrated learners' ability to detect patterns in continuous streams of sensory input, across modalities and materials, in different stages of development, and under a range of learning conditions (see, Frost, Armstrong, Siegelman, & Christiansen, 2015 for review). This has led to vast interest in this ability – commonly labelled "Statistical Learning" (SL) – and its relation to other cognitive functions.

Yet despite the presumed role of SL across cognition and its numerous experimental demonstrations, key questions regarding its underlying computations are still mostly unanswered. One major controversy is whether learners extract global clusters from the continuous stream, or whether they are primarily tuned to local co-occurrences of pairs of elements. To exemplify, consider a stream consisting of the patterns A-B-C and D-E-F. According to the global view, successful learning means creating representations of the full patterns "A-B-C" and "D-E-F". Learning such patterns could occur through transitional probabilities (TPs) tracking, in which low TP between adjacent elements signal the pattern boundaries. This is a common interpretation of the seminal findings of Saffran et al. (1996), showing that infants recognize “words” in a continuous speech stream. Alternatively, full pattern extraction could also occur without tracking TPs. Such an account assumes that the continuous stream is parsed into repeatedly encountered “global clusters”, or chunks, where representations of chunk candidates are reinforced or decayed given consistent repetitions of the chunks in the stream (e.g., the PARSER model; Perruchet & Vinter, 1998; and see Perruchet & Pacton, 2006; Saffran & Kirkham, 2018; Thiessen, 2017 for discussion).

Yet another possible account of the computation underlying learning in a continuous stream of elements is to assume that learners simply register co-occurrences of local adjacent elements, akin to Hebbian learning. These “atomic” units of learning may eventually cluster into larger more complex chunks with lengthy exposures, however, the local co-occurrences are the primary object of learning, rather than the full global clusters (see, e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Swingley, 2005, for discussion). Thus, in the simple example above, learning the stream consisting A-B-C and D-E-F entails the *independent* learning that element B follows A, that C follows B, E follows D, and F follows E. One major attempt at delineating

between these different accounts was undertaken using the Phantom-word paradigm developed by Endress and Mehler (2009). In this study the familiarization stream included the following 6 patterns, all with TPs=0.5 (each letter stands for one element): *A-B-C, D-B-E, A-F-G, H-F-J, H-I-E, D-I-J*. Based on the structure of these patterns, the sequence A-B-E constitutes a "phantom-word": it maintains a local-TP structure similar to the original six patterns (i.e., TP=0.5), but it never appeared in the familiarization stream as a full chunk. The rationale behind this design is that it can potentially differentiate between learning via local co-occurrences versus full global patterns: If SL relies on the assimilation of the local co-occurrences between elements, phantom sequences like A-B-E would be treated similar to "word" patterns, since they consist of two local elements A-B, and B-E, with identical TP structure. If, however, SL is based on the extraction of larger global patterns from the stream, "words" should be preferred over phantom-words, since the three elements of phantom-words never fully appeared together during familiarization.

This debate has attracted significant attention since it touches upon a fundamental issue in SL theory: understanding the computations involved in learning the regularities embedded in a continuous input stream. Yet, almost ten years since the original report by Endress and Mehler, there are no clear conclusions regarding the nature of SL computations, because studies using this paradigm have provided mixed evidence. For example, from three large-scale multi-experiment investigations one supported local computations (Endress & Mehler, 2009), another supported learning of full patterns (Perruchet & Poulin-Charronnat, 2012), and the third presented mixed evidence across experiments (Endress & Langus, 2017). Notably, these contrasting results were observed despite the use of the same procedure, and in some cases, the exact same stimuli (see Experiment 1 in Perruchet & Poulin-Charronnat, 2012, vs. Endress & Mehler, 2009).

Why similar manipulations with identical stimuli lead to such mixed findings? One possible factor is methodological. The phantom-word paradigm (as well as related experimental procedures) measure success in a recognition test administered only at the end of familiarization (e.g., Giroux & Rey, 2009; Orbán et al., 2008; Perruchet et al., 2014; but see Rey, Minier, Malassis, Bogaerts, & Fagot, 2018 for an exception). Such "offline" tests examine the post-hoc outcomes of SL, which may differ from the representations that are available to learners as they actually learn the statistical structure of the input. Importantly, the testing procedure consists of repeated

presentations of “words”, “phantom-words”, and “part words” which merge with the learned representations, contaminating the assessment of learning (see Siegelman, Bogaerts, Christiansen, & Frost, 2017; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018, for extended discussion). This would inevitably introduce variability in the experimental outcomes. In addition, offline measures are often characterized by mediocre reliability, potentially also contributing to inconsistent findings across studies (and see Siegelman, Bogaerts, & Frost, 2016, for discussion).

However, an alternative and more intriguing account is deeply theoretical. One hypothesis to consider is that the previously mixed results actually reflect a true mix of learning strategies. Specifically, it is possible that not all participants in the experiment employ identical computations for learning, but that different *individuals* employ different computations, reflecting individual sensitivity to local co-occurrences of elements vs. larger patterns. This may contribute to overall different learning scores across different samples of participants. Importantly, examining only group-level mean performance, as was typically the case in previous studies, by-definition cannot reveal such an inter-individual mix.

The goal of the current study is to simultaneously address these methodological concerns and theoretical hypothesis. First, our study refrains from using only an offline test of SL performance. Instead, we focus on an online measure of learning, which monitors response latencies to predictable versus unpredictable stimuli throughout the familiarization phase. Second, we employ an alternative analytical approach – *Bayesian Latent Mixture Modeling*, that speaks to the issue of whether learning proceeds through local co-occurrences versus global patterns on the average, but critically, examines also the extent to which specific *individuals* employ these computations. To preview our findings, we demonstrate that our novel approach, applied to online SL data, leads to new important insights regarding SL computations. Specifically, we show that indeed SL computations cannot be described as based only on local co-occurrences or full patterns, since different individuals display behavior consistent with different computations while processing a continuous stream of visual elements.

Methods

General analytical approach: Bayesian modeling and latent-mixture models. The central analysis in this paper uses a hierarchical Bayesian approach to account for response latencies during a self-paced visual statistical learning task. In this task,

participants are presented with a continuous stream of shapes (which consists of regular patterns) and are required to advance the shapes at their own pace. Learning of regularities is reflected by faster responses to predictable vs. unpredictable stimuli. Importantly to our investigation, the data from this task were fit to a Bayesian model that examined whether an individual's pattern of responses reflects reliance on full patterns versus local co-occurrences (see details below).

In general, Bayesian models are based on the specification of a *generative* model that presumably gives rise to the observed data. To do so, one specifies the relevant latent parameters, prior distributions regarding these parameters (reflecting researchers' a-priori knowledge), and relations between the various parameters as well as between parameters and observed data. Then, the data are used to update the priors and estimate the latent parameters. The output of such a model is therefore a *posterior distribution* for each latent parameter, reflecting researchers' belief regarding each parameter, given priors and data.

To illustrate, consider an IQ test, conducted to estimate a given person's latent score (θ). In a case where we do not have a-priori expectations regarding this person's true score, we may assign a prior that follows the IQ's distribution in the general population: a normal distribution with an expected value of 100 and SD of 15 (i.e., $\theta \sim N(100,15)$). Different a-priori expectations would be reflected in different prior distributions. For example, if a more specific a-priori expectation regarding this person's IQ exists (e.g., IQ of a person who was sampled from a group of gifted individuals), a more restrictive prior with higher expected value can be chosen to reflect that knowledge (e.g., θ~N(130, 15)). If, however, we do not have any knowledge regarding the population from which an observation is sampled, we may want to use a uniform prior distribution, assigning an equal a-priori probability for each value in some wide range (referred to as an 'uninformative prior').

After specifying the priors, the observed data are used to compute a likelihood function (i.e., p(D | θ)) for each value of θ. Bayes theorem is then applied to update the priors given the likelihood values in light of the observed data. This process results in a posterior distribution (p(θ | D), reflecting researchers' updated beliefs regarding the latent parameter, given the priors and after having seen the data. The output of this analysis is a full distribution (as opposed to a single point estimate) for each parameter. Various measures can then be extracted from this distribution, such as its mode, median,

or mean (reflecting a central tendency for the estimated parameter), SD (reflecting the uncertainty in the estimation), or different interval measures (e.g., 95% credible interval, e.g., Chen & Shao, 1999).

Practically, since in most cases there is no simple analytical solutions for computing posterior distributions, Monte Carlo Markov Chains (MCMC) sampling procedures are used to estimate them (see, e.g., Lee & Wagenmakers, 2013, for details). Here we used JAGS (Depaoli, Clifton, & Cobb, 2016), and the *rjags* package in R (Plummer, 2016), to run MCMC samples. In all estimations we used three separate MCMC chains with random starting points. Each chain included 3000 iterations (after 1000 burn-in iterations). To check whether the 3 chains converged to a similar distribution we used the Gelman-Rubin diagnostic measure (Gelman & Rubin, 1992). Values under 1.1 are generally interpreted as high agreement across chains and good model convergence.

For the purposes of our main research question, we use a sub-type of Bayesian modeling: *Latent-mixture modeling* (e.g., Ortega, Wagenmakers, Lee, Markowitsch, & Piefke, 2012). In this approach, two competing models are first specified. In our case, we thus first define two latent models – model A that depicts a local learning process, and model B that reflects SL based on full patterns (see Results section below). We then pit these two models against each other, by defining a larger model that includes the two competing models and a *classification parameter*. This classification parameter examines, for each individual, whether his/her data are more likely given model A compared to model B. We get as output for each individual *i* a posterior distribution for the classification parameter. This distribution reflects the certainty in classification of subject i as following model A (in comparison to model B). Importantly, such models allow for individual differences or *mixture* in the data (hence their name): that is, a situation in which some individuals in the sample are classified as following model A, whereas others as B.

Participants Seventy-six students of the Hebrew University (24 males) participated in the study for payment or course credit. Participants had a mean age of 23.5 (range: 18-36), and had no reported history of reading disabilities, ADD or ADHD. One subject was removed from further analysis since he did not follow the instructions and did not advance the stream of shapes. Data from two additional subjects were discarded due to abnormally slow mean RTs in the self-paced portion of the task: more

than 2 SDs from the sample mean. Analyses below are therefore based on the remaining 73 participants.

Design, Materials, and Procedure Our design is closely similar to the self-paced visual SL task, a paradigm that was shown to produce a reliable and valid online measure of visual SL performance (Siegelman et al., 2018; see also Karuza, Farmer, Fine, Smith, & Jaeger, 2014). The only major change from this previous study was that the regular patterns here were quadruplets rather than triplets. As in a typical SL task, this task consisted of a familiarization phase, followed by a test phase. Materials included 24 complex visual shapes (identical to Siegelman et al., 2018). For each participant shapes were randomly organized to create *six quadruplets*, with a TP of 1 between shapes within patterns. As explained below, the rationale for these larger units was to allow for better differentiation between the two types of underlying computations. The familiarization stream consisted of 24 blocks, where all six quadruplets appearing once (in a random order) in each block.

Before familiarization, participants were told that they would be shown a sequence of shapes, appearing on the screen one after the other. Participants were instructed that some of the shapes tend to follow each other and that their task is to try to notice these co-occurrences. Following Siegelman et al. (2018), and in contrast to standard SL tasks, stimuli did not appear at a fixed presentation rate. Rather, participants were asked to advance the stream of shapes at their own pace, by pressing the space bar each time they wanted to advance to the next shape. RTs for each press were recorded and served as a basis for computing an online measure of SL: the difference in log-transformed RTs between unpredictable and predictable shapes (i.e., between shapes in position 1 within quadruplets vs. the mean RT of shapes in position 2, 3 and 4). Note that importantly, the self-paced data also served as input to the Bayesian models.

Following familiarization, participants completed a two-alternative forced choice (2-AFC) offline test, consisting of 36 trials. In each trial, participants were sequentially presented with two four-item sequences of shapes: (1) a target: four shapes that formed a quadruplet during familiarization (TP=1), and (2) a foil: four shapes that appeared in the familiarization, but never together (TP=0). Foils were constructed without violating the position of the shapes within the original quadruplets (e.g., from the four quadruplets ABCD, EFGH and IJKL, MNOP, a possible foil could be AFKP, but not BGLM). During the offline test, shapes appeared in a fixed presentation rate of

800ms, with an ISI of 200ms between shapes within targets/foils, and a blank of 1000ms between the two sequences. Each of the six targets appeared six times throughout the test, against all six foils (and thus each foil also appeared six times throughout the test, against all quadruplets). 2-AFC test trials were presented in a random order. At the start of the test, participants were instructed that in each trial they would see two groups of shapes and that their task was to choose the group that they were more familiar with as a whole. The offline test score ranged from 0 to 36, according to the number of correct identifications of targets over foils. Given the 2-AFC format, chance performance corresponds to a score of 18/36.

Results

Outlier removal Prior to all analyses we removed RTs outside the range of 2 SD from the participant's mean (4.8% of all trials)[1]. Note also that, to account for variance in baseline RTs, all analyses were conducted on log-transformed RTs (rather than raw RTs). The use of a log-scale allows us to better compare differences in response latencies across individuals with different baselines (see Siegelman et al., 2018, for details).

Basic Findings Before turning to the main research question, we first review some basic findings from the self-paced SL task, following Siegelman et al. (2018). Table 1 presents mean response latencies to shapes in position 1, 2, 3, and 4 within quadruplets. As predicted, there was a significant effect of position on log-transformed RTs (repeated measures ANOVA: $F(3, 216)=11.93$, $p<.001$). Subsequent paired t-tests revealed a difference between shapes in the first versus second position ($t(72)=3.41$, $p=.001$), first versus third position ($t(72)=3.99$, $p<.001$), and first versus fourth ($t(72)=4.38$, $p<.001$). In contrast, there was no evidence for an RT difference between shapes in second versus third positions ($t(72)=0.52$, $p=.60$) and third versus fourth ($t(72)=-0.1$, $p=.99$)[2]. Figure 1 presents the log-transformed RTs to shapes in positions 1, 2, 3, and 4 over the course of the familiarization phase.

[1] This outlier removal criterion was a-priori selected to match that of Siegelman et al. (2018). It is important to emphasize, however, that our results are not limited to this approach and generalize to a more conservative procedure of outlier removal. In the Supplementary Material we thus repeat the main analyses below, only removing trials with RTs shorter than 100ms or longer than 5000ms, showing qualitatively similar results.

[2] All p-values here are two-tailed, and are reported without correction for multiple comparisons. It is worth noting however that applying a Bonferroni correction does not change the overall pattern of results, as all significant tests remain significant also under a stricter threshold.

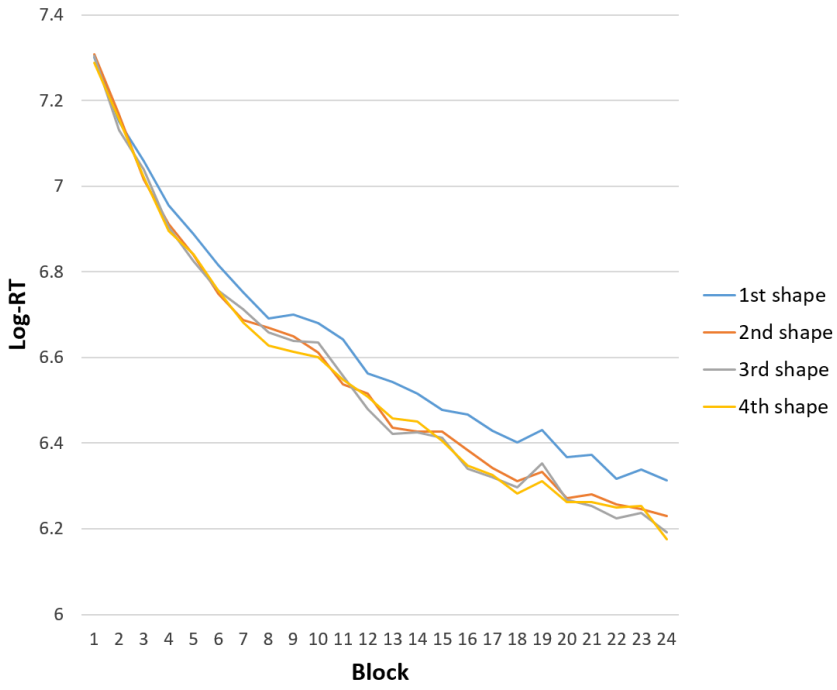

Figure 1. Response latencies to shapes in first, second, third, and fourth position within a quadruplet over familiarization blocks.

Next, we examined the time-course of SL during familiarization, as reflected by the change in the online measure (log-RT difference between unpredictable and predictable shapes) across the 24 blocks in the familiarization stream (Figure 2). Replicating Siegelman et al. (2018), this trajectory followed a logarithmic function. One-sample t-tests revealed significant learning (log-RT difference larger than zero, $p<0.05$) in all blocks from block 9 until the end of familiarization, showing stable group-level learning already after 9 repetitions (cf. significant learning from block 7 onwards in Siegelman et al., 2018). We also calculated the reliability of the online measure of learning using a split-half procedure (i.e. the correlation of log-RT difference between odd and even quadruplets) finding a very high estimate of $r=0.9$.

Lastly, we examined the individual-level correlation between the online SL measure and the 2-AFC offline test. As in Siegelman et al., (2018), a positive significant correlation was found: $r=0.33$, $p=0.004$. Overall, these basic findings replicate Siegelman et al.'s previous findings and re-validate the self-paced SL paradigm using patterns with four as opposed to three elements.

Table 1. Means and SEs for RTs and log-transformed RTs for shapes in first, second, third and fourth positions within quadruplets.

	1st position	**2nd position**	**3rd position**	**4th position**
Raw RT (*SD*)	960.5 (*60.6*)	884.3 (*47.8*)	880.5 (*46.9*)	878.1 (*47.4*)
Log-transformed RT (*SD*)	6.62 (*0.062*)	6.55 (*0.056*)	6.54 (*0.054*)	6.54 (*0.055*)

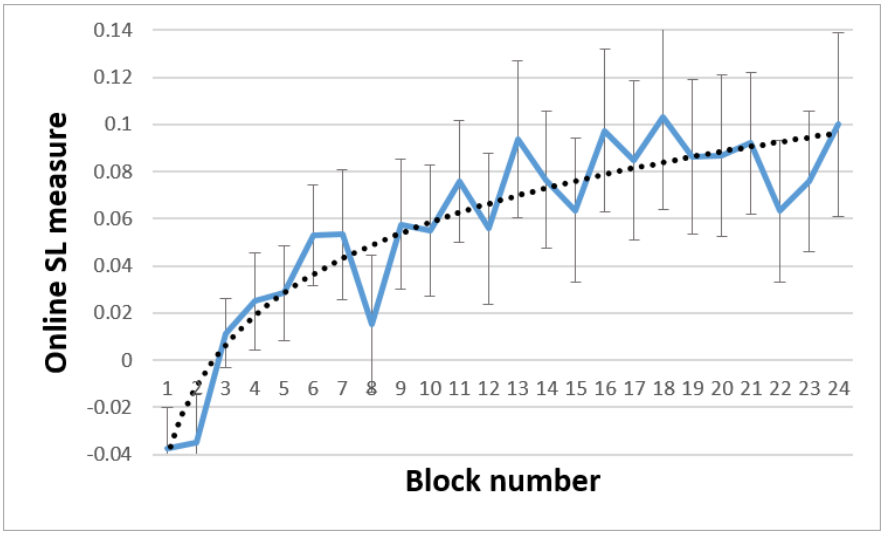

Figure 2. Learning trajectory as reflected by the change in the online measure (i.e., difference between log-RT to predictable vs. unpredictable shapes) throughout familiarization blocks. Error bars represent standard errors. The dashed line represents the best logarithmic fit.

Bayesian mixture-model. As described above, the first step in latent-mixture modeling is to specify two competing models, depicted in Figure 3. The full-pattern model assumes that RTs to predictable shapes within a pattern are uniformly faster than RTs to the first (unpredictable) shape. The local co-occurrence model assumes that RTs within a pattern may be faster or slower given *the independent learning* of co-occurrences of shapes. We follow a graphical notation (based on Lee & Wagenmakers, 2013) that represents latent parameters using white nodes, and observed data using grey nodes. Priors for latent parameters are listed to the right of each model.

The top panel depicts a local co-occurrence model. The input for this model is the log-transformed RTs for shapes in position 1, 2, 3, and 4, in each quadruplet *j*, for each participant *i* (averaged across blocks). RTs for each position are assumed to be drawn from a normal distribution, with some expected value for each position: thus, the parameter $\mu_{i,j}^{1}$ reflects the expected log-RT for the shape in position 1 for participant *i* and for the quadruplet *j*, $\mu_{i,j}^{2}$ reflects the expected log-RT for the shape in position 2 for

participant *i* for quadruplet *j*, and so on. For simplicity, standard deviations are assumed to be equal in all positions within a participant. Importantly, the expected values are determined through another set of latent variables: $Z^1_{i,j}, Z^2_{i,j}, Z^3_{i,j}$. These are Bernoulli trials that reflect whether a given participant *i* learned some local co-occurrence in quadruplet *j*: Z=1 stands for successful learning of this co-occurrence and Z=0 reflects no learning. Importantly, there are three such Bernoulli trials for each quadruplet: $Z^1_{i,j}$ reflects learning of the co-occurrence between the first and the second elements within a quadruplet, $Z^2_{i,j}$ reflects the co-occurrence between the second and the third elements; and $Z^3_{i,j}$ reflects the co-occurrence between the third and the fourth elements. The probability of these Bernoulli trials is determined via another parameter k_i, which reflects the percent of co-occurrences learned by a participant *i* (out of the full array of local co-occurrences in the stream). The parameter Δ_i reflects the speed-up in log-RTs given a learned co-occurrence – that is, given that a participant *i* learned some local co-occurrence A-B, Δ_i is the speed-up in log-RT for the shape B, compared to the shape A. The expected value of the shape in position 1 (i.e., an unpredictable shape) is always set to some baseline RT, μ_i^{base}, estimated for each participant. Then, the parameters Z's and Δ are used to determine the expected values of RTs in positions 2, 3, and 4. Specifically, the expected RT for shape in position 2 in quadruplet *j* for participant *i* would be similar to the baseline RT in the quadruplet (reflecting an unpredictable shape) if the participant did not learn their co-occurrence (i.e., when $Z^1_{i,j}$=0). In contrast, if the participant did learn this transition ($Z^1_{i,j}$=1), the expected RT for position 2 would be the expected RT for position 1 minus the speed-up parameter Δ_i. Similarly, the expected RT for position 3 would be equal to that the baseline RT if the participant did not learn the co-occurrence of 2 and 3 ($Z^2_{i,j}$=0). If, however, the participant did learn the transition between position 2 and 3 ($Z^2_{i,j}$=1), the expected RT for position 3 would be faster by Δ compared to that of position 2. The same holds for the transition between position 3 and 4.

To emphasize, since this model simulates learning of local co-occurrences, $Z^1_{i,j}, Z^2_{i,j}, Z^3_{i,j}$ are independently estimated for each subject in each quadruplet. *This is because this model assumes that a participant can either learn, or not learn, each local co-occurrence within each pattern, regardless of other transitions*. As a result, this model posits that RTs may be faster or slower even within a quadruplet based on the

specific learned co-occurrences. For example, in the pattern ABCD, if only the co-occurrence BC was learned, there will be faster responses to shape C, but slower expected RTs to shapes in positions A, B, and D. Note also that perfect learning in this model (k=100%, learning of all possible co-occurrences) would result in graded RTs as a function of position, where position 4<position 3<position 2<position 1 (since learning all co-occurrences entails speed up to position 2 vs. 1, position 3 vs. 2, and position 4 vs. 3).

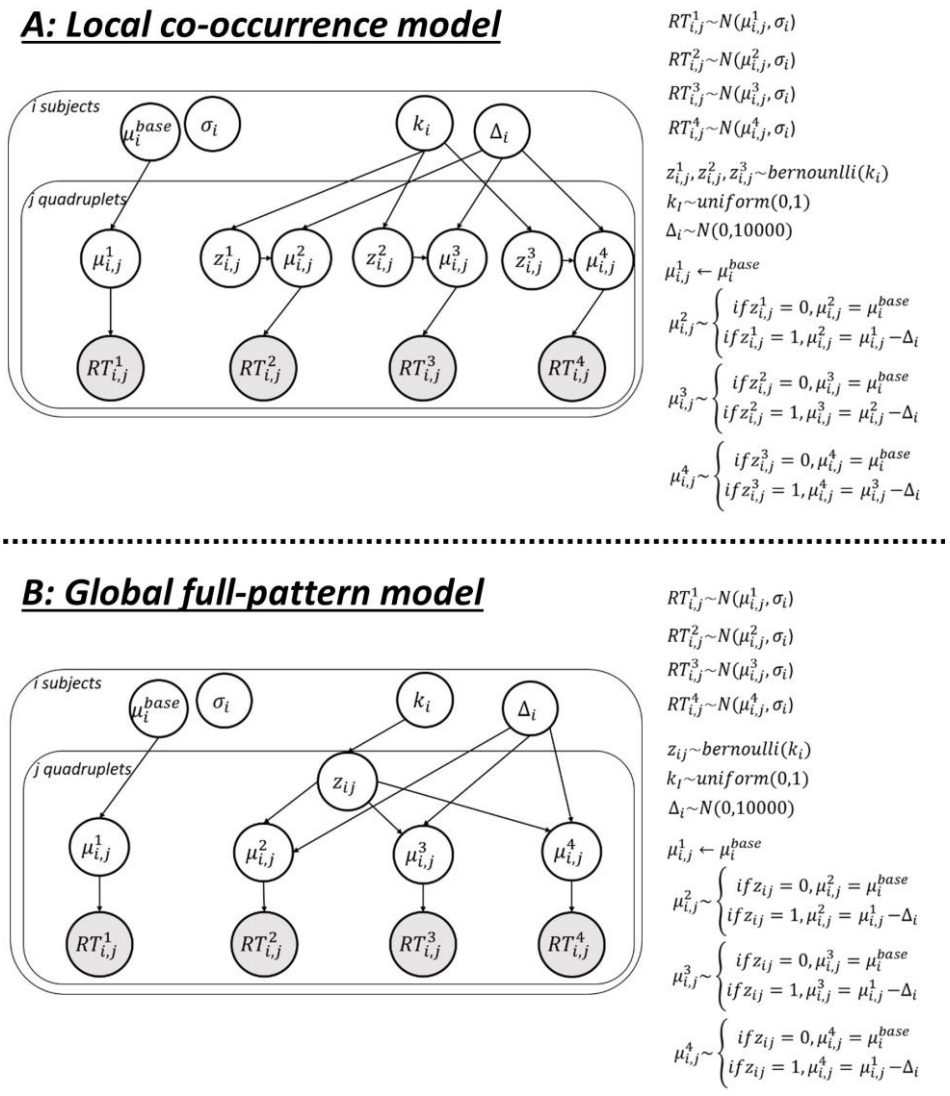

Figure 3. Graphical depiction of the two competing models. The top panel depicts the local co-occurrence model. The bottom panel shows the full pattern model. Note that input to both models were log-transformed RTs.

As a side note, even without the use of the latent-mixture model, which is the central aim of the current investigation, some interesting insights can be gained simply

by running this first model and examining the resulting posterior distributions. For example, the proportion of learned co-occurrences for each participant can be drawn from the model by examining the posterior distribution of k_i. Figure 4, panel A, presents two illustrative posterior distributions of this parameter: an individual who learned a large proportion of the embedded co-occurrences (mean=74%), and an individual who learned a smaller portion (mean=35%). We can also estimate the full distribution of proportion of learned co-occurrences across individuals. To do so, we take the mean of the posterior distribution for k_i for each subject, and then plot the distribution of these mean k_i's across subjects. Figure 4, panel B, presents the resulted histogram, which shows that on average participants learn 48.2% of the co-occurrences embedded in the stream (SD=12.1%).

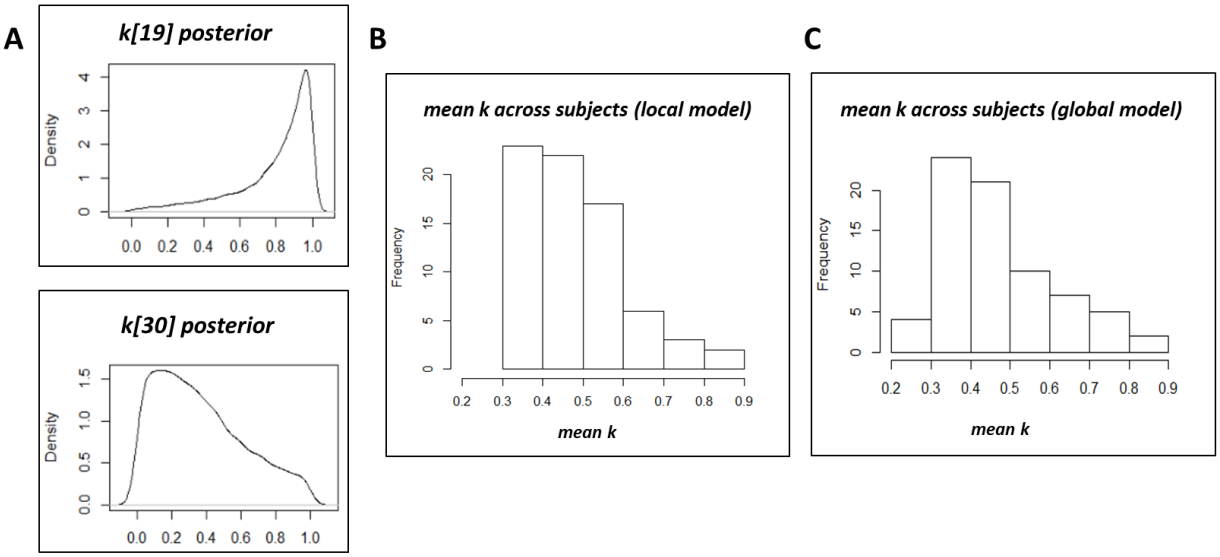

Figure 4. Parameter estimation based on the separate models (local co-occurrences or full patterns). Panel A: examples of posterior distributions of k for two subjects under the local model. Panel B: mean of k across participants, under the local co-occurrence model. Panel C: mean of k across participants, under the full-pattern model.

Returning to our central research question, the bottom panel of Figure 3 depicts a global full pattern SL model. The input for this model is identical to the local co-occurrence model: namely, log-RT for each quadruplet for each participant, in positions 1 to 4 (averaged across blocks), and so is its overall architecture. The critical difference between this and the local co-occurrence model is that it only has a single Z parameter per quadruplet for each participant. This reflects the fact that according to the global

full-pattern model, each quadruplet can be either learned, or not learned, as a whole. Consequently, the interpretation of the parameters k_i and Δ_i changes in comparison to the previous model: k_i now depicts, for each subject *i*, the percent of learned *patterns* (as opposed of local co-occurrences), and Δ_i depicts the speed-up given a learned pattern of all shapes in position 2, 3, and 4. The specification of the expected values of each position within-quadruplet is now different too. For each subject *i*, quadruplet *j*, the expected values of positions 2, 3, and 4 ($\mu_{i,j}^2$, $\mu_{i,j}^3$, $\mu_{i,j}^4$) would be identical to that of the baseline RT (which is identical to the expected RT of position 1) if the pattern was not learned (that is, if $Z_{i,j}$=0). In contrast, if this pattern was learned ($Z_{i,j}$=1), the expected RTs for positions 2-4 in this quadruplet would be set to the first position RT minus the speed-up parameter Δ_i. Note that under this model, positions 2-4 within a quadruplet always have identical expected response latencies. This would also be the case under perfect learning of all patterns (i.e., k=100%) under this model. Again, as a side note, simply running this model on the RT data can already provide some insights. For example, Figure 4, Panel C, shows the histogram of mean k_i, now reflecting the average percent of learned *patterns*, across participants (mean=47%, SD=14.2%).

Most central to the current investigation, after specifying these two models we combined them to a single latent-mixture model by adding a classification parameter s_i: A Bernoulli trial estimated as either 1 or 0 in each iteration of the model. This classification parameter reflects the group membership of each participant *i*: where *s*=1 reflects a classification of the participant as a *local co-occurrence learner*; *s*=0 reflects a classification as *full-pattern learner*. Note that the classification parameter (s_i) is simply an additional parameter to be estimated in a larger Bayesian model that includes both the local and global models. Thus, the model estimates a posterior distribution for the group classification parameter from the specified prior and the likelihood function calculated given the data. As a result, the mean of this classification parameter across MCMC iterations reflects the model's certainty in classifying participant *i* as a local learner (versus a full pattern learner). The a-priori distribution of s_i=1 was defined as a uniform distribution from 0 to 1, meaning that there was no a-priori assumption regarding the probability of a given subject to be classified as a local (or global) learner. Note also that the model was characterized by good convergence on the s_i parameter: in all subjects the point estimate of Gelman-Rubin diagnostic measure was smaller than

1.1, and in all but one participant the upper boundary of the 95% CI of the measure was also smaller than 1.1.

Figure 5 presents examples of posterior distributions of the latent parameter s_i from three individuals. On the left panel of this figure, an example for a local co-occurrence learner: reflected by classification as a local learner in 84% of the model iterations. In contrast, the middle panel shows an example for a global learner, which was classified as a full pattern learner in 93% of iterations. The right panel presents an additional interesting case: a participant that was classified either as a local or global learner in ~50% of the iterations, thus showing no clear tendency for neither model. Looking at this distribution alone, it is unclear whether this is because this subject used both two strategies interchangeably, or whether s/he just did not learn any of the statistical properties and therefore could not be classified successfully (but see General Discussion for an additional investigation, comparing group classification to offline performance).

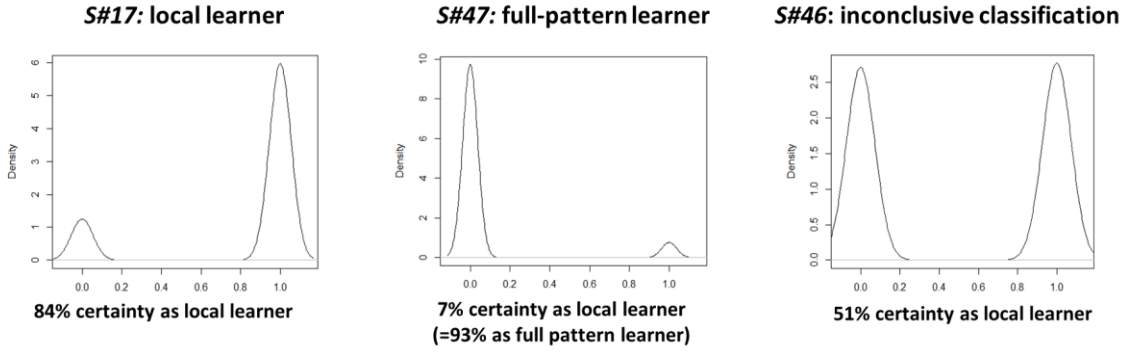

Figure 5. Examples of posterior distributions of group classification for three subjects. Left: example of a local co-occurrence learner; middle: example of a full-pattern learner; right: example of a subject that was classified in ~50% as each model.

The critical question for our current investigation has to do with the distribution of group membership (s_i) across participants. We thus next extracted the mean value of s_i for each individual (reflecting the model's overall tendency to classify subject *i* as a local vs. global learner). Figure 6 presents the distribution of mean s_i across individuals. On average, mean group classification was equal to 47.8%. This value is very close to 50% suggesting that, overall, there is no clear group-level tendency to either local or full-pattern learning. Yet, a closer inspection of Figure 5 leads to two more important conclusions. First, the distribution of group classification was close to symmetrical: despite the fact that slightly more participants were classified as local learners (40/73

subjects with s_i>50%) as opposed to full-pattern learners (33/73 subjects with s_i<50%), this ratio was close to what is expected in a fully symmetrical distribution (36.5 subjects out of 73). Most importantly, whereas nearly symmetrical, the distribution was not normal around its mean, displaying substantial inter-individual differences (also apparent with the high SD of 28.1%). Namely, whereas some individuals were classified with a high certainty as local co-occurrence learners, others were clearly classified as full patterns learners. We wish to emphasize that many subjects clearly followed either the local or the global model: 42% of the subjects (31/73) are twice as likely to be local learners according to the model (mean of s_i>66.67%), and 34% (25/73) are twice as likely to be global learners according to the model (mean of s_i<33.33%). This suggests that the majority of subjects clearly exhibit an overall tendency to learn either locally or globally[3]. Together, the results thus point to inter-individual mixture in the reliance on local co-occurrences versus full patterns. We return to this point in the General Discussion, below.

[3] It is worth noting that while the local and global models are mostly similar in their specification, the two models diverge slightly in their complexity: the local model has two more parameters compared to the global model. This raises a possible concern that the distribution of group classification might be slightly biased towards the more complex (i.e. flexible) model, if the larger number of parameters to be estimated leads to a higher chance of overfitting. To ensure that this is not the case, we ran a simulation in which we sampled hypothetical subjects under a null hypothesis of no learning (i.e. no difference between positions 1, 2, 3, and 4, other than random noise). Virtually all simulated subjects had a mean s_i around 0.5, as expected under no signal (no actual learning). This suggests that there is no bias in classification towards one model. The full details and results of this simulation are presented in the Supplementary Material.

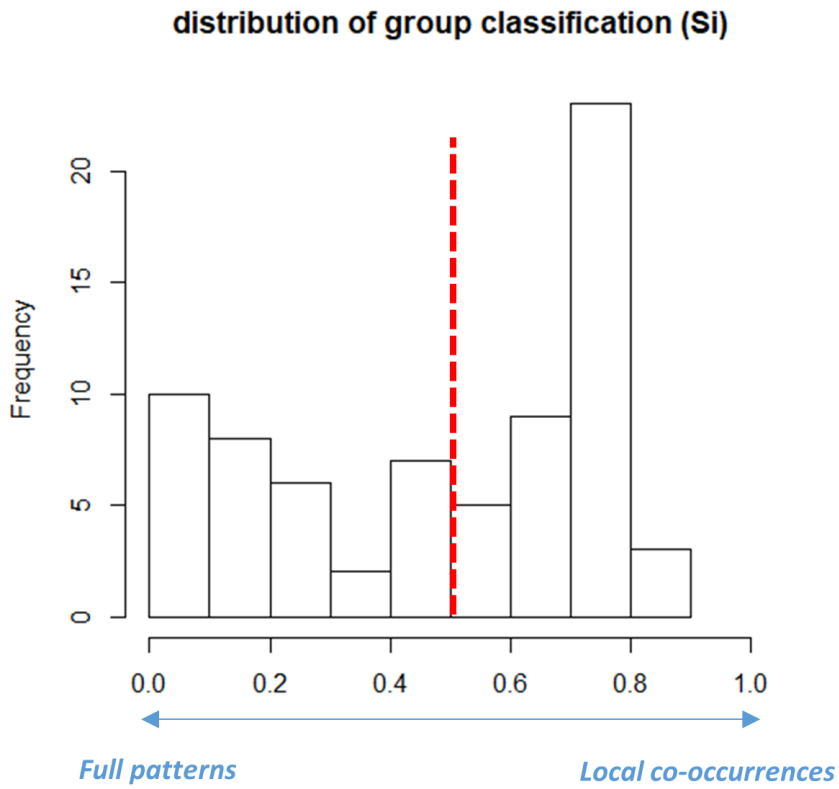

Figure 6. Distribution of mean group classification across participants. Dashed line represents 0.5. Values closer to 1 reflects classification as local learning; values closer to 0 reflects full pattern learning.

Following up on this finding, we next examined the time-course of learning: that is, whether there is a trajectory towards reliance on local co-occurrences versus global patterns as learning proceeds (see, e.g., Rey et al., 2018). For example, one possibility is that learners start by attending to local co-occurrences, but begin to merge them and attend to larger units after extensive exposure (see Batterink & Paller, 2017 for a related discussion). To examine this issue, we re-ran the latent-mixture analysis on data from each 6 consecutive repetition blocks (i.e., blocks 1-6, 7-12, 13-18, and 19-24). Figure 7 presents the distribution of mean s_i across individuals in these four quarters of familiarization. As can be seen, in the first two quarters of the familiarization phase (blocks 1-6 and 7-12) there was no clear tendency to rely on either local co-occurrences or global patterns, with the majority of subjects having a mean classification value around 0.5 (73% and 77% of subjects with a mean s_i between 1/3 and 2/3 in the first and second quarter, respectively). Only in the third and fourth quarter of the familiarization phase a clear classification into group membership emerged, with the majority of subjects having a clear group distinction (62% and 70% of subjects with

mean s_i smaller than 1/3 or larger than 2/3 in the third and fourth quarters). Importantly, the distributions of group membership in blocks 13-18 and in blocks 19-24 were similar, and both resembled the group membership distribution based on the full familiarization phase (Figure 6 above). These results suggest that there is no clear shift from one strategy to another over the course of learning. Rather, participants increasingly lean towards one strategy or the other as learning proceeds. To further examine this issue, we also calculated the correlations between each individuals' group membership estimation (mean of s_i) between the four quarters. As shown in Table 2, there were no significant correlations between group classification in the first two quarters, or between the classification in the first and second quarter to that of later quarters (all r's < 0.2). In contrast, there was a strong positive correlation between classification in the third and fourth quarter (r=0.67). This again suggests that in the first two quarters subjects do not have a reliable and clear reliance on either strategy, and only in the later stages of learning they lean towards a global or local strategy, which remains consistent for the reminder of the learning phase.

Table 2: Correlations between individuals' group classification in the four quarters of the familiarization phase. *p*-values are shown in parenthesis; significant correlations are in bold.

	Blocks 1-6	Blocks 7-12	Blocks 13-18	Blocks 19-24
Blocks 1-6	***	0.01 (0.98)	0.04 (0.74)	0.09 (0.46)
Blocks 7-12		***	0.15 (0.21)	0.2 (0.1)
Blocks 13-18			***	**0.67** (<0.001)
Blocks 19-24				***

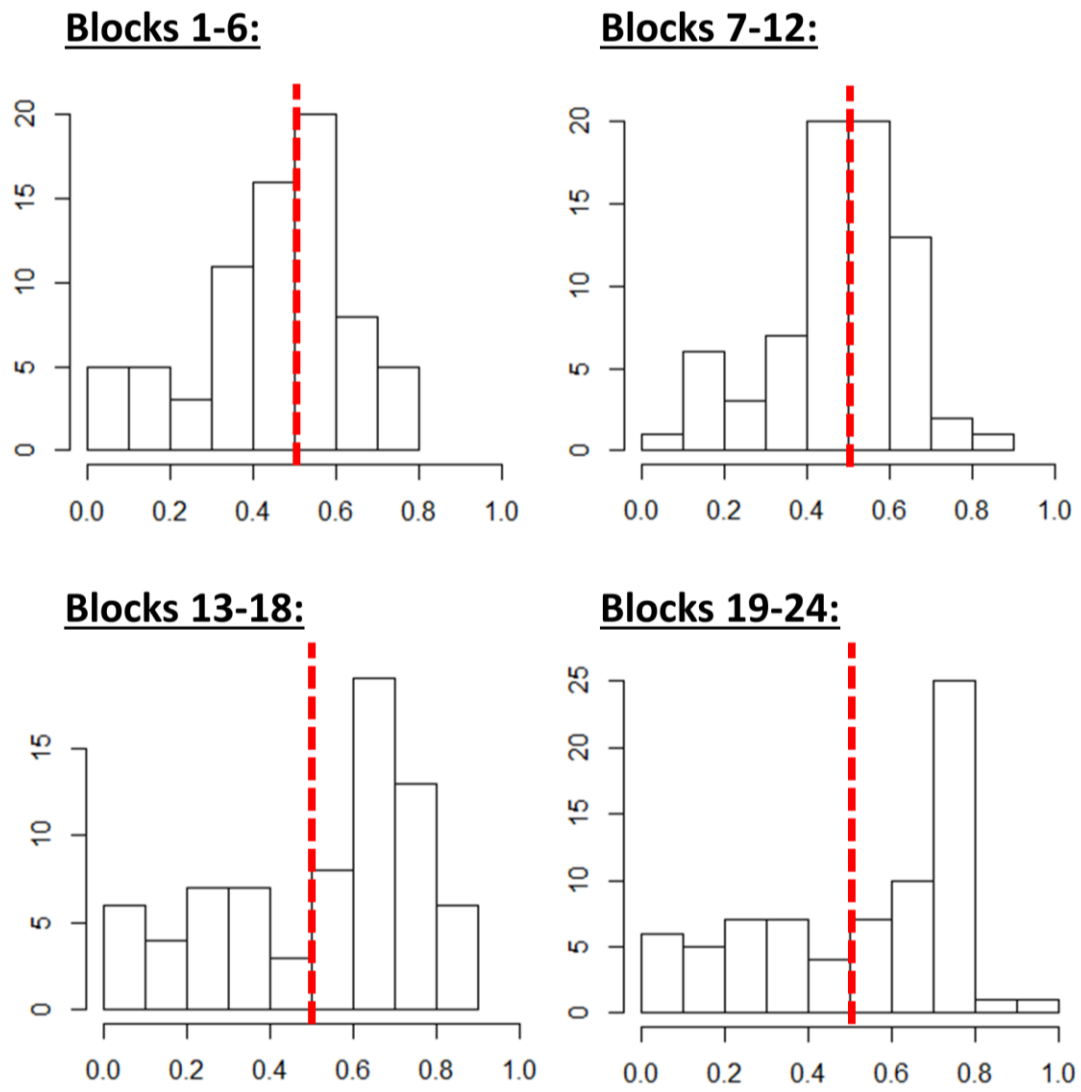

Figure 7. Distribution of mean group classification over the four quarters of the familiarization phase (blocks 1-6, 7-12, 13-18, 19-24). Dashed line represents 0.5. Values closer to 1 reflects classification as local learning; values closer to 0 reflects full pattern learning.

General Discussion

What is learned in visual SL, the local co-occurrences between elements or global patterns? This critical question was the center of multiple previous studies as it represents a fundamental building block of a theory of how complex patterns, embedded in a continuous input stream, are learned. Most investigations so far have searched for a binary answer, in the form of either account A, or account B. The current investigation suggests, however, that the answer to this question is neither *A* nor *B*, rather it is *both*, differing between individuals. This is reflected in the results of a Bayesian model, providing strong evidence for an inter-individual mixture of local co-

occurrences and full patterns in SL, showing different reliance on the two types of computations across different participants. This provides a novel perspective regarding the computations underlying SL, suggesting that multiple types of computations co-exist (at least across different individuals). As such, the current results may go a long way in explaining the inconclusive and inconsistent results observed in previous studies probing this issue.

Our local and global generative models are not meant to directly represent computational models previously proposed in the literature. However, they are undoubtedly conceptually related to some of them. Specifically, the simple recurrent network model (SRN; Elman, 1990; Mirman, Graf Estes, & Magnuson, 2010) is based on prediction of adjacent co-occurrences (at least in a network where there is only one memory layer). In contrast, chunking models, such as PARSER (Perruchet & Vinter, 1998), propose that high frequency sequences are clustered together as chunks, and thus they priorities the larger units embedded in the stream (see also Giroux & Rey, 2009; Slone & Johnson, 2018). Our generative models provide an important insight regarding the contrast between these two broad types of learning architectures, suggesting that models with only one type of computations may be an over-simplification of SL behavior. Indeed, a recent model (TRACX(2), Mareschal & French, 2017) is based on a combination of local-TP learning and chunking, and is more compatible with the current findings (even though in this model there is no direct reference for inter-individual mixture). More generally, our results call for caution when interpreting data on the group level as supportive of contrasting model predictions, and for a careful examination of individual-level behavior patterns.

Of course, in formulating our Bayesian models, it was necessary to make multiple assumptions about the parameters that drive behavior. As researchers move towards being more computationally explicit about their specific theoretical accounts, the formal versions of their accounts will inevitably differ to some degree from the exact implementations that were tested here. Importantly, the current framework provides a clear way of incorporating and testing such modifications in a formal manner. Different assumptions regarding either the local or global model can be reflected by changes to the generative models, which can then be pitted against one another using a latent mixture model. We illustrate this capacity in the Supplementary Material, wherein we change the exact assumptions underlying how RTs decrease across successive correctly predicted elements in the local model. In this alternative

local model, a learned transition, regardless of whether it is in the first, second, or third position, always results in a speed-up relative to baseline (and not relative to the preceding shape). In this case, the results were very similar, but non-identical to those reported above. This also demonstrates the robustness of our main conclusion – inter-individual mixture of local and global learning - across a variation of our proposed local model.

On a more concrete methodological level, the current work also joins recent studies in exemplifying the usefulness of online SL measures – that is, measures that track learning as it unfolds. It shows that using an online SL task, especially when combined with a generative model, offers new insights into SL computations. In the current case, this approach revealed new information regarding local versus global SL, an information that typical offline measures are generally blind to. This becomes very apparent when looking at the correlation between the individual-level classification as local/full-pattern learner (s_i), and offline test scores, presented in Figure 8. As can be seen, there was an overall negative correlation between group classification and offline test performance (r=-0.42, p<.001). Examining the scatter plot suggests that it stems from the fact that full-pattern learners (i.e., those for whom mean group classification approaches 0) show in the vast majority of cases perfect or near-perfect offline performance. Importantly, however, Figure 6 shows that individuals can achieve such perfect offline test accuracy either via learning the local co-occurrences or via assimilation of full patterns (in the 2-AFC test, a quadruplet can be preferred over the foil already given one learned co-occurrence). This is reflected by the fact that high offline test scores are present both for participants who exhibit global learning ($s_i \rightarrow 0$) and those who show local co-occurrence learning ($s_i \rightarrow 1$).

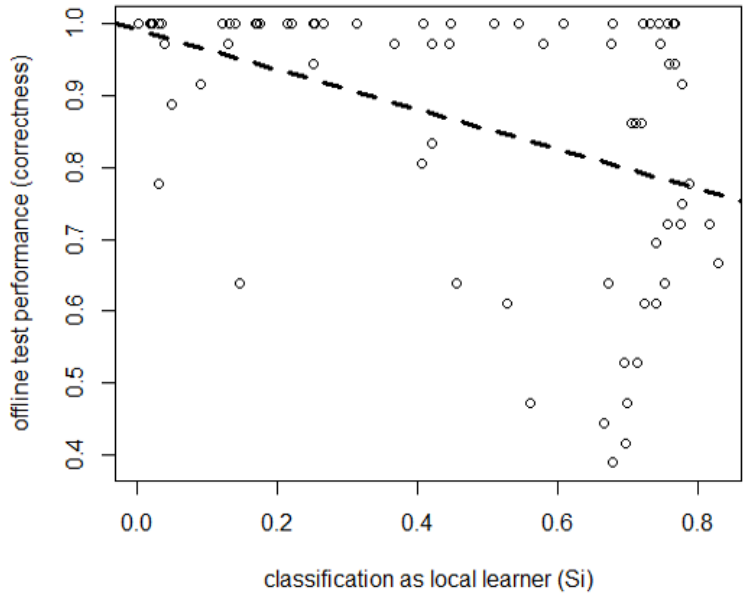

Figure 8. Correlation between group classification (x-axis) and offline test performance (y-axis). Dashed line reflects best linear fit.

This last point also raises an important general implication for research on individual-differences in SL. Our results suggest that individuals differ in the underlying computations they use to learn a new set of statistical properties. This is a feature that was so far overlooked in studies of individual differences in SL, which focused only on differences in the overall success in assimilating the statistical properties from the input. Hence, our study provides an additional potential layer of individual-differences in SL: the tendency to learn via tracking of local co-occurrences versus reliance on larger chunks. Given recent claims regarding the importance of chunking in language acquisition and processing (Christiansen & Chater, 2016; Page & Norris, 2009), it would be interesting for future research to investigate whether a tendency to rely on larger chunks during SL has a unique predictive value in accounting for variance in linguistic abilities. That is, in contrast to the common experimental approach which estimates the correlation between some overall measure of SL performance and some linguistic outcome (see Siegelman et al., 2017), our findings raise the possibility that differences in the *underlying computations* through which each individual extracts regularities from the input carry unique explanatory power. To re-iterate, previous works on individual differences in SL overlooked this possibility due to their reliance on coarse-grained measures of SL (either offline or online) that only examine the extent of sensitivity of an individual to a set of statistical regularities without considering the specific computations that yielded it.

Lastly, although our work shows strong promise for advancing the understanding of SL computations, some open questions should be underlined. The first question arises from the fact that a non-negligible number of individuals were not clearly classified to either the local or global model. Interestingly, some of these subjects still presented some successful SL computations: Figure 6 shows that a number of subjects without clear classification into local/global computations (s_i around 0.5) nevertheless exhibited high offline performance. How to classify such subjects remains an open question, and it is possible that more data would have led to a clearer classification of these subset of participants. Thus, future research should aim to develop more refined models and designs with a large enough number of patterns to allow an even higher detection rate of the different types of computations. Second, in contrast to many previous studies, we focused here on visual, rather than auditory SL. Future research should examine whether similar inter-individual mixture occurs also in the auditory modality. Such research will also have to deal with an outstanding methodological challenge, and come up with reliable online SL measures in the auditory domain (see Batterink, 2017; Batterink & Paller, 2017; Kuppuraj, Duta, Thompson, & Bishop, 2018, for possible avenues). Third, our design used larger embedded patterns: four-element long (i.e., quadruplets). This stands in contrast to typical SL studies using mostly triplets (or sometimes pairs) of elements. The rationale behind this design was to have better differentiation between the two models, by having a larger number of transitions within patterns. Future work is left with examining whether the current results generalize to other learning situations, either with another fixed length of patterns, or non-uniform distribution of pattern lengths (e.g., Hoch, Tyler, & Tillmann, 2013). Fourth, our models only account for the learning of adjacent contingencies, disregarding the assimilation of non-adjacent dependencies, despite recent evidence that the two types of computations can occur in parallel (Vuong, Meyer, & Christiansen, 2016). Future models can be used to account for such concurrent learning of different types of information.

Taken together, our theoretical and methodological approach, as well as our insightful pattern of results, have shed important new light on debates surrounding the computations underlying SL. It stresses the importance of assessing learning online, taking into account critical differences in the computations underlying learning across different individuals, and in developing formal models of a theory's assumptions. Going forward, the computational framework used here can also serve as a foundation

for comparing the performance of alternative theoretical accounts in explicit, quantifiable terms, allowing for the assessment of how major qualitative differences and subtle quantitative differences across models could refine our understanding of SL computations. This approach should therefore prove valuable in moving beyond underspecified verbal accounts to a fully fleshed out account of SL phenomena.

Supplementary material

Code and raw data are available via Open Science Framework at:
https://osf.io/enp6q/?view_only=d39b4988c58b476b8190729b0a3f5f8f
[DOI 10.17605/OSF.IO/ENP6Q]

Acknowledgements

This paper was supported by the ERC Advanced grant awarded to Ram Frost (project 692502-L2STAT), and the Israel Science Foundation (Grant 217/14 awarded to Ram Frost), and NSERC grant DG-502584 to Blair Armstrong. Noam Siegelman is a Rothschild Yad-Hanadiv post-doctoral fellow. Louisa Bogaerts received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 743528 (IF-EF), at the Hebrew University. We wish to thank Per-li Piro for her help in data collection. We also thank Amy Perfors and an anonymous reviewer for their helpful comments.

References

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science*, *28*(7), 921–928. https://doi.org/10.3837/tiis.0000.00.000

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004

Chen, M. H., & Shao, Q. M. (1999). Monte carlo estimation of bayesian credible and hpd intervals? *Journal of Computational and Graphical Statistics*, *8*(1), 69–92. https://doi.org/10.1080/10618600.1999.10474802

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*(62), 1–72. https://doi.org/10.1017/S0140525X1500031X

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics*, *41*(6), 628–649. https://doi.org/10.3102/1076998616664876

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. https://doi.org/10.1016/j.cogpsych.2016.11.004

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351–367. https://doi.org/10.1016/j.jml.2008.10.003

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. https://doi.org/10.1016/j.cognition.2010.07.005

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception.

Cognitive Science, *33*(2), 260–272. https://doi.org/10.1111/j.1551-6709.2009.01012.x

Hoch, L., Tyler, M. D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic Bulletin & Review*, *20*(1), 142–147. https://doi.org/10.3758/s13423-012-0309-8

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., & Jaeger, T. F. (2014). On-line Measures of Prediction in a Self-Paced Statistical Learning Task. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 725–730).

Kuppuraj, S., Duta, M., Thompson, P., & Bishop, D. (2018). Online incidental statistical learning of audiovisual word sequences in adults: A registered report. *Royal Society Open Science*, *5*(2), 171678. https://doi.org/10.1098/rsos.171678

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. *Bayesian Cognitive Modeling: A Practical Course*. https://doi.org/10.1017/CBO9781139087759

Mareschal, D., & French, R. M. (2017). TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160057. https://doi.org/10.1098/rstb.2016.0057

Mirman, D., Graf Estes, K., & Magnuson, J. S. (2010). Computational modeling of statistical learning: Effects of Transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486. https://doi.org/10.1111/j.1532-7078.2009.00023.x

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745–2750. https://doi.org/10.1073/pnas.0708424105

Ortega, A., Wagenmakers, E. J., Lee, M. D., Markowitsch, H. J., & Piefke, M. (2012). A bayesian latent group analysis for detecting poor effort in the assessment of malingering. *Archives of Clinical Neuropsychology*, *27*(4), 453–465. https://doi.org/10.1093/arclin/acs038

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one

phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–8. https://doi.org/10.1016/j.tics.2006.03.006

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica*, *149*, 1–8. https://doi.org/10.1016/j.actpsy.2014.01.015

Perruchet, P., & Vinter, A. (1998). PARSER: A Model for Word Segmentation. *Journal of Memory and Language*, *39*, 246–263. https://doi.org/10.1006/jmla.1998.2576

Plummer, M. (2016). rjags: Bayesian graphical models using MCMC. *R Package Version 4-6*. https://doi.org/http://cran.r-project.org/package=rjags

Rey, A., Minier, L., Malassis, R., Bogaerts, L., & Fagot, J. (2018). Regularity Extraction Across Species: Associative Learning Mechanisms Shared by Human and Non-Human Primates. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12343

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., & Frost, R. (2016). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, 1–15. https://doi.org/10.3758/s13428-016-0719-z

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "Learning" in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science*, *42*(53), 692–727. https://doi.org/10.1111/cogs.12556

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants

represent visual sequences in terms of chunks. *Cognition*, *178*, 92–102. https://doi.org/10.1016/j.cognition.2018.05.016

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001

Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160056. https://doi.org/10.1098/rstb.2016.0056

Vuong, L. C., Meyer, A. S., & Christiansen, M. H. (2016). Concurrent learning of adjacent and nonadjacent dependencies. *Language Learning*, *66*(1), 8–30.