COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12803

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12803

What Determines Visual Statistical Learning Performance? Insights From Information Theory

Noam Siegelman,[a,b] Louisa Bogaerts,[a] Ram Frost[a,b,c]

[a]*Department of Psychology, The Hebrew University of Jerusalem*
[b]*Haskins Laboratories*
[c]*Basque Center of Cognition, Brain and Language (BCBL)*

Received 21 January 2019; received in revised form 17 October 2019; accepted 5 November 2019

Abstract

In order to extract the regularities underlying a continuous sensory input, the individual elements constituting the stream have to be encoded and their transitional probabilities (TPs) should be learned. This suggests that variance in statistical learning (SL) performance reflects efficiency in encoding representations as well as efficiency in detecting their statistical properties. These processes have been taken to be independent and temporally modular, where first, elements in the stream are encoded into internal representations, and then the co-occurrences between them are computed and registered. Here, we entertain a novel hypothesis that one unifying construct—the rate of information in the sensory input—explains learning performance. This theoretical approach merges processes related to encoding of events and those related to learning their regularities into a single computational principle. We present data from two large-scale experiments with over 800 participants tested in support for this hypothesis, showing that rate of information in a visual stream clearly predicts SL performance, and that similar rate of information values leads to similar SL performance. We discuss the implications for SL theory and its relation to regularity learning.

Keywords: Statistical learning; Rate of information; Information theory; Visual processing

1. Introduction

Statistical learning (SL)—learning from the distributional properties of sensory input across time and space—is taken to be involved in a wide range of basic and higher-order cognitive functions (Frost, Armstrong, Siegelman, & Christiansen, 2015). In the last two decades, therefore, extensive research has focused on demonstrating that regularities in different inputs are rapidly perceived and learned by newborns (e.g., Bulf, Johnson, &

Correspondence should be sent to Noam Siegelman, Haskins Laboratories, 300 George St. #900, New Haven, CT 06511. E-mail: noam.siegelman@yale.edu

Valenza, 2011), children (e.g., Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), and adults (e.g., Saffran, Newport, & Aslin, 1996), as well as non-human species (e.g., Toro & Trobalón, 2005). What remains controversial, however, is the precise nature of the computations underlying SL (e.g., Siegelman, Bogaerts, Armstrong, & Frost, 2019; and see Frost, Armstrong, & Christiansen, 2019, for extensive discussion).

In a typical SL experiment, participants are presented with a continuous stream of elements. This stream consists of repeating embedded patterns, which can also be described as an array of transitional probabilities (TPs) between the basic elements in the stream. Research on SL has unequivocally demonstrated that humans are sensitive to the statistical properties of the input, which allows them to recognize the patterns embedded in the stream in a subsequent test phase, in the visual (e.g., Kirkham, Slemmer, & Johnson, 2002), auditory (e.g., Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), or tactile (Conway & Christiansen, 2005) modality.

Statistical learning performance, however, cannot be explained by considering only the statistical properties of the input. Consider, for example, the visual statistical learning (VSL) task, in which the input stream consists of abstract shapes (e.g., Turk-Browne, Junge, & Scholl, 2005), fractal stimuli (e.g., Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014), or cartoon aliens (e.g., Arciuli & Simpson, 2012). In order to learn the pattern of co-occurrences between the stream's elements, participants must (a) acquire robust representations of these elements, and (b) compute their pattern of co-occurrences. This approach served as the basis for a recent account of SL performance (Frost et al., 2015). In this schematic model, SL outcomes are the product of the interplay between encoding-related (and modality-specific) mechanisms that form representations from the input, and higher-order learning-related (and domain-general) mechanisms that compute the transitional statistics of the encoded representations. Critically, this model naturally assumes some form of temporal modularity; first, elements in the stream must be encoded into internal representations, and then the co-occurrences between these internal representations can be computed and registered. Moreover, the model argues for neuro-biological separability—with encoding of stimuli occurring in lower cortical regions (visual, auditory and somatosensory cortices) that feed into statistical computations in higher cortical regions such as the medial temporal lobe (MTL) system (and see Schapiro et al., 2014; Schapiro, Turk-Browne, Botvinick, & Norman, 2017, for discussion). The model also outlines implications for understanding individual differences in SL. It suggests that these may arise either from variance in efficacy of encoding stimuli in a given modality, or from variance in the efficacy in computing their TPs.

This view, however, was challenged by a recent experimental work that aimed to test the model, focusing on the demarcation line between encoding and learning (Bogaerts, Siegelman, & Frost, 2016). The study used a within-subject design with an independent manipulation of exposure duration (ED) of each shape in the stream (i.e., the amount of time that the stimulus is physically available for processing, a parameter affecting the efficacy of encoding the individual shapes), and the extent of the TPs in the stream (a parameter related to the registration of the co-occurrences between shapes). In contrast to the model proposed by Frost et al. (2015), which considered these two processes to be

independent, temporally modular, and therefore, additive, a substantial interaction between ED and TPs was found, so that the impact of ED was modulated by TPs and vice versa. This interaction suggested that the ability to encode the individual shapes and the sensitivity to their co-occurrences are not independent processes, but rather that the statistical properties of the stream may facilitate encoding, and conversely, that optimal conditions of encoding can serve to enhance sensitivity to the statistical structure of the input. The interaction of ED and TPs was further demonstrated by tracking individual differences in performance. Participants who showed greater sensitivity to changes in ED tended to show greater sensitivity to changes in TPs, and vice versa. These sets of findings are consistent with other studies demonstrating the interdependence of perceptual processes and regularity learning. For example, Barakat, Seitz, and Shams (2013) reported that the presence of statistical regularities enhances the detection of individual visual elements even when they appear outside the context of the learned regularities. In a similar vein, Grieco-Calub, Simeon, Snyder, and Lew-Williams (2017) showed that auditory SL performance is impaired when the familiarization stream comprises spectrally degraded speech sounds. From a broader perspective, the interaction between encoding and learning aligns with the well-documented benefits of context on stimulus encoding, such as the classic effect of schema congruency on subsequent memory of new events (Hintzman, 1986).

However, whereas the account provided in Bogaerts et al. (2016) adequately describes the data, it does not truly explain it. Finding that two processes interact does not tell us much about what drives this interaction and what computations underlie it. Importantly, it does not elucidate the nature of individual differences in visual SL abilities. How should these be explained and predicted? Why is sensitivity to rate of presentation correlated with sensitivity to extent of stimuli predictability? The present paper targets, therefore, the intriguing interaction between encoding of elements in vision and computing their transitional statistics. Our following experimental work aimed to offer a novel computational principle for understanding visual SL, building upon the principles of Information Theory. To preview our findings, we suggest that one unifying computational principle can account for SL abilities, jointly explaining the effects of encoding- and learning-related factors, as well as their interaction. Note, that although we target here the learning of abstract visual shapes as a test case, our computational principle can be extended to a range of learning situations across cognition.

1.1. A unifying principle—Rate of information

In the current work, we adopt an information theoretic approach to SL. We therefore aim to quantify the input to which learners are exposed in terms of the available information it carries. In information theory, the basic measure for the amount of the available information in an input is *entropy*. The basic idea behind entropy is that when the input is more random, events in the stream are less predictable, and therefore, each event carries more information by itself (also referred to as higher surprisal). In contrast, if events in an input stream are highly predictable, each event on average carries little information

(or has lower surprisal) because it can be predicted from other events. Operationally, in the context of SL, the average amount of information found in each element, given the element that preceded it, can be quantified using Markov entropy (measured in bits/elements; see, for example, Nastase, Iacovella, & Hasson, 2014)[1]:

$$\text{Markov entropy} = -\sum_{i=1}^{n} p(i) \sum_{j=1}^{n} p(j|i) \times \log p(j|i).$$

This formula quantifies the average extent of new information provided by each element in the stream, by summarizing the average predictability of each element, weighted by its overall frequency.[2] To illustrate, consider an input stream with four elements: A, B, C, and D. In an input stream in which these four elements appear randomly one after the other (e.g., D B A C C B C A D B, etc.), all elements in the stream are unpredictable, and therefore each event provides maximal new information, leading to high entropy. In contrast, if the same four elements are organized in a stream consisting only of the pairs AB and CD (e.g., A B A B C D A B C D, etc.), the second element in each pair does not carry any new information, because it is fully predictable by the element preceding it, leading to low entropy. Indeed, calculating the Markov entropy for these two streams reveals higher entropy for the random stream (2 bits/element) versus the structured stream (0.5 bits/element).

In the context of SL, the higher the TPs between elements within-patterns (i.e., a stream that consists of more regular patterns), the lower the entropy value. Importantly, however, while Markov entropy and TP are correlated, they are far from being equivalent. From the inception of the SL literature, TP was used as a way of summarizing the degree of structure in the presented input and was calculated simply by averaging the probabilities of transitions within patterns (e.g., Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996). This has led later experimental work to manipulate the degree of structure in the input by changing this parameter (see, e.g., Bogaerts et al., 2016; Hay, Pelucchi, Estes, & Saffran, 2011). In contrast to TP, Markov entropy is calculated *on the full matrix of transitions of all elements in the stream* (all transitions i, j), including both between- and within-pattern transitions. Thus, Markov entropy is impacted by all parameters that affect the full matrix of transitions between elements (e.g., within-pattern TPs, the ways in which noise is introduced to the input, the overall number of patterns, the length of patterns, etc.), and thus quantifies the average amount of information that is provided by each element. Note, however, that in the present design, changes of entropy values resulted from manipulation of TPs alone, and not via other parameters.

However, looking only at the entropy level, the average extent of new information provided by each element is not enough to predict learning outcomes. Another important factor that affects information processing is the *time* available to process each element. Therefore, a unifying computational principle for SL performance must also take into account the time dimension. In the context of SL paradigms, a straightforward measure for the time dimension is *rate of presentation*—the number of elements that appear in

each second. A faster stream of elements (with higher rate of presentation) will be more difficult to process as there is less time to process the extent of novel information carried by each element. In contrast, a slower stream of information will be easier to process, therefore leading to better learning outcomes.

A single computational principle, which takes into account both the entropy and time dimensions, is, therefore, *Rate of Information* (RI). As shown in the formula below, RI is simply the multiplication of Markov entropy (average extent of new information per element, measured in bits/element) by the rate of presentation (the time available to process each element, measured in elements/second).

$$\text{Rate of Information}\left(\frac{\text{bits}}{\text{second}}\right) = \text{entropy}\left(\frac{\text{bits}}{\text{element}}\right) \times \text{rate of presentation}\left(\frac{\text{element}}{\text{second}}\right). \quad (1)$$

Since RI is the multiplication of bits/element by element/second, the units of this measure are *bits/second*. Intuitively, therefore, RI quantifies the *extent of new information per second* to be processed (or, in more formal terms, the amount of bits of new information to be processed per second). In the context of SL, *it measures how much new information is available in each second of the familiarization stream*. Importantly, note that RI takes into account both the average extent of predictability of elements as well as the time available to process each unit given its predictability. Therefore, it could potentially explain the effects of both encoding-related and learning-related factors, as well as their interaction. Specifically, here we propose that RI could account for both the effect of TPs between elements (with higher TPs leading to lower entropy, therefore lower RI, and vice versa), the effect of ED (with longer ED for each element leading to slower rate of presentation—less shapes per second, therefore lower RI, and vice versa), and their interaction.

How would we know whether it is indeed RI that drives VSL performance? The RI hypothesis generates two straightforward predictions. First, if indeed RI drives SL performance, then a clear monotonic function, where higher levels of RI lead to worse SL performance, should be found (as documented in other domains in the Cognitive Sciences, e.g., de Fleurian, Blackwell, Ben-Tal, & Müllensiefen, 2017; Fan, 2014; Miller, 1956). Whereas, in principle, the shape of this function is not unequivocally clear a priori, in the context of VSL performance, a logarithmic function seems the most probable: At lower levels of RI (i.e., streams with lower amounts of information to be extracted) we expect a substantial impact of changes in the amount of information on SL performance, which is expected to decrease at higher levels of RI as SL performance gets closer to chance level. We therefore hypothesize that SL performance would be predicted from log-transformed RI. A second even stronger prediction is that similar SL performance should be observed in learning conditions with identical RI levels regardless of the specific task parameters. In the context of the VSL paradigm, it is thus predicted that even if two learning conditions have different combinations of TPs and ED (i.e., a condition with high ED and low TP vs. a condition with low ED and high TP), a similar SL performance should be observed as long as their RI level is identical.

In light of these two predictions, the current work employed two experimental strategies. In order to test the first prediction—whether RI indeed monotonically predicts SL performance—we examined VSL performance across a large number of conditions, with different RI levels (through a re-analysis of Bogaerts et al., and Experiment 1). This allowed us to test whether there is a logarithmic relation between SL performance and RI. In order to examine the second prediction—similar VSL performance for conditions with similar RI regardless of specific ED and TPs parameters—we compared performance in a series of paired conditions with identical RI, but different combinations of ED and TPs (Experiment 2).

1.2. Preliminary evidence—Re-analysis of Bogaerts et al. (2016)

Before turning to the new experimental work, in the current section we examine some preliminary evidence for the RI hypothesis, through a reanalysis of the data from Bogaerts et al. (2016). The data are based on a sample of 42 subjects, each participating in nine VSL conditions. In each of these conditions, the VSL task included eight pairs of shapes. Using a within-subject factorial design, Bogaerts et al. (2016) found significant effects of ED (a factor with three levels: 200, 600, 1,000 ms per shape), within-pattern TPs (0.6, 0.8, 1), and their interaction on VSL performance. Here we re-examined these results, now asking whether VSL performance is indeed predicted not from ED and TPs (and their interaction), but rather from a single factor: RI. We therefore calculated the RI level in each of the nine conditions employed by Bogaerts et al. (2016; see Table 1), then, we examined the relation between VSL performance and RI across the nine conditions. As shown in Fig. 1, we observed a very high logarithmic fit between VSL and RI ($R^2 = 0.84$). This preliminary analysis already supports, to some extent, the first prediction regarding RI, showing that it indeed accounts for performance in a VSL task. Moreover, it clearly shows that this possible relation is characterized by a logarithmic function. This is further supported by the fact that a linear fit between VSL and RI was substantially lower, $R^2 = 0.66$.

Table 1
Rate of information levels for the 9 conditions from Bogaerts et al. (2016)

Condition Parameters	Rate of Information (bits/sec)	SL Performance (% correct)
ED = 1,000, TP = 1	1.50	72.8
ED = 1,000, TP = 0.8	2.22	65.6
ED = 600, TP = 1	2.50	67.6
ED = 1,000, TP = 0.6	2.71	61.8
ED = 600, TP = 0.8	3.68	59.7
ED = 600, TP = 0.6	4.53	59.5
ED = 200, TP = 1	7.50	58.4
ED = 200, TP = 0.8	11.08	56.3
ED = 200, TP = 0.6	13.53	54.8

Note. ED, exposure duration; SL, statistical learning; TP, transitional probabilities.

Admittedly, however, this reanalysis cannot be taken to provide anything more than preliminary support, due to some limitations. First, it was based on a small number of experimental conditions—only nine overall. Second, RI levels in these conditions were not well distributed over a wide range of possible values, as there were many conditions with a small amount of information per second (6/9 conditions had a RI of 4.53 or less). The goal of Experiment 1 was therefore to examine whether a logarithmic function between VSL scores and RI is indeed obtained, over a larger number of experimental conditions, spanning a wide range of RI values.

2. Experiment 1

In Experiment 1, we examined VSL performance across 24 RI levels, created through different combinations of ED and TPs. We predicted that RI would account for VSL performance, and that it would do so better compared to other possible factors. Therefore, in operational terms, we predicted that a clear logarithmic function between VSL performance and RI would (again) be observed across the whole range of RI levels, and, importantly, that a regression model with RI as a single predictor would best fit the data, compared to models with ED, TPs, and their interaction as predictors.

2.1. Methods

2.1.1. Participants

A total of 217 students of the Hebrew University (43 males) participated in the study for payment or course credit. Participants had a mean age of 22.8 years (range: 18–35)

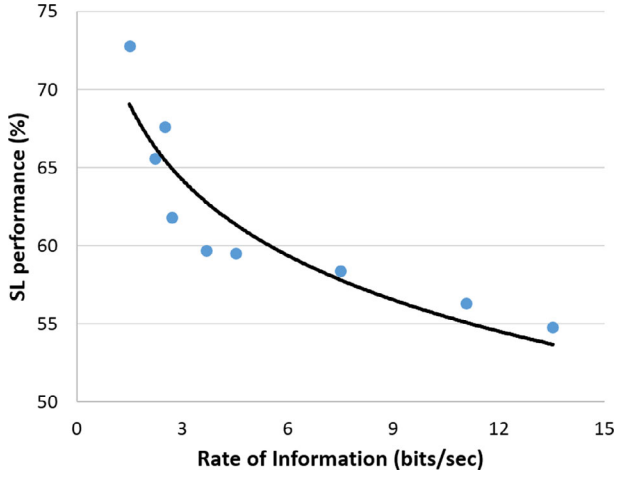

Fig. 1. Statistical learning (SL) performance by rate of information (RI) (bits/sec) in the nine conditions from Bogaerts et al. (2016). Line shows best logarithmic fit.

and reported no history of learning or reading disabilities, ADD, or ADHD. They were randomly assigned into one of the 24 VSL conditions (resulting in a sample of 8–12 subjects in each condition).

2.1.2. Design, materials, and procedure

In designing the experiment, we calculated the RI levels of 72 possible VSL conditions, resulting from all possible combinations of nine possible ED values (from 200 ms to 1,000 ms in intervals of 100 ms, i.e. 200 ms, 300 ms, 400 ms, …, 1,000 ms), and eight TPs values (from 0.3 to 1 in intervals of 0.1, that is, 0.3, 0.4, 0.5, … 1). The RI levels of these possible conditions are presented in Table 2. From these 72 possible combinations, we selected 24 conditions (highlighted cells in Table 2), by selecting 2–3 conditions from each value of ED and TPs (i.e., 2–3 conditions from each row/column in Table 2), so as to evenly cover a wide range of possible RI levels.

All other task parameters were similar to that of Bogaerts et al. (2016). The VSL task in all conditions employed the same 22 shapes (e.g., Turk-Browne et al., 2005; see Appendix A). For each subject, 16 of the 22 shapes were randomly chosen and organized to create eight pairs (the remaining 6 shapes were used for the "lie detection" trials, see below). Similar to typical SL tasks, the first part of the VSL was a familiarization phase, in which the eight pairs were presented continuously, one after the other, in a random order. Each pair appeared 24 times during familiarization, with the only constraint that the same pair could not appear twice in a row. In all conditions, there was a 100 ms interval between shapes, both within- and between-pairs. The manipulation of TPs was done by including random noise in conditions with TPs smaller than 1: for example, in a condition with TPs = 0.8, for each pair AB during familiarization, shape B appeared after shape A 80% of the time, while in 20% of the time shape B was randomly replaced by another shape X, while avoiding immediate repetition of shapes. The familiarization stream lasted between 2 and 7 min (depending on the ED condition). Prior to familiarization, participants were instructed to carefully attend the stream as they would later be tested, but they were not told that the stream consisted of pairs or any form of sequential regularities.

Table 2
Rate of information levels for 72 possible combinations of ED and TPs. Highlighted cells are the conditions selected for Experiment 1

		Exposure Duration (ED)								
		200	300	400	500	600	700	800	900	1,000
Transitional probability (TPs)	**0.3**	16.22	10.82	8.11	6.49	5.41	4.64	4.06	3.60	3.25
	0.4	15.46	10.3	7.72	6.19	5.16	4.42	3.86	3.43	3.09
	0.5	14.57	9.70	7.29	5.82	4.85	4.16	3.64	3.23	2.91
	0.6	13.53	9.01	6.76	5.41	4.53	3.87	3.39	3.01	2.71
	0.7	12.42	8.26	6.20	4.97	4.13	3.54	3.09	2.75	2.47
	0.8	11.08	7.39	5.53	4.42	3.68	3.16	2.77	2.46	2.22
	0.9	9.54	6.34	4.78	3.82	3.17	2.73	2.39	2.12	1.91
	1.0	7.50	5.00	3.75	3.00	2.5	2.14	1.87	1.67	1.50

Following the familiarization phase, a 38 2-Alternative Forced Choice (2-AFC) test phase started. Thirty-two of the test-trials examined SL abilities, by contrasting (a) "true pairs"—two shapes that appeared as a pair during the familiarization phase, and (b) "foils"—two shapes that did not appear as a pair during familiarization. Foils were constructed without violating the position of the shapes within the original pairs (e.g., for two true pairs AB and CD, possible foils could be AD or CB, but not AC or DB). Scores in the SL task ranged then from 0 to 32, according to the number of correct identifications of pairs during the test phase. The remaining 6 test-trials were "lie detection" trials, examining whether subjects attended the familiarization stream. These trials contrasted "true pairs" with a pair containing a novel shape that did not appear at all during familiarization (see Bogaerts et al., 2016; Romberg & Saffran, 2013, for a similar procedure).

2.1.3. General procedure

Participants completed the VSL task from their home computers, through an online platform. Participants were instructed to do the task alone in a quiet room, to avoid external distractions (turn-off their cell phone, and turn off music) and to have only the experiment window open on their computer.

2.2. Results and discussion

Performance on the lie detection trials was high (86.6%, significantly different than chance-level of 50%, $t(216) = 29.15$, $p < .001$), showing that subjects (at the group level) indeed kept their attention on the familiarization stream. Note that, in contrast to previous studies (e.g., Bogaerts et al., 2016; Romberg & Saffran, 2013), we (a priori) opted not to use performance on lie detection trials as a method to screen participants at the individual level, due to the small number of trials (only 6 per subject), which might lead to a high error rate (i.e., screening subjects who actually did attend to the stream).

Across the 24 conditions, performance on the SL trials was significantly different than chance (56%, $t(216) = 6.89$, $p < .001$). Fig. 2 shows VSL performance as a function of RI level. As can be seen, the results of Experiment 1 replicate the logarithmic function between RI and VSL performance from the preliminary analysis based on the results of Bogaerts et al. (2016). Admittedly, Fig. 2 presents a noisier picture compared to Fig. 1 above (across the 24 conditions $R^2 = 0.35$, substantially lower than the fit across the 9 conditions in Fig. 1 where $R^2 = 0.84$). This additional noise is most likely due to the smaller number of subjects represented by each data point (n varied between 8 and 12 for each data point in Fig. 2, compared to $n = 42$ in Fig. 1) and the between-subject design in Experiment 1 (as opposed to the within-subject design in Bogaerts et al., 2016).

Next, we turned to examine whether indeed (log-transformed) RI outperforms other possible predictors in accounting for VSL performance. To do so, we ran a series of logistic mixed effect models (Jaeger, 2008). All models had correctness in each SL

trial as the dependent variable, the trial number (1–32) and the position of the true-pair within the 2-AFC trial (first or second) as control variables, and a by-subject random intercept. Each of the models had different fixed effects, representing the different possible predictors of SL performance (see Table 3). We predicted that the model with only log-transformed RI would best account for the data. We compared the models by looking at Bayesian Inference Criteria (BIC) values, which are a standard way of model comparison in mixed effect models (e.g., Müller, Scealy, & Welsh, 2013). A lower BIC represents a better model, with the natural exponent powered by half the difference between the BIC of two models to be interpreted similarly to a Bayes Factor (i.e., $e^{\Delta BIC/2} \approx$ Bayes Factor; e.g., Masson, 2011). Thus, for example, a $\Delta BIC = 5$ between model A and model B means that model A account for the data $e^{5/2} = 12.18$ times better than model B.

As predicted, the logarithmic model with only log-transformed RI level (top row of Table 3) best fit the data, as reflected by its lowest BIC value compared to all other models. The minimum difference between this model and any other model without Rate Information as a predictor (i.e., models 3–5 in Table 3) was $\Delta BIC = 8$, meaning that (log-transformed) RI accounted for the data $e^{8/2} = 54.5$ times better compared to any of these models. Note also that the logarithmic transformation of RI improved model fit by $\Delta BIC = 5$ (see model 1 vs. model 2 in Table 3), showing strong evidence for a logarithmic (rather than a linear) relation between VSL and RI. Taken together, the results of Experiment 1 show that (log-transformed) RI can indeed account for VSL performance, and that it does so better than other predictors. Importantly, RI had better explanatory power than a model with separate predictors for ED, TPs, and their interaction. This suggests that the main effects for ED and TPs, as well as their interactive

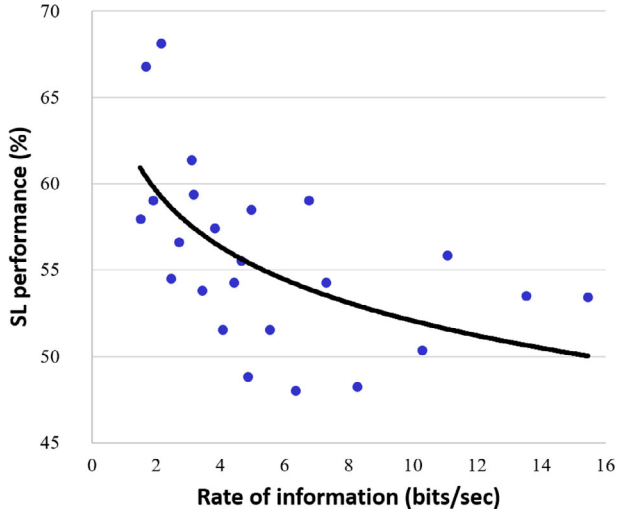

Fig. 2. Statistical learning (SL) performance by rate of information (RI) (bits/sec) in the 24 conditions of Experiment 1. Line shows best logarithmic fit.

Table 3
Model comparison results for Experiment 1

Model	BIC
1) **VSL ~ log (Rate of Information)**	**9,460**
2) VSL ~ Rate of Information	9,465
3) VSL ~ ED + TPs + ED:TPs	9,475
4) VSL ~ ED:TPs	9,469
5) VSL ~ ED + TPs	9,468
6) VSL ~ ED + TPs + log (Rate of Information)	9,475
7) VSL ~ ED + log(TPs) + ED:log(TPs)	9,476

Note. ED, exposure duration; TP, transitional probabilities; VSL, visual statistical learning.

effect, can all be explained by a single computational principle—RI. Theoretically, this suggests that both encoding-related and learning-related effects (as well as their interaction) operate along a single axis—the amount of information per second in the input stream.

An even stronger test of the RI account can be achieved by comparing the model with log-transformed RI (model 1 in Table 3 above) to an alternative model that instead considers separately the mathematical constituents of RI: rate of presentation and Markov entropy. Mathematically, the RI model, SL ~ a + b*log(RI), can be expressed as SL ~ a + b*log(entropy*presentation_rate), which is equivalent to: SL ~ a + b*log(entropy) + b*log(presentation_rate). Critically, in this specification of the model, the coefficient of entropy and rate of presentation are equal. The question is, therefore, whether this model better predicts the data compared to a model in which the two coefficients are not constrained to be equal (i.e. SL ~ a + $\underline{\mathbf{b_1}}$log(entropy) + $\underline{\mathbf{b_2}}$log(presentation_rate)). Examining this unconstrained model revealed that it had a BIC value of 9463, indeed higher than that of the log(RI) model where BIC = 9460, meaning that the unified log-RI accounted for the data $e^{3/2} = 4.48$ times better than the unconstrained model. This provides additional evidence that RI, as a unified construct, accounts best for the data.

3. Experiment 2

Our aim in Experiment 2 was to extend the findings of Experiment 1 and show that not only RI determines VSL performance in a logarithmic manner, but importantly, that similar RI levels lead to similar VSL performance, regardless of ED and TPs values. Our experimental strategy was to employ a within-subject design with different pairs of ED and TPs combinations that have identical RI levels. In two sub-experiments (Experiments 2a and 2b), each participant thus performed two VSL tasks with the same RI level: one with fast stimulus presentation (low ED) and a high degree of regularity (high TP), and the other with slow stimulus presentation (high ED) and a low TP. We examined whether these led to similar performance.

4. Experiment 2a

4.1. Methods

4.1.1. Participants

An additional sample of 250 Hebrew University students was recruited to participate in Experiment 2 for course credit or payment. Participants were randomly assigned to one of five experimental conditions (n = 50 in each condition). From this initial sample, 243 participants (56 males, n = 47–50 in each condition) completed their participation in the study (i.e., participated in both experimental sessions; see below), and their data are used for all subsequent analyses. Participants' mean age was 22.89 years (range: 18–35), and they reported no history of reading or learning disabilities, ADD, or ADHD.

4.1.2. Design, materials, and procedure

Each experimental condition in Experiment 2 consisted of a pair of VSL tasks with similar RI level but different combinations of ED and TPs. To select the five pairs, we again examined all 72 possible combinations of 9 ED levels (200–1,000 ms in intervals of 100) and 8 TP levels (0.4–1 in intervals of 0.1). From these possible conditions, we then selected five pairs, each consisting of two ED-TP combinations with similar RI values (see Table 4). Importantly, in each pair, one condition had a combination of lower value of ED and higher value of TPs (labeled *low ED/high TP* condition), while the other had a combination of a higher ED and lower TPs (labeled *high ED/low TP*).

Experimental parameters in the VSL tasks were identical to Experiment 1. Each participant performed *both* conditions within each pair condition, at least 7 days apart (mean: 14.24 days; range: 7–50 days). As in Experiment 1, participants took the tasks from home, through an online platform, and were requested to avoid any external distractions.

4.2. Results and discussion

Performance on the lie detection trials was high—84.1%. This value significantly differed from chance ($t(242)$ = 30.99, p < .001), showing that again participants kept their

Table 4
Highlighted cells show the conditions selected for Experiment 2a. Each color represents one condition pair

		Exposure Duration (ED)								
		200	300	400	500	600	700	800	900	1.000
Transitional probability (TPs)	**0.3**	16.22	10.82	8.11	6.49	5.41	4.64	4.06	3.60	3.25
	0.4	15.46	10.3	7.72	6.19	5.16	4.42	3.86	3.43	3.09
	0.5	14.57	9.70	7.29	5.82	4.85	4.16	3.64	3.23	2.91
	0.6	13.53	9.01	6.76	5.41	4.53	3.87	3.39	3.01	2.71
	0.7	12.42	8.26	6.20	4.97	4.13	3.54	3.09	2.75	2.47
	0.8	11.08	7.39	5.53	4.42	3.68	3.16	2.77	2.46	2.22
	0.9	9.54	6.34	4.78	3.82	3.17	2.73	2.39	2.12	1.91
	1.0	7.50	5.00	3.75	3.00	2.5	2.14	1.87	1.67	1.50

attention to the familiarization stream. Overall SL performance (across conditions) also differed from chance level (56.2%, $t(242) = 9.67$, $p < .001$), reflecting successful SL at the group level.

Fig. 3 (left panel) presents VSL performance in each of the five condition pairs as a function of RI. As can be seen, the pattern of results across RI levels replicates those of Experiment 1, with an observed logarithmic relation between VSL performance and RI. More important, we then examined VSL performance within each condition pair, to assess whether it was similar across the *low ED/high TP* and the *high ED/low TP* combinations. In line with our predictions, a series of paired *t* tests showed no significant difference between the two combinations across all five condition pairs (from lowest to highest RI levels: pair 1: $t(47) = 0.15$, $p = 0.88$; pair 2: $t(49) = 1.46$, $p = .15$; pair 3: $t(48) = 1.81$, $p = .08$; pair 4: $t(46) = 0.65$, $p = 0.52$; pair 5: $t(46) = 0.63$, $p = .53$). Note, however, that the use of t-tests in this context is limited. Our a-priori prediction in this experiment was that a similar VSL performance should be observed in each of the five condition pairs—that is, we predicted a *lack* of an effect. *t* tests, as all other null-hypothesis-significance-testing (NHST) procedures, cannot provide support for the null hypothesis (i.e., no difference between combinations), but only for the alternative (some difference between combinations; see, e.g., Dienes, 2014). We therefore next calculated, for each condition pair, a Bayes Factor (BF), quantifying the ratio between the likelihood of the data under the alternative hypothesis and the likelihood of the data under the null (i.e., $BF = \frac{p(D|H_1)}{p(D|H_0)}$). We predicted that across the five pairs, support for the null hypothesis would be observed, as reflected by BFs smaller than 1. We used "Bayesian paired *t* tests" (Rouder, Speckman, Sun, Morey, & Iverson, 2009) using the BayesFactor package in R

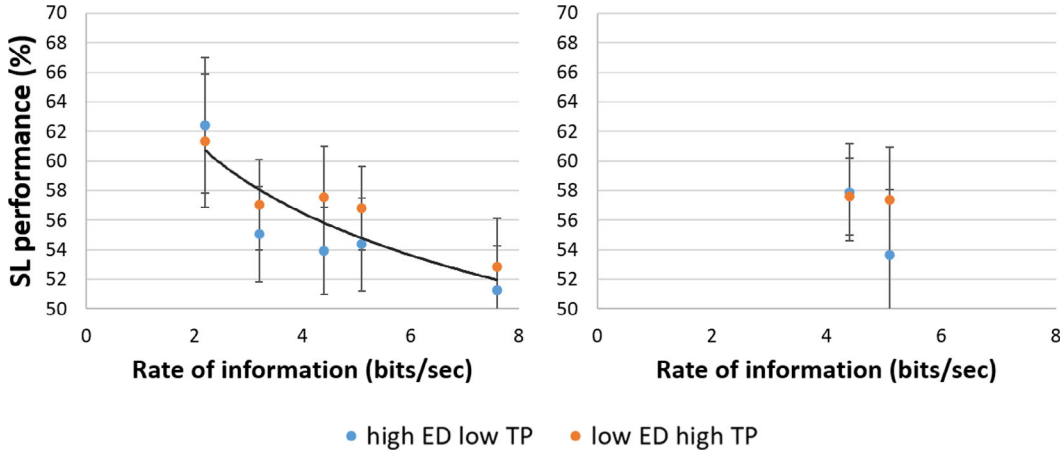

Fig. 3. Statistical learning (SL) performance by rate of information (RI) (bits/sec) across condition pairs in Experiment 2a. Each pair of conditions is represented by two points in the same RI level: Blue points represent high ED/low TP combinations, whereas orange points represent low ED/high TP combinations. Left: Results of the original five condition pairs. Right: Results of the follow-up study: replication of pairs 3 and 4. Error bars depict 95% confidence intervals.

(Morey, Rouder, & Jamil, 2014), with default priors (i.e., Cauchy distribution with a width of $\sqrt{2}/2$ on effect size). As predicted, in all five pairs, a BF smaller than 1 was observed, reflecting support for the null hypothesis (from lowest to highest RI levels: pair 1: BF = 0.19; pair 2: BF = 0.19; pair 3: BF = 0.7; pair 4: BF = 0.41; pair 5: BF = 0.16). Similarly, a combined analysis of all data across all five pairs revealed substantial support the null, meaning, as predicted, no difference in VSL performance between high ED/low TP and high ED/low TP combinations across RI levels (BF = 0.2).

Note, however, that the strength of evidence for the null was somewhat weaker in two out of the five condition pairs with medium RI levels. Specifically, in pairs 3 and 4, the null hypothesis predicted the data only 1.42 or 2.44 times better than the alternative, respectively, values that are typically considered as "weak" evidence for the null over the alternative (e.g., Wetzels et al., 2011). We therefore conducted a follow-up study in which we re-ran these two condition pairs with a new sample of 100 Hebrew University students. Eighty-nine of them (20 males, M_{age} = 22.4, age range: 19–30) completed their participation (n = 42 in pair 3, n = 47 in pair 4, mean interval between tests 11.3 days with range: 7–36 days). Performance on the lie detection trials for this sample was 86.9% (significantly different from chance: $t(88) = 23.01$, $p < .001$) and mean SL performance again differed from chance level 56.6%, $t(88) = 7.25$, $p < .001$). The right panel of Fig. 3 presents VSL performance in the two replicated conditions as a function of RI. Again, no significant difference was observed between high ED/low TP vs. low ED/high TP combinations in the two pairs (pair 3: $t(41) = 0.12$, $p = .9$; pair 4: $t(46) = 1.7$, $p = .1$). Bayes Factor analysis also provided support for the null hypothesis in both conditions, although the extent of support again differed between the two pairs: pair 3: BF = 0.17; pair 4: BF = 0.60. Given the variance in findings between pairs in Experiment 2b, we aimed to further estimate the actual effect of ED and TP combinations across a wider range of RI levels.

5. Experiment 2b

Experiment 2b again used conditions with identical RI levels, but different combinations of ED and TP. The experiment had a total of five condition pairs: three "replication pairs" from Experiment 2a (which were not already replicated in the follow-up study) and two "new pairs," with different RI values. The goal of Experiment 2b was therefore to collect further data, both replicating and extending Experiment 2a.

5.1. Methods

5.1.1. Participants

Since we exhausted most of the Hebrew University student pool in Experiments 1 and 2a, in Experiment 2b we turned to a different platform—Amazon Mechanical Turk. This also allowed us also to extend the validity of our findings to a wider and heterogeneous

sample of participants. We first recruited 500 participants through Mechanical Turk, who were randomly assigned into one out of the five condition pairs (n = 100 in each condition), and participated in one (random) combination of ED and TP within this pair. After 7 days, we recontacted all participants via e-mail and offered them the chance to participate again in the experiment, sending them a link to the other ED/TP combination of the condition pair to which they were assigned. A total of 279 participants out of the original 500 agreed to participate also in the second session (between 43 and 66 in each condition pair)—their data were used in all analyses presented below.

5.1.2. Design, materials, and procedure

The five condition pairs in Experiment 2b are presented in Table 5. As noted above, the three "replication pairs" conditions were pairs from Experiment 2a (those that were not already replicated in the follow-up study): pair 1 from Experiment 2a, RI ≅ 2.2; pair 2, RI ≅ 3.2; and pair 5, RI ≅ 7.6. In addition, two new pairs were selected (RI ≅ 3.65,RI ≅ 6.3), to cover more of the RI range. All other task parameters were identical to Experiment 2a. The mean time interval between sessions within each condition pair was 10.9 days (range: 7–20 days).

5.2. Results and discussion

As in previous Experiments, performance on the lie detection trials was high—72.8% (significantly different from chance: $t(278) = 19.88$, $p < .001$) showing that participants kept their attention to the familiarization stream. Admittedly, lie detection performance was somewhat lower than the observed values in Experiments 1 and 2a (in which lie detection performance was around 85%). This is probably due to the difference in participant pool (Hebrew University students in Experiments 1 and 2a vs. Mechanical Turk workers in Experiment 2b). Nonetheless, overall SL performance (across conditions) significantly differed from chance level (55.17%, $t(278) = 9.44$, $p < .001$), reflecting successful SL at the group-level. Fig. 4 presents VSL performance in the five condition pairs

Table 5
Condition pairs in Experiment 2c. Each color represents one condition pair. Yellow, orange, and purple pairs are replications of pairs 1, 2, and 5 in Experiment 2a, respectively (RI ≅ 2.2, RI ≅ 3.2, and RI ≅ 7.6). The two other pairs are new conditions

		Exposure Duration (ED)								
		200	300	400	500	600	700	800	900	1,000
Transitional probability (TPs)	**0.3**	16.22	10.82	8.11	6.49	5.41	4.64	4.06	3.60	3.25
	0.4	15.46	10.3	7.72	6.19	5.16	4.42	3.86	3.43	3.09
	0.5	14.57	9.70	7.29	5.82	4.85	4.16	3.64	3.23	2.91
	0.6	13.53	9.01	6.76	5.41	4.53	3.87	3.39	3.01	2.71
	0.7	12.42	8.26	6.20	4.97	4.13	3.54	3.09	2.75	2.47
	0.8	11.08	7.39	5.53	4.42	3.68	3.16	2.77	2.46	2.22
	0.9	9.54	6.34	4.78	3.82	3.17	2.73	2.39	2.12	1.91
	1.0	7.50	5.00	3.75	3.00	2.5	2.14	1.87	1.67	1.50

as a function of RI levels. Below we describe these results, discussing separately the replication pairs, and the new pairs.

5.2.1. VSL performance: Replication pairs

We start by examining VSL performance in the "replication pairs" from Experiment 2a: pairs 1 (RI ≅ 2.2), 2 (RI ≅ 3.2), and 5 (RI ≅ 7.6). As in Experiment 2a, no significant difference was observed between high ED/low TP and low ED/high TP combinations in the three pairs, and Bayes Factor analysis revealed some support for the null hypothesis (BF < 1) in all three pairs (pair 1: $t(42) = 1.88$, $p = .07$, BF = 0.82; pair 2: $t(51) = 1.36$, $p = 0.18$, BF = 0.36; pair 5: $t(58) = 0.52$, $p = .61$; BF = 0.16).

5.2.2. VSL performance: New pairs

In line with previous experiment, VSL performance in the two new pairs (new-pair 1: RI ≅ 3.65, new-pair 2: RI ≅ 6.3) showed no significant difference between high ED/low TP and low ED/high TP combinations, and the Bayes Factor analysis showed some support for H0 in both condition pairs (new-pair 1: $t(58) = 1.73$, $p = .08$, BF = 0.58; new-pair 2: $t(65) = 0.12$, $p = .91$; BF = 0.14).

Taken together, the results of Experiment 2a and 2b suggest that indeed RI provides an overall good predictor of VSL performance. Across the three sub-experiments, on a total 12 sub-samples (all $n > 40$), not one statistically significant difference was found between VSL conditions with identical RI level, having different combinations of ED and TPs. Bayes Factor analyses complemented this finding: When looking at the experimental conditions separately, Bayes Factors for all 12 RI pairs had a BF < 1 value, signaling some

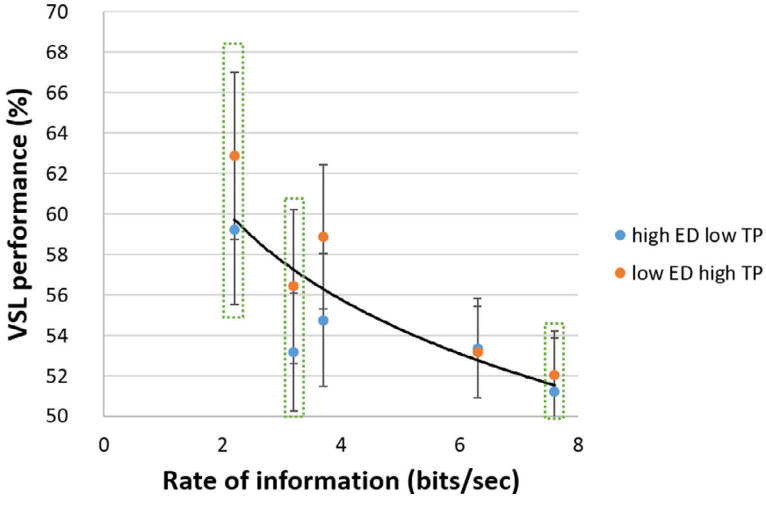

Fig. 4. Visual statistical learning (VSL) performance by rate of information (RI) (bits/sec) in the five condition pairs of Experiment 2b. Blue points represent high ED/low TP combinations, whereas orange points represent low ED/high TP combinations. Replication pairs are marked with dashed rectangles. Error bars depict 95% confidence intervals.

support for the null hypothesis. Moreover, in 6 out of the 12 pairs, we found that BF < 1/3, signaling substantial support for the null. Our results thus conform with the predictions of RI determining VSL performance. They suggest that similar VSL performance is obtained within the same rate of information, regardless of ED and TP combinations.

5.3. Aggregated analysis

Although our results show that RI does a good job at predicting VSL performance, some data points we obtained seem to be discrepant with the RI account. Specifically, some of the pairs comparisons resulted in BFs that are typically considered "inconclusive" (i.e., BF larger than 1/3; e.g., Wetzels et al., 2011). This raises the question of whether such variance in findings simply reflects noise, or whether it points to some more fundamental bias of RI as a predictor of VSL performance. In light of the multiple comparisons throughout Experiment 2, and the error measurements characteristics to the 2-AFC test (see Siegelman, Bogaerts, & Frost, 2017 for discussion), it is possible that the discrepant data points are due to noisy performance measures. However, when considering the combined results of Experiments 2a and 2b, it seems that the lowED/highTP combination had a slight advantage in VSL performance compared to the highED/lowTP combination (see Figs. 3 and 4, above). Indeed, when merging the data across all conditions in all sub-experiments across all condition pairs, a 2% difference in performance between combinations emerged, which was found to be significant (lowED/highTP mean = 56.8%, highED/lowTP mean = 54.8%, $t(610) = 3.33$, $p = .001$; BF = 10.7). It thus seems that when considering VSL performance across all conditions, the effect of ED seems perhaps less pronounced than that of TPs—that is, a change from a low-ED to a high-ED condition impacts VSL performance less than a parallel change from a low-TP to a high-TP condition. In more theoretic terms, it is possible that while combining the two measures together, ED has a lesser influence on VSL compared to TPs. What could be the source of imbalance between ED's and TPs' impact on VSL performance?

One possible source could be a nonlinear impact of ED on task performance. As documented in many fields (e.g., Carver, 1973; Sable, Gratton, & Fabiani, 2003; Sherman & Turvey, 1969), presentation rate tends to have a nonlinear effect on performance, where a change in very fast EDs (e.g., a change in ED from 200 ms to 300 ms) have a greater impact on performance than a parallel change in slower EDs (e.g., a change from 900 ms to 1,000 ms, which is negligible). This would suggest that for obtaining an even better fit of the data, the simple multiplication of entropy (bits/element) by rate of presentation (elements/second) for computing RI (see Eq. 1, above) should incur some correction. Mathematically, this can be done by having a power smaller than 1 for the rate of presentation part of the equation (noted as x in formula (2) below), which aims to capture the decrease in the effect of presentation rate with increasing baselines:

$$Modified\_RI = \text{entropy} * \text{rate of presentation}^{x} \ (\text{where } x < 1) \tag{2}$$

To examine whether such modification improves the predictive power of RI, we conducted an aggregated analysis of all data from Experiments 2a and 2b (n = 611 subjects

overall). The data were analyzed using logistic mixed-effect models, with correctness in each trial as the dependent variable, as well as a by-subject random intercept. We then used a bootstrapping procedure[3] to find the value of x (the exponent of rate of presentation) that best predicts the data. The average exponent of rate of presentation which produced the best model fit was $x = 0.7$ ($SD = 0.11$). The value being smaller than 1 with a small *SD* suggests that, indeed, a modification of the RI calculation was beneficial. To cross-validate this proposed modification of the RI calculation, we returned to the results of *Experiment 1* and examined whether a better model fit would be reached by a (log-transformed) modified-RI, compared to the (log-transformed) "simple RI" we used before. A slightly lower BIC value was found when using modified-RI as a predictor, with BIC = 9,458, as compared to the original "linear-RI," which had a BIC = 9,460. This suggests that modified-RI accounted for the data $e^{2/2} = 2.72$ times better than the "simple RI" model.[4] To illustrate this finding, Fig. 5 presents SL performance in the different conditions of Experiment 1 as a function of log-transformed modified-RI, which accounts for 38.5% of the variance across conditions (cf. Fig. 1 above which presents the same results with the "linear-RI," and where $R^2 = 35\%$). To avoid making a claim based on overfitting the available data, however, we should caution that firm conclusions regarding the exact power of ED for RI computations require further investigation (and see Section 6, below).

6. General discussion

Learning the regularities embedded in a continuous input stream has been taken to be a basic cognitive ability demonstrated across modalities, age, and species. In the visual modality, it has been shown across a wide range of stimuli differing in complexity. In the present paper, we address the question of how precisely SL computations are done, aiming to offer a theoretical account regarding the possible nature of computed representations. Assuming that learning the embedded regularities between elements in an input stream requires encoding the elements' representations as well as computing their TPs, our starting point was the surprising interaction between encoding constraints and TPs computations (Bogaerts et al., 2016). This interaction ruled out an account based on two separate stages of processing, one in which stimuli are encoded, the other which subsequently computes the TPs between the elements (Frost et al., 2015), suggesting that these processes occur in parallel. But how can we account for such parallel processing? What is the nature of the computed representations during learning? Here we offer a unified principle that subsumes this interaction and considers in parallel both TPs (the amount of information carried by each element in the stream) and the time allocated to compute this information. As such, we suggest that the representations of the amount of information encompassed in each element in the stream (entropy), and how they are built over time, belong to one unified construct—*Rate of information*. This account blurs the distinction between the encoding of elements and the extraction of the statistical properties embedded in the stream (cf. Frost et al., 2015). Instead, it suggests that learning patterned

regularities in an input can be described as a process of information extraction, a process constrained either by extent of unpredictability of elements or by challenging presentation parameters, and these two components can be interchangeable. This view thus conforms with neurobiological findings, showing entropy-related computations already in early sensory cortices (Nastase et al., 2014; Tobia, Iacovella, & Hasson, 2012; and see Hasson, 2017 for a detailed discussion of possible underlying neurobiological mechanisms).

We provide substantial evidence supporting this theoretical claim, based on data from more than 800 subjects overall. In Experiment 1 we show across a wide range of TPs and ED combinations that RI monotonically determines VSL performance in a function that is logarithmic. Our findings demonstrate that log-transformed RI provides by far the best fit of VSL performance, predicting learning outcomes significantly better than the joint combination of ED, TP, and ED by TP interaction. In Experiments 2a and 2b we take this finding a step further, testing the strong prediction that similar RIs will result in similar performance regardless of ED and TP combinations. Using Bayesian analysis, we show that RI offers the most probable account for VSL performance.

Considering SL theory, an important question is whether our findings from the visual domain can be extended to auditory inputs. From a pure information theory perspective, entropy can be assessed in any input regardless of modality. Indeed, recent studies targeting speech production suggest that the extent of word surprisal determines speech rate so that unpredictable (informative) words are spoken more slowly (e.g., Cohen Priva, 2017). However, whereas it is possible that speakers naturally adjust their speech rate given word surprisal, the extrapolation of RI calculations to the auditory modality is not trivial and requires targeted research. First, the visual and auditory cortices vastly differ in the parameters defining efficiency in computing representations from continuous input streams (Conway & Christiansen, 2006; Emberson, Conway, & Christiansen, 2011; Frost et al., 2015). Second, and perhaps more important, when linguistic stimuli compose the input stream (as is the case in the Saffran et al., 1996 task), participants have

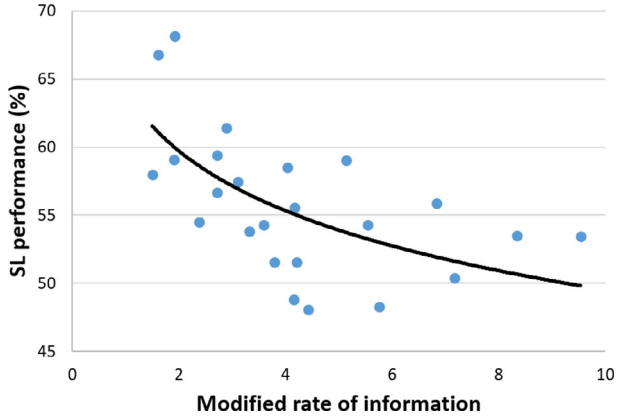

Fig. 5. Results of the cross-validation procedure, depicting statistical learning (SL) performance by modified rate of information (RI) in the 24 conditions of Experiment 1. Line shows best logarithmic fit.

well-established linguistic representations they bring to the learning task (see Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018). Such prior expectations may either increase or decrease the input's initially perceived entropy, based on the overlap between the input and prior linguistic exposure, as well as potentially alter the impact of rate of presentation on SL computations. Thus, how RI impacts auditory SL with linguistic elements (as well as other facets of SL, see Siegelman, Bogaerts, Christiansen, & Frost, 2017) remains an outstanding question.

The current results also provide a new perspective to the role of SL in language acquisition and processing. Under the common view, efficient language use relies on SL abilities as it requires sensitivity to regularities embedded in the input (see, e.g., Gómez & Gerken, 2000; Romberg & Saffran, 2010). The current work refines this view, suggesting that the joint shared mechanism behind SL and language is information extraction. According to this perspective, languages can be viewed as a mean for transferring novel information in a given time unit (see also, e.g., Cohen Priva, 2017; Linzen & Jaeger, 2015), which is then extracted through SL computations. In the same vein, our account also refines the interpretation of individual-differences studies showing a correlation between SL performance and linguistic abilities (e.g., Arciuli & Simpson, 2012; Frost, Siegelman, Narkiss, & Afek, 2013). It suggests that the source of these correlations is different information extraction capacities of different individuals, where some individuals can process more information per time unit compared to others. This view thus resembles the classic work by George Miller, who argued that individuals exhibit varying working memory performance as they differ from one another in the maximal amount of information they can process (Miller, 1956).

Our findings, however, raise additional questions that require further investigation. Recall that to account for the decreasing effect of ED on RI in slower presentation rates, we employed a transformation on RI using a power smaller than 1 on presentation rate. This transformation seems to fit best the range of EDs employed in our study, EDs that are typically employed in SL experiments. One possibility to consider, however, is that very long EDs may have an adverse effect on learning (leading to worse VSL performance), because they implicate higher demands on working memory. Investigating this possibility requires data collected in conditions with EDs longer than 1 s, combined with an exploration of different possible mathematical transformations that allow for non-monotonicity. Relatedly, we should stress that the exact mathematical transformation of ED needs to be further validated using other visual stimuli that differ in complexity from those used here. We should also note that in our present investigation we did not employ a separate and independent measure to estimate a-priori participants' encoding abilities. Rather, we manipulated encoding and learning related factors within the SL task. Our RI account highlights the interplay of encoding and assimilation of statistics in accounting for SL performance. A future extension of this work could therefore examine whether and to what extent individual differences in SL performance can be traced back to individuals' encoding abilities.

Another important outstanding question concerns the impact of RI on the dynamics of learning. Our current RI formulations only consider the expected mean of RI in an input stream, and examine its power in accounting for learning outcomes following a relatively

long exposure phase. What remains unexplored is how RI modulates performance continuously throughout learning. Future work is left with the task of assessing the impact of RI dynamically, measuring how the rate of information presented to learners at each time point during familiarization is related to their gradual accumulation of knowledge, eventually leading to final SL performance. This requires, however, more advanced models that estimate changes in information as stimuli unfold (see, e.g., Kidd, Piantadosi, & Aslin, 2012), combined with online measures of SL which continuously tap into learning abilities throughout the familiarization phase (see Siegelman, Bogaerts, Kronenfeld, & Frost, 2018, for discussion).

Lastly, the current work also opens broad theoretical questions regarding the role of SL across cognition, and in language in particular. We found that over the course of an experimental session, adults show above-chance learning of streams involving few bits of information to be learned (0–10 bits in our experimental setting). This amount of information is smaller tenfold compared to the amount of information that needs to be extracted by learners in real-world tasks. A recent work estimated that an average human has to extract and store around 1.5 million bits of information during language acquisition (Mollica & Piantadosi, 2019). This discrepancy in scale raises questions regarding whether and how SL findings scale up to account for real-world learning tasks in general, and language acquisition and processing in particular. Assuming that SL plays a key role in language acquisition and use, SL research should explicitly consider the exact mechanisms that are available to learners, as well as the full array of statistical and non-statistical cues that are available to them in the input, to provide a comprehensive account of how learners assimilate the huge amounts of information that they eventually learn.

To conclude, we suggest that learning regularities in the visual modality is best described by considering the rate of novel information that has to be processed in the input stream. The advantage of RI as a theoretical construct is that it offers a unifying computational principle that combines the extent of informativeness (or entropy) in the stream and the time that is allocated to process it. This approach holds the promise of generating precise computational accounts and, therefore, testable predictions for understanding how regularities in an input stream are learned.

Acknowledgments

This paper was supported by the ERC Advanced grant awarded to Ram Frost (project 692502-L2STAT), the Israel Science Foundation (Grant 217/14 awarded to Ram Frost), and by the National Institute of Child Health and Human Development (RO1 HD 067364 awarded to Ken Pugh and Ram Frost, and PO1 HD 01994 awarded to Haskins Laboratories). Noam Siegelman is a Rothschild Yad-Hanadiv post-doctoral fellow. Louisa Bogaerts received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 743528 (IF-EF). We wish to thank Uri Hasson for helpful discussions, and Blair Armstrong for useful comments on a draft of this paper.

Notes

1. Note that Markov Entropy focuses on the amount of information of each element given the TPs in the stream. This type of entropy is specifically relevant to learning situations in which TPs between elements are the available cues for learning, such as in typical SL tasks on which we focus here. In other learning situations, other measures of entropy should be applied. For example, some learning situations require the learning of the overall distribution of elements, rather than the TPs between them (see Thiessen, Kronstein, & Hufnagle, 2013, for discussion). In these situations, the information in the input should be quantified by Shannon Entropy, which focuses on the (marginal) probability of each element (regardless of the TPs between elements).
2. Entropy can be calculated using Log in different bases (2, *e*, or 10). Here we use Log in the basis of 2 in all calculations following the convention in Information Theory.
3. In each run, we randomly sampled $n = 200$ subjects. We then ran a series of mixed effect models, each time with a different exponent x, running from $x = 0.2$ to $x = 1$, in intervals of 0.05 (i.e., 0.2, 0.25, 0.3, . . ., 1). For each x, a mixed model was run, with log-transformed modified-RI as a fixed-effect. After running on all values from 0.2 to 1 for this sub-sample, the optimal value of x for this run was selected, based on the value that resulted in a lowest BIC value for the model (representing best model fit). We reiterated this bootstrapping procedure $k = 100$ times.
4. To re-validate the RI account and its superior predictive power compared to models comprising of ED and TP, we also compared the model with the modified RI to models with power-transformed ED (i.e., $ED^{0.7}$), TP, and their interaction. Thus, we ran two additional models with ED powered by 0.7, with either TP or log-transformed TP, and their interaction, namely: (1) $VSL \sim ED^{0.7} + TP + ED^{0.7}{:}TP$, and (2) $VSL \sim ED^{0.7} + \log(TP) + ED^{0.7}{:}\log(TP)$. Both models produced BIC values substantially higher compared to the modified RI model: BIC = 9,474 and BIC = 9,475, respectively (ΔBIC of at least 16 compared to the modified RI model; see main text).

References

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*(2), 286–304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Barakat, B. K., Seitz, A. R., & Shams, L. (2013). The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted. *Cognition*, *129*(2), 205–211. https://doi.org/10.1016/j.cognition.2013.07.003

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review*, *23*(4), 1250–1256. https://doi.org/10.3758/s13423-015-0996-z

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*, 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Carver, R. P. (1973). Effects of increasing the rate of speech presentation upon comprehension. *Journal of Educational Psychology*, *65*(1), 118–126. https://doi.org/10.1037/h0034783

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, *160*, 27–34. https://doi.org/10.1016/j.cognition.2016.12.002

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science*, *17*, 905–912. https://doi.org/10.1111/j.1467-9280.2006.01801.x

de Fleurian, R., Blackwell, T., Ben-Tal, O., & Müllensiefen, D. (2017). Information-theoretic measures predict the human judgment of rhythm complexity. *Cognitive Science*, *41*(3), 800–813. https://doi.org/10.1111/cogs.12347

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, *64*, 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Fan, J. (2014). An information theory account of cognitive control. *Frontiers in Human Neuroscience*, *8*, 680. https://doi.org/10.3389/fnhum.2014.00680

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

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

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*(7), 1243–1252. https://doi.org/10.1177/0956797612472207

Gómez, R., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186. https://doi.org/10.1016/S1364-6613(00)01467-4

Grieco-Calub, T. M., Simeon, K. M., Snyder, H. E., & Lew-Williams, C. (2017). Word segmentation from noise-band vocoded speech. *Language, Cognition and Neuroscience*, *32*(10), 1344–1356. https://doi.org/10.1080/23273798.2017.1354129

Hasson, U. (2017). The neurobiology of uncertainty: implications for statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160048. https://doi.org/10.1098/rstb.2016.0048

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382–1411. https://doi.org/10.1111/cogs.12274

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352. https://doi.org/10.1037/h0043158

Mollica, F., & Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science*, *6*(3), 181393. https://doi.org/10.1098/rsos.181393

Morey, R. D., Rouder, J. N., & Jamil, T. (2014). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9, 8.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*(2), 135–167. https://doi.org/10.1214/12-STS410

Nastase, S., Iacovella, V., & Hasson, U. (2014). Uncertainty in visual and auditory series is coded by modality-general and modality-specific neural systems. *Human Brain Mapping*, *35*, 1111–1128. https://doi.org/10.1002/hbm.22238

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906–914. https://doi.org/10.1002/wcs.78

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, *37*, 1290–1320. https://doi.org/10.1111/cogs.12050

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sable, J. J., Gratton, G., & Fabiani, M. (2003). Sound presentation rate is represented logarithmically in human cortex. *European Journal of Neuroscience*, *17*(11), 2492–2496. https://doi.org/10.1046/j.1460-9568.2003.02690.x

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, Jenny R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Sherman, M. F., & Turvey, M. T. (1969). Modality differences in short-term serial memory as a function of presentation rate. *Journal of Experimental Psychology*, *80*(2 Part 1), 335–338. https://doi.org/10.1037/h0027276

Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? Insights from Bayesian modeling. *Cognition*, *192*, 104002. https://doi.org/10.1016/j.cognition.2019.06.014.

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), 20160059. https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432. 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

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*, 792–814. https://doi.org/10.1037/a0030801

Tobia, M. J. M., Iacovella, V., & Hasson, U. (2012). Multiple sensitivity profiles to diversity and transition structure in non-stationary input. *NeuroImage*, *60*(2), 991–1005. https://doi.org/10.1016/j.neuroimage.2012.01.041

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics*, *67*, 867–875. https://doi.org/10.3758/BF03193539

Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology-General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Raw data are available at: https://osf.io/6brs2/.