Towards a theory of individual differences in statistical learning

Noam Siegelman [1], Louisa Bogaerts [2], Morten H. Christiansen [3, 4], and Ram Frost [1, 4, 5]

[1] The Hebrew University of Jerusalem, Israel

[2] CNRS and University Aix-Marseille, France

[3] Cornell University, Ithaca, NY, USA

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

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

Abstract

In recent years statistical learning (SL) research has seen a growing interest in tracking individual performance in SL tasks, mainly as a predictor of linguistic abilities. We review studies from this line of research and outline three presuppositions underlying the experimental approach they employ: (1) that SL is a unified theoretical construct, (2) that current SL tasks are interchangeable, and equally valid for assessing SL ability, and (3) that performance in the standard forced-choice test in the task is a good proxy of SL ability. We argue that these three critical presuppositions are subject to a number of theoretical and empirical issues. First, SL shows patterns of modality- and informational-specificity, suggesting that SL cannot be treated as a unified construct. Second, different SL tasks may tap into separate sub-components of SL, that are not necessarily interchangeable. Third, the commonly used forced-choice tests in most SL tasks are subject to inherent limitations and confounds. As a first step we offer a methodological approach that explicitly spells out a potential set of different SL dimensions, allowing for better transparency in choosing a specific SL task as a predictor of a given linguistic outcome. We then offer possible methodological solutions for better tracking and measuring SL ability. Taken together, these discussions provide a novel theoretical and methodological approach for assessing individual differences in SL, with clear testable predictions.

Keywords: *Statistical learning; Individual differences; Online measures; Predicting linguistic abilities*.

Introduction

Over the past two decades, extensive research has focused on statistical learning (SL), demonstrating sensitivity to complex distributional properties in the input. Starting from the seminal work of Saffran and colleagues [1], numerous studies have shown that humans display remarkable sensitivity to distributional regularities in the auditory [2], visual [3], and tactile [4] modalities, with verbal [5] or non-verbal [6] stimuli, comprising adjacent or non-adjacent [7] dependencies, over both time and space [8], even without overt attention [9], and from a very young age [10]. Sensitivity to the input's statistical structure has become an important theoretical construct in explaining a wide range of human capacities such as language learning, perception, categorization, segmentation, transfer and generalization (see [11], for discussion).

Whereas all of the above studies focused on demonstrating that a given sample of participants shows evidence of learning the distributional properties of a sensory input, recent years has seen a growing interest in tracking *individual performance* in SL tasks. This line of study is relatively new. Its initial motivation was to confirm the theoretical link between SL and language acquisition. However, more generally, the study of individual differences holds the promise of providing critical insights regarding the mechanisms of SL and could enable more powerful studies ([11–13]; see also [Arciuli, this issue]). Note that "*individual differences" in* the context of SL can in principle refer to any quantitative or qualitative differences between individual learners (i.e., differences in both the extent and the speed/trajectory of learning, individual variation in the sensitivity to multiple statistics within the same input, etc.). Nevertheless, individual differences other than overall performance differences have to date rarely been investigated. We return to this issue further on, when considering the limitations of the currently used offline learning measures. For now, the important point is that these recent SL studies that tracked individual performance aimed to show that language learning relies, at least in part, on being sensitive to the statistical properties of a linguistic environment, and that individual variation in sensitivity to such regularities predict linguistic abilities. Within this research program SL and artificial grammar learning (AGL) tasks were shown to correlate with literacy skills in L1 [14,15], literacy acquisition in L2 [16], comprehension of syntax [17], sentence processing [13,18,19], semantic and phonological lexical access [20], vocabulary development [21,22], and speech perception [23,24]. Conversely, other studies aimed to show that participants with language deficits such as children with specific language impairment ([20,25], but see [26]), dyslexics readers [27,28], and agrammatic aphasia patients [29], display poor SL abilities.

This research is characterized by a prototypical experimental approach. First, a SL or AGL task that has been shown to produce above chance performance in the group level is selected, and imported into the study as is or with minor modifications. Typically, the tasks involve a visual or an auditory familiarization stream (representing an artificial grammar or a stream comprising set of transitional probabilities), which is followed by a test phase. Second, individual performance in the task is registered for each participant (often the number of correct two-alternative forced-choice [2AFC] decisions in distinguishing presented visual or auditory sequences from foils at the test phase). Third, given the aim of the study (e.g., reading, syntactic processing, speech recognition, etc.), participants' capability in the respective linguistic domain is independently measured through well-established relevant language tests. Fourth, the participants' SL scores are used as predictors of their linguistic test performance. Table 1 presents a set of recent studies that followed this approach, including our own, along with the correlations they obtained.

	SL task(s)	Operational SL measure	Linguistic measure	Studied population	Number of participants	Obtained correlation
Arciuli & Simpson [14]	Visual SL	Success in 64 2AFC trials	Reading skills (reading sub-test WRAT-4)	adults	37	r=0.34
				6-12yo children	38	r=0.33
Conway et al., 2010 [24]	Visual AGL	Difference in span between grammatical and ungrammatical sequences in test	Speech perception in noise	adults	23	r=0.46
	Auditory AGL				22	r=0.42
	Visual AGL				64	r=0.31
Frost et al., 2013 [16]	Visual SL	Success in 32 2AFC trials	Learning scores in nonword decoding, word reading, and morphological priming	adult L2 learners	27	r=0.44 to 0.57
Kidd & Arciuli, 2015 [17]	Visual SL	Success in 64 2AFC trials	Auditory syntax comprehension task	6-8yo children	68	Pearson's r not reported. SL predicts comprehension of passives and relative clauses (logistic mixed-effect models).
Mainela-Arnold & Evans, 2014 [20]	Auditory SL	Success in 2AFC test	Gating task (lexical-phonological skills), word-definition task (lexical-semantic)	8-12yo children with SLI	20	r=0.2 for both linguistic tasks.
				8-12yo typically developing children	20	r=0.28 (phonological); r=0.1 (semantic)
Misyak & Christiansen, 2012 [13]	Two auditory AGL tasks: adjacent and non-adjacent	Success in 2AFC test	Comprehension of different types of grammatically complex sentences.	adults	30	r=0.11 to r=0.49
Misyak et al., 2010 [32]	Auditory non adjacent AGL, combined with SRT	Differences in the ability to predict the final non-adjacent dependent element after training	Self-paced reading of sentences involving object relative clauses	adults	20	r=0.59
Shafto et al., 2012 [21]	Visual SL	RT difference of eye movements towards predictable stimuli between learning and test.	Early receptive vocabulary skills	7.5 month-old infants	58	r=0.28
Spencer et al., 2014 [15]	Auditory SL and visual AGL	Success in 4 2AFC test trials for SL; Difference in span between grammatical and ungrammatical for AGL	A series of 10 tasks related to early literacy skills	4-10yo children	553	ranging from 0 to 0.2

Table 1 *Summary of recent individual differences studies predicting linguistic abilities from SL performance*

Although never explicitly specified, individual differences studies of this kind typically involve three critical preliminary presuppositions which underlie the logic of this experimental strategy. First, since there is no agreed taxonomy of possible types of SL, it is treated by default as a unified theoretical construct, *a general capacity for picking up regularities* (with the exception of [13,30]; see, e.g., [31], for discussion). Second, and relatedly, the tasks which are selected for the study from the arsenal of tasks employed in this domain, are naturally assumed to equally represent a good operational proxy of this unified theoretical construct, so that the selection of one specific task for the study is not a matter of deep theoretical concerns (though see [13,30,32])[1]. Third, the performance score of the test phase in the task is naturally assumed to be a valid and reliable measure of the operational proxy, and therefore, a valid and reliable measure of the postulated ability for picking up regularities.

In the following, we will argue that these three critical presuppositions are subject to a number of both theoretical and empirical issues. Although previous studies of individual differences in SL have yielded important initial insights into how SL might be involved in various aspects of cognition, to get a deeper understanding of the extent and precise nature of these relationships we need to address these issues head on.

Is SL a general unified capacity?

Most studies of SL do not provide an explicit computational account of learning but, rather, tend to adopt a more abstract notion of the underlying computations in the form of domain-general learning. Typically, the underlying computational system is assumed to be a "unified capacity" instantiated by a unitary learning system that is applied across different modalities and domains. This may be a reasonable first approximation, given

[1] Admittedly, some coarse-grained taxonomy between AGL and SL tasks exist, so that AGL tasks are typically selected to examine syntactic abilities (e.g., [18]).

that the ability to extract statistical structure from the input is found across a wide range of stimuli as well as different domains, as reviewed above. Indeed, in the simple and abstract sense, there is something common to all these behavioral phenomena: registering regularities in the environment. However, advances in cognitive science require moving from abstract verbal theorizing to refined mechanistic computational theories. From this perspective, it seems that current empirical evidence suggests that the differences in computations across different SL phenomena, largely outweigh their superficial abstract similarity.

Modality specificity: Whereas SL has been demonstrated in all sensory and sensory-motor areas, current evidence systematically suggests qualitatively different patterns of performance in different modalities (see [11], for review). Importantly, tracking individual abilities in different SL tasks reveals significant reliability of capacity within modality, but zero correlation in performance across modalities [33]. Admittedly, one should be cautious drawing firm conclusions from a lack of correlations in a single study, especially given the relatively low reliability of some of the studied SL tasks (which limits the extent of expected correlations between SL measures, see [12,33]). Importantly, however, this result concurs with other findings showing qualitative differences in SL ability in the auditory, visual, and tactile modalities [4,34], opposite effects of presentation parameters on visual vs. auditory SL performance [35], lack of learning transfer across modalities (e.g., [36]), and interference in learning two artificial grammars within modality, but no interference across modalities [37]. This large body of evidence suggests that individual capacity of learning regularities differs across domains. This state of affairs should not come as a surprise. Recent imaging data suggest that in spite of the suggested role of the medial temporal lobe (MTL) memory system in SL (e.g., [38,39]), substantial SL computations occur already in the early visual and auditory cortices (e.g., [40,41]). The visual and auditory cortices involve different representations, and the set of computations characterizing these cortical areas is naturally constrained by the specific characteristics of the processed input. Thus, both the neurobiological and the behavioral evidence are inconsistent with the presupposition that SL is a unified capacity.

Informational specificity: Although SL can be abstractly defined as "learning the statistical properties of the continuous sensory input", from an informational perspective there are different kinds of "statistical properties" which are the object of learning (see [42], for discussion; see also [Hasson, this issue]). First, there is ample evidence that humans are sensitive to transitional statistics in continuous input, allowing them to detect even small changes in Transitional Probabilities (TPs) [43][2]. Second, there is evidence that humans also aggregate information about the relative frequency of events (e.g., [44]), as well as their variance in the stream (e.g., [45]), showing sensitivity to distributional statistics. Cue-based statistics as revealed in spatial contextual-cuing (e.g., [46]), or temporal cuing (e.g., [47]), is yet another form of learned regularities. In some cases, multiple cues either within or across modalities are needed to learn more complex probabilistic patterns [48]. As Thiessen et al. discuss in their expansive review [42], different kinds of statistical information do not necessarily implicate different sets of computations. Nevertheless they argue that a complete account of statistical learning must explain not only the learning of distributional

[2] That learners display sensitivity to TPs does not necessarily entail that the underlying computational mechanism of SL explicitly represents TPs between sequential elements. Indeed, an alternative theoretical accounts assume that the seeming sensitivity to transitional statistics emerges from chunking due to the repetition of groups of elements (e.g., [31,79–81]; see also [82]).

statistics (i.e., the frequency and variance of exemplars) but also transitional statistics (i.e., learning the co-occurrences of elements in the stream).

Whether one or more kinds of computations are needed to cover the range of SL behaviors requires additional investigation, mainly through computational modeling, but also through correlational designs. For example, it has been suggested that learning non-adjacent contingencies follows specific constraints that do not exist while learning adjacent contingencies [7]. Indeed, supporting findings show that individual SL ability to learn adjacent contingencies is uncorrelated with their ability to learn non-adjacent contingencies even within modality [13,33,49][3].

In sum, current empirical evidence is largely inconsistent with SL being a unified capacity involving a single set of computations. This has immediate implications for any correlational study aiming to tie specific cognitive abilities to SL. We suggest that such studies need to consider SL as a componential ability, requiring researchers to explicitly specify the theoretical link between the specific cognitive construct they investigate and its relation to the specific relevant SL computations.

2. Are all SL tasks equally valid for assessing SL ability?

To date there are no agreed-upon constraints on which tasks should be selected as proxies for SL capacity. This is exemplified by the different tasks employed in correlational studies tying SL to other cognitive capacities, with often very little discussion regarding the theoretical logic governing the specific task selection (but see, e.g., [13], for such discussion). The problem with this state of affairs is twofold. First,

[3]Importantly, though, comparing potentially different kinds of computations in correlational designs requires careful attention to the detailed probability structure of such computations. For instance, when controlling for probability of occurrence between dependencies, Vuong, Meyer, & Christiansen [76] found that adjacent and nonadjacent dependencies could be learned simultaneously.

without a clear understanding of the specific SL components that are being tapped by a given task, well-defined empirical predictions regarding its predictive validity cannot be generated. Second, understanding the relation between specific SL components and the proxies selected to tap them is necessary for integrating different findings, so as to make sense of the wide range of obtained results. In order to develop such integrative theory of the relations between SL computational components and linguistic capacities (as well as other cognitive capacities), we must first explicitly spell out the different components of SL capacity that, according to current evidence, is a multi-faceted construct.

One promising way to develop a theory regarding the inner structure of a complex construct is to define it in the form of a mapping sentence in line with Facet Theory, a systematic approach to theory development and data collection (e.g., [50,51]). In Facet Theory, the first and most important step in investigating a complex theoretical construct (in our case, SL), is to formulate a mapping sentence, which defines the full domain of the studied phenomena given existing data. A mapping sentence includes *content facets* that represent the different dimensions of the construct. It further outlines for each content facet a set of possible values (categorical or continuous) which could be relevant to the specific facet. This divides the full range of behavioral phenomena into theoretically distinct sub-types [51]. Importantly, one of the unique characteristics of Facet Theory is that it is taken to be a continuous effort of trial and error, where constructing a mapping sentence that outlines the various facets of a theoretical construct resembles an ongoing process of hypotheses testing and updating. An initial sentence is typically offered as a starting hypothesis (see [33]), and it is subsequently modified given novel empirical data regarding the inter-correlations between the suggested facets and their postulated values. Following this strategy, we define a preliminary mapping sentence below that concurs with a wide range of SL phenomena already reported in the literature, and outlines a potential set of different dimensions:

Statistical Learning is the ability to pick-up $^{(1)}\begin{Bmatrix} transitional \\ distributional \end{Bmatrix}$ *statistics from the sensory environment, in the* $^{(2)}\begin{Bmatrix} visual \\ auditory \end{Bmatrix}$ *modality, when contingencies are* $^{(3)}\begin{Bmatrix} adjacent \\ non-adjacent \end{Bmatrix}$, *over* $^{(4)}\begin{Bmatrix} verbal \\ non-verbal \end{Bmatrix}$ *material, across* $^{(5)}\begin{Bmatrix} time \\ space \end{Bmatrix}$, $^{(6)}\begin{Bmatrix} with \\ without \end{Bmatrix}$ *motor involvement, thereby shaping behavior.*

This suggested mapping sentence offers then six preliminary content facets to account for SL phenomena [4]. The first three facets: the type of statistics extracted (transitional vs. distributional), the input modality (visual vs. auditory)[5], and the type of contingencies (adjacent vs. non-adjacent), were included in light of empirical evidence (reviewed in the previous section), and which have been suggested to involve non-overlapping sets of computations. Facets (4) and (5) are additional hypothetical dimensions that we offer to account for SL capacity, since they reflect ecologically separable phenomena: SL studies show that it occurs for both verbal and non-verbal material (e.g., [6]), and that statistical contingencies are extracted across both time and space (e.g., [8], though with different biases, see [34]). Admittedly, to date there is little unequivocal evidence showing that these phenomena are governed by non-overlapping computations and necessarily result in different learning constraints. Nevertheless, our recent investigation of SL capacities demonstrates no correlation in performance with verbal vs. nonverbal stimuli within modality [33]. Similarly, no interference was found

[4] Note that computations related to different values within facet of SL may operate in parallel. Indeed, there is compelling evidence that can learners can exploit more than one source of statistical information at the same time (e.g., [49,76,83]), although sometimes at the cost of interference [84].
[5] Because sensory information related to SL phenomena is mostly visual or auditory, the tactile modality is omitted for the sake of simplicity.

in learning two different sets of regularities at the same time, when they comprised verbal and nonverbal materials (nonwords vs. tones, [37]) [6] . Indeed, recent neurobiological findings suggest that neural temporal coding is independent of the spatial dimension, and that specific time cells represent the flow of time (see [52]).

Importantly, from a theoretical perspective, including facets (4) and (5) in the mapping sentence has the advantage of shaping future investigation, so as to examine empirically the extent of their relative overlap and interaction (see [43], for discussion). Facet number (6) – motor involvement – is yet another dimension that requires further investigation. Statistics of an input can be extracted without any motor involvement (such as in the case of most SL or AGL tasks). However, some SL tasks specifically involve active motor responses to stimuli (such as in the case of motor sequence statistical learning, best exemplified by the Serial Reaction Time (SRT) task, e.g., [53]). Whether such motor activity results in non-overlapping sets of computations in extracting statistical structure, is then another open question awaiting future research (see, e.g., [54] for a discussion).

Mapping sentences typically start small and grow bigger as empirical investigation progresses. Our initial proposed mapping sentence, therefore, does not preclude the possibility that other dimensions may be relevant for understanding SL ability. Possible additional candidate facets could be, for example, basic perceptual dimensions (color, line orientation, etc.; e.g., [37]), full vs. quasi regularity (see [43], for discussion), implicit vs. explicit learning settings (e.g., [55]), or, relatedly, unsupervised vs. supervised learning settings (see [56] for a discussion of the role of

[6] Here we *do not* argue that verbal stimuli are special in the sense that they require a hardwired specific neurobiological mechanism. Rather, verbal stimuli (e.g., syllables) differ from non-verbal stimuli (e.g., tones) in the sense that they involve extensive prior exposure, which inevitably effects learning.

feedback in perceptual category learning). An additional factor that was shown to affect SL performance is rate of presentation – with opposite effects of both the inter stimulus interval and the actual stimulus duration on SL performance in the visual versus auditory modality ([34,35]; but see [57]). Whether rate of presentation constitutes a separate facet, or simply affects peripheral aspects to SL such as the encoding of individual elements, with different constraints in different modalities (see [11]), deserves further investigation.

Defining a mapping sentence as a working hypothesis for studying individual differences in SL enables theoretical discussions regarding how and why specific SL components modulate specific sub-components of other cognitive abilities, given their overlapping hypothesized computations. This makes the logic of choosing specific SL tasks for a given study more transparent, and allows a clear interpretation of the findings. For example, different components of linguistic phenomena most likely involve more than one type of underlying SL computations. Acquiring phonotactic constraints of a language requires registering both transitional and distributional statistics[7] of phonemes in the speech stream via the auditory modality [58], while learning to read in L1 or L2 involves assimilating transitional statistics of letter sequences in the visual modality, but also aggregating systematic correlations between letters and sound, and between letter sequences and meaning through morphological form (see [59], for discussion). The mapping sentence above thus allows for more refined discussions of the components involved in each linguistic capacity and its relation to SL.

[7] Note that distributional and transitional statistics overlap given that to compute transitional probabilities (e.g., between phonemes), the learner needs to keep track of the frequency of phonemes and phoneme pairs (or bigrams). For example, the forward transitional probability of phoneme Y following phoneme X is computed as Frequency (XY) / Frequency (X), requiring the learner to register both the distribution of biphone pairs (XY) and that of the individual phonemes (X).

Importantly, a mapping sentence for SL not only dissects the outcome cognitive phenomena in terms of their different statistical computations, but also points to tasks that could (or should) be used to measure SL as predictors of a specific ability. To date, the arsenal of tasks tapping SL capacity is impressively varied: in addition to those reviewed in Table 1, tasks such as the Serial Reaction Time (e.g., [60]), Contextual Cuing (e.g., [61]), Tone Detection (e.g., [62]), or Hebb Repetition Task (e.g., [63]), are all considered to be proxies of SL, since they all involve learning statistical regularities. The advantage of a mapping sentence is that it provides a priori criteria for selecting one of the many available tasks for a given study, specifying the inter-relations between them. For example, in contrast to tasks such as visual SL or SRT that tap the extraction of transitional statistics, tasks such as Contextual Cuing require registering the distribution of stimuli to learn the repeated patterns, whereas tasks such as AGL involve both learning of units defined by transitional statistics (see, e.g., [64]), as well as their distributional statistics [42].

So far we have advocated a research strategy that requires researchers to be very explicit about what specific computations involved in a given SL task and their predicted outcomes. However, if the target of research is to assess the *overall* SL capacity of an individual as defined by the mapping sentence, as well as its predictive validity, the proposed mapping sentence provides specific guidelines for developing novel SL tasks to cover a wide range of SL components. Here we propose that if SL is indeed a multi-faceted construct involving different types of computations with substantial non-overlapping variance, then this capacity should be measured and assessed by a variety of different tasks. Much like in the measurement of other complex constructs (e.g., the *g* factor measured by WAIS, [65]), accurate estimation of multi-faceted constructs involves a large battery of tasks, each covering different parts of the variance. But note, that in contrast to general intelligence, which has been mapped through decades of extensive research, the dimensions of SL as an individual ability are yet to be empirically established. Our mapping sentence attempts to offer a preliminary approximation of the possible facets of SL, serving as springboard for such research. At this point, we argue that current evidence points to SL as a multi-faceted individual ability. Selecting tasks as proxies for this ability thus requires an integrative approach with explicit discussions of the specific components which are being tapped.

3. Are standard task test scores a good proxy of SL ability?

The vast majority of studies tracking individual differences in SL employ the same tasks that were originally designed for group-level studies. Here the underlying assumption is that the outcome measure of performance in the task would serve as a good proxy or indicator of the theoretical construct: individual SL ability. We see two problems with this assumption. First, from a methodological perspective although the typical SL tasks can reliably estimate the mean performance of the sample as a whole, they are often not sensitive enough to estimate a given individual's SL ability. Second, as we outline bellow, from a theoretical perspective, the structure of the tasks often intermixes outputs of different SL computations. This practice is likely to confound cognitive capacities that are orthogonal to SL, while also potentially lead to interference effects that mask the true capacity of SL.

Psychometric weakness. A task that is suitable for measuring individual capacity must show substantial between-individual variance and this variance must be highly reliable. If not, the task cannot differentiate between good and bad learners, and cannot reliably predict other cognitive capacities. As we have recently argued [12], most SL tasks that have been used for group-level studies do not withstand psychometric scrutiny. This is due to a number of shortcomings, such as insufficient number of test trials, the difficulty of the task which results in a large part of the sample performing at chance, and the lack of variability in test item difficulty. Together, these psychometric weaknesses lead to tasks tapping mainly error variance rather than variance related to SL capacity (see [12], for extended discussion, and possible solutions). Whereas this state of affairs did not hinder demonstrations of learning across a full sample of participants, they constitute a formidable obstacle to individual differences studies.

Structural confounds: At present, most SL tasks are based on a passive familiarization phase, in which stimuli representing a set of regularities are presented to participants (e.g., a continuous stream of shapes or syllables organized in pairs or triplets in visual and auditory SL, a sequence of "grammatical" sequences in AGL, etc.). Once the familiarization phase is over, it is followed by a test phase that estimates participants' learning of the statistical properties of the previously presented stream, typically through a series of 2AFC responses. We will refer to these measures as *offline measures of performance*, since they do not track the discovery of regularities from the stream while it unfolds, but attempt to assess the extent of learning once it is over.

The theoretical challenges that offline measures implicate are presented in Figure 1 which outlines the components of individual performance in the classical visual SL (VSL) task (e.g., [3,16,35,66,67]) is measured (see [68] for a related approach).

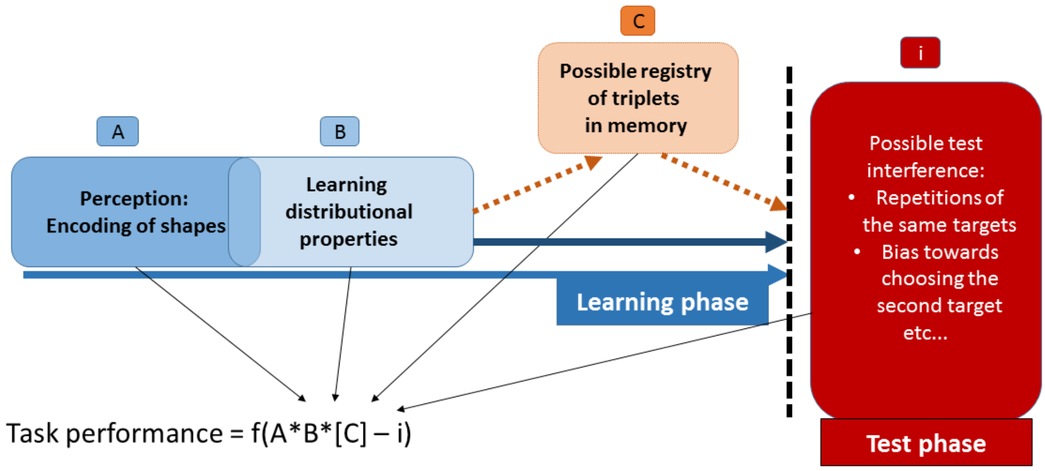

Figure 1. *The factors contributing to SL task performance, as measured by standard offline measures.*

As an example, consider a common variant of the VSL task, in which 24 abstract shapes are organized into eight triplets. During a familiarization phase these triplets are repeatedly presented in a continuous stream. The only source of information regarding the composition of the triplets in the stream lies in the transitional probabilities (TPs) between the shapes in the sequence: TPs between shapes within a triplet is 1, whereas TPs of shapes between triplets is 1/7, for 8 triplets without immediate repetition of a triplet. Following familiarization, the test phase begins. It consists of a series of 2AFC trials, each contrasting one of the triplets presented during learning with a "foil" – a group of three shapes that never appeared together in the familiarization phase (TPs=0). In each trial of the test, one foil and one triplet are presented, and participants are asked to decide which group of shapes appears more familiar, given the stream they have seen. The final score that represents SL individual ability is the number of correct responses in the test phase.

Figure 1 depicts a coarse-grained account of possible factors and processes underlying the final observed performance in the task. On the left side of the figure (in blue), we describe the processes involved in the familiarization phase, while on the right side (in red) we list several additional factors affecting the outcome of the test phase. Considering the computation of regularities, we note that participants first have to perceive and encode the individual elements of the stream (factor A). Since individuals substantially differ in the resolution of their perceptual system, their differential ability to generate perceptual representations under specific exposure constraints, would inevitably contribute to the variance of performance in the subsequent test-phase (see also [11,43]). While encoding the individual events, participants further have to discover the statistical regularities in the stream (factor B), which is, in our present context, the most central factor for SL research. These two factors in combination (A and B) may result in sufficient sensitivity to the statistical regularities to perform above chance in the test phase. In addition, some experiments include instructions or other demand characteristics that may lead participants to try and detect repeated patterns (here, triplets) and memorize them for later recollection. Although this type of strategy is not required for successful SL, such additional explicit memorization efforts will add yet another task component (factor C) with considerable individual differences (e.g., [69]; see [Gomez, this issue] for discussion).

Critically, *the underlying implicit assumption behind the use of offline measures is that the accurate signature of learning can be retroactively traced*, so that the test score would reflect the two, possibly three factors (A, B, and C) contributing to SL abilities, and these only. However, as we will argue, similar to many offline tests in other fields in cognitive science (e.g., [70]) , the testing phase in SL tasks inevitably interferes with what was learnt during familiarization, obscuring the ability to accurately measure the net SL ability. Here we label the interference component (factor i). This leads us to the following operationalization of measuring performance in the VSL task, using offline measures:

Task performance = f (A*B [*C] – i)

By this formula, performance in the VSL task would be some function of the multiplication of the ability to encode shapes (A), the ability to encode their co-occurrences (B), and, in some cases, the ability to store the extracted triplets (or patterns) in memory (C), minus (i) the degree of test interference. We opted for multiplication of the first two/three factors A, B, C, rather than simple additivity, since zero ability in any of the components (inability to encode shapes, inability to extract regularities, or inability to store items in memory) would inevitably result in zero learning.

Note that this operationalization applies not only to the VSL task, but can be generalized (with some obvious modifications given the exact task design) to other SL tasks involving offline measurements. It enables us, however, to explicate the critical shortcomings of this method to assess SL capacity.

Shortcomings of current offline measures*:*

The first problem that arises is that performance in offline tests intermixes encoding efficiency, learning statistical regularities, and possibly individual memory constraints. Since the offline test is administered only after the learning phase in which those processes occur, it cannot differentiate between the relative contribution of these factors to the final learning score. Naturally, one could dismiss this caveat arguing that SL capacity inherently reflects the joint contribution of these components. However, in terms of predictive validity, in order to theoretically tie SL performance to specific cognitive abilities, knowing where exactly a potential weakness lies (encoding, learning, or explicit memorization), is crucial. This is especially critical for an explanatory theory regarding how SL results in specific cognitive impairments, such as SLI, dyslexia, etc.

The second problem is that offline measures, being set at the end of learning, *do not provide any information regarding the learning dynamics across time*. Since no data are collected during the learning phase, offline measures simply miss a large part of the action (this is a key part of the motivation for the AGL-SRT task developed by [19,30]; see also [71] for an individual-differences SRT study). Learning dynamics are important for two reasons. First, they provide insights regarding the speed of learning (i.e., how fast a given individual is in picking-up the statistical properties of the input), in addition to the extent of his/her learning (i.e., how much of the underlying structure did he/she learn in a pre-defined time period). From a theoretical perspective, both speed and degree of learning are useful markers of a participant's SL ability. In addition, learning dynamics can provide valuable insights regarding the shape of the learning trajectory – for example, it can be used to examine whether knowledge is acquired gradually (reflected by a linear/logarithmic learning trajectory), or whether learning is characterized by a sudden burst in performance (i.e., step function).

The third problem is that *the post-hoc nature of the offline test inevitably introduces testing interference and confounds*. For example, to allow for sufficient test items and to improve the resolution of performance scores, patterns and foils are typically repeated throughout the test phase several times. These repetitions effects interfere with learning, thereby blurring the methodological separation between intended learning during familiarization, and unintended learning that occurs during the test phase[8]. It is

[8] The potential for learning during test has long been known in the implicit learning literature and thus a number of AGL studies have employed no-learning control groups to factor out potential effects of such learning on test performance (e.g., [4]). Note also that in some paradigms researchers have tried to mitigate the effect of learning/interference during the test phase by interleaving several tests with re-familiarization phases (e.g., in perceptional adaptation paradigms, see [85]).

impossible to know whether responses reflect information acquired during learning or of overriding information presented by the repeated test items (see [12], for discussion).

The promise of online measures

The main motivation for using online measures is to track learning throughout the familiarization phase as it unfolds, which alleviates most of the caveats introduced by offline measures. As such, online measures of SL carry the promise of better resolution on multiple levels: First, from a theoretical perspective, they can differentiate cognitive processes that relate to the perceptual encoding of input elements and the learning of their distributional properties, from processes that use this information during a subsequent test. This makes it possible to identify the contribution of each of these components to SL performance. Second, online measures provide information regarding learning dynamics, reflecting how *fast* each individual learns the statistical properties of a stream, as well as indicating his/her learning *trajectory*. Third, by gathering a maximal amount of information (by tapping the full learning session), and by avoiding the interference introduced by the test phase, online measures have the promise of higher 'psychometric resolution'- resulting in more reliable measurements.

Operationally, we define online measures as examining participants' responses throughout the learning process. A typical example is the classic SRT task, where implicit learning of a repeated sequence of digits is monitored. The online measure, the time taken to press a given key corresponding to a given digit, reflects the underlying assumption that faster motor responses are expected for predicted sequences compared with random ones. Since predicted events result in faster responses, the trajectory of learning can be traced in this task.

These principles, however, can be easily applied to classical SL tasks. Consider for example the above VSL task. A simple modification can be introduced into the task to yield useful online information (see [72], for an action-sequence version, and [73], for visual AGL). Rather than asking participants to passively watch a stream of visual shapes which appeared on the screen at a fixed rate of presentation, they are asked to advance the stream of shapes by themselves, at their own pace, by pressing the spacebar (much like in the self-paced reading paradigm, [74]). The assumption is that learning the transitional probabilities between shapes in the triplets will result in faster bar pressing for predicted shapes (second and third shapes of the triplet), relative to unpredicted shapes (the first shape of each triplet). This makes it possible to track the detailed time-course of learning. RT differences between predicted and unpredicted stimuli have also been demonstrated in other tasks in auditory [75] and audio-visual SL [19,30,32,76]. Importantly, online measures of SL have been found to correlate with sentence processing in L1 [19,32], providing preliminary evidence regarding its predictive validity.

But note that the development of online measures of SL still requires extensive research. First, it is yet to be shown whether the existing online measures of learning provide reliable measures of *individual* performance, since no studies to date have examined the reliability of such measures (see by contrast, the reliability coefficients of offline measures recently reported by [33]). Second, existing studies present mixed reports regarding the correlations between online and the standard offline measures of SL (high correlations reported in [73], but zero correlations reported in [19,75,77]). Low correlations between offline and online measures in the same task could reflect theoretical issues (e.g., tapping explicit vs. implicit knowledge, [78], or tapping different components of SL variance, [19]). However, such state of affairs might also be due to an inherent low reliability of online measures, either because they are unstable or inaccurate. A third issue in the development of online measures is that some online tasks may actually contaminate learning – for example, it was shown that in the SL click-detection paradigm (first proposed by [75]), the mere presence of clicks in the familiarization stream hinders learning due to its taxation on attentional resources [77]. These issues need to be resolved by further research if the promise of the higher-resolution online measures is to be realized in future SL studies.

Concluding remarks

The theoretical interest in SL originally emerged as potential domain-general alternative to domain-specific approaches to language. Rather than assuming an innate and modular human capacity for processing linguistic information, SL was offered as a general mechanism for learning and processing any type of sensory input. In line with this view, individual performance in SL tasks was systematically shown to correlate with an array of linguistic abilities. Here we have suggested that further advances in this research enterprise require a deep *mechanistic* understanding of the precise interrelationship(s) between linguistic performance and SL ability, where SL as a theoretical construct is unpacked, no longer treated as a unified “black-box” entity. On this view, empirical and modeling work should provide a-priori hypotheses regarding the set of computations that underlie the learning of specific statistical regularities, within different types of input, in different modalities, taking into account their neurobiological constraints. This will allow for clear and testable fine-grained predictions that tie particular linguistic (and potentially other cognitive) abilities to specific SL computations. In the same vein, different experimental tasks impose different constraints on learning, thereby implicating different learning mechanisms.

Transparent discussions regarding the specific computations involved in each SL task, its relations to other SL paradigms, and the strategies that learners might use to learn a given statistical structure are necessary for establishing the theoretical link between performance in the task, and the cognitive function it is supposed to predict. On the methodological level, such finer-grained hypotheses would call for more refined measures of SL, that track SL performance more directly, providing a richer set of data regarding the processes involved in SL. In line with these aims, the current paper offers a preliminary taxonomy of SL phenomena and outlines methodological guidelines, that can serve such future research.

Acknowledgements

This article was supported by the Israel Science Foundation (Grant No. 217/14, awarded to R.F.), and by the National Institute of Child Health and Human Development (Grant Nos. RO1 HD 067364, awarded to Ken Pugh and R.F., and PO1-HD 01994, awarded to Haskins Laboratories). L.B. is a research fellow of the Fyssen Foundation.

Authors' contributions

All four authors contributed to the writing of this paper.

References

1. Saffran, J. R., Aslin, R. N. & Newport, E. L. 1996 Statistical Learning by 8-Month-Old Infants. *Science (80-.).* **274**, 1926–1928. (doi:10.1126/science.274.5294.1926)

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

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

4. Conway, C. M. & Christiansen, M. H. 2005 Modality-constrained statistical learning of tactile, visual, and auditory sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 24–39. (doi:10.1037/0278-7393.31.1.24)

5. Pelucchi, B., Hay, J. F. & Saffran, J. R. 2009 Statistical Learning in a Natural Language by 8-Month-Old Infants. *Child Dev.* **80**, 674–685. (doi:10.1111/j.1467-8624.2009.01290.x)

6. Gebhart, A. L., Newport, E. L. & Aslin, R. N. 2009 Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychon. Bull. Rev.* **16**, 486–490. (doi:10.3758/PBR.16.3.486)

7. Newport, E. L. & Aslin, R. N. 2004 Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cogn. Psychol.* **48**, 127–162. (doi:10.1016/S0010-0285(03)00128-2)

8. Fiser, J. & Aslin, R. N. 2001 Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* **12**, 499–504. (doi:10.1111/1467-9280.00392)

9. Evans, J., Saffran, J. & Robe-Torres, K. 2009 Statistical learning in children with Specific Language Impairment. *J. Speech, Lang. Hear. Res.* **52**, 321–335.

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

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

12. Siegelman, N., Bogaerts, L. & Frost, R. 2016 Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behav. Res. Methods* (doi:10.3758/s13428-016-0719-z)

13. Misyak, J. B. & Christiansen, M. H. 2012 Statistical learning and language: An individual differences study. *Lang. Learn.* **62**, 302–331. (doi:10.1111/j.1467-9922.2010.00626.x)

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

15. Spencer, M., Kaschak, M. P., Jones, J. L. & Lonigan, C. J. 2014 Statistical learning is related to early literacy-related skills. *Read. Writ.* **28**, 467–490. (doi:10.1007/s11145-014-9533-0)

16. Frost, R., Siegelman, N., Narkiss, A. & Afek, L. 2013 What predicts successful literacy acquisition in a second language? *Psychol. Sci.* **24**, 1243–52. (doi:10.1177/0956797612472207)

17. Kidd, E. & Arciuli, J. 2015 Individual Differences in Statistical Learning Predict Children's Comprehension of Syntax. *Child Dev.* **87**, 184–193.

(doi:10.1111/cdev.12461)

18. Kidd, E. 2012 Implicit statistical learning is directly associated with the acquisition of syntax. *Dev. Psychol.* **48**, 171–184. (doi:10.1037/a0025405)

19. Misyak, J. B., Christiansen, M. H. & Tomblin, J. B. 2010 On-line individual differences in statistical learning predict language processing. *Front. Psychol.* **1**, 31. (doi:10.3389/fpsyg.2010.00031)

20. Mainela-Arnold, E. & Evans, J. L. 2014 Do statistical segmentation abilities predict lexical-phonological and lexical-semantic abilities in children with and without SLI? *J. Child Lang.* **41**, 327–51. (doi:10.1017/S0305000912000736)

21. Shafto, C. L., Conway, C. M., Field, S. L. & Houston, D. M. 2012 Visual Sequence Learning in Infancy: Domain-General and Domain-Specific Associations With Language. *Infancy* **17**, 247–271. (doi:10.1111/j.1532-7078.2011.00085.x)

22. Singh, L., Steven Reznick, J. & Xuehua, L. 2012 Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Dev. Sci.* **15**, 482–495. (doi:10.1111/j.1467-7687.2012.01141.x)

23. Conway, C. M., Karpicke, J. & Pisoni, D. B. 2007 Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with hearing adults. *J. Deaf Stud. Deaf Educ.* **12**, 317–334. (doi:10.1093/deafed/enm019)

24. Conway, C. M., Bauernschmidt, A., Huang, S. S. & Pisoni, D. B. 2010 Implicit statistical learning in language processing: Word predictability is the key. *Cognition* **114**, 356–371. (doi:10.1016/j.cognition.2009.10.009)

25. Hsu, H. J., Tomblin, J. B. & Christiansen, M. H. 2014 Impaired statistical learning of non-adjacent dependencies in adolescents with specific language

impairment. *Front. Psychol.* **5**, 1–10. (doi:10.3389/fpsyg.2014.00175)

26. Plante, E., Bahl, M., Vance, R. & Gerken, L. 2010 Children with specific language impairment show rapid, implicit learning of stress assignment rules. *J. Commun. Disord.* **43**, 397–406. (doi:10.1016/j.jcomdis.2010.04.012)

27. Gabay, Y., Thiessen, E. D. & Holt, L. L. 2015 Impaired statistical learning in developmental dyslexia. *J. Speech, Lang. Hear. Res.* **58**, 934–945. (doi:10.1044/2015)

28. Pavlidou, E. V., Kelly, M. L. & Williams, J. M. 2010 Do Children with Developmental Dyslexia Have Impairments in Implicit Learning? *Dyslexia* **16**, 143–161. (doi:10.1002/dys.400)

29. Christiansen, M. H., Louise Kelly, M., Shillcock, R. C. & Greenfield, K. 2010 Impaired artificial grammar learning in agrammatism. *Cognition* **116**, 382–393. (doi:10.1016/j.cognition.2010.05.015)

30. Misyak, J. B., Christiansen, M. H. & Tomblin, J. B. 2010 Sequential expectations: The role of prediction-based learning in language. *Top. Cogn. Sci.* **2**, 138–153. (doi:10.1111/j.1756-8765.2009.01072.x)

31. Perruchet, P. & Pacton, S. 2006 Implicit learning and statistical learning: one phenomenon, two approaches. *Trends Cogn. Sci.* **10**, 233–8. (doi:10.1016/j.tics.2006.03.006)

32. Misyak, J. B. & Christiansen, M. H. 2010 When ‘more’ in statistical learning means ‘less’ in language: Individual differences in predictive processing of adjacent dependencies. In *Proceedings of the 32nd Annual Cognitive Science Society Conference* (eds R. Catrambone & S. Ohlsson), pp. 2686–2691. Austin, TX: Cognitive Science Society.

33. Siegelman, N. & Frost, R. 2015 Statistical learning as an individual ability:

Theoretical perspectives and empirical evidence. *J. Mem. Lang.* **81**, 105–120.

34. Conway, C. M. & Christiansen, M. H. 2009 Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *Eur. J. Cogn. Psychol.* **21**, 561–580. (doi:10.1080/09541440802097951)

35. 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. *Q. J. Exp. Psychol.* **64**, 1021–1040. (doi:10.1080/17470218.2010.538972)

36. Redington, M. & Chater, N. 1996 Transfer in artificial grammar learning: A reevaluation. *J. Exp. Psychol. Gen.* **125**, 123–138. (doi:10.1037/0096-3445.125.2.123)

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

38. Schapiro, A. C., Gregory, E. & Landau, B. 2014 The Necessity of the Medial-Temporal Lobe for Statistical Learning. *J. Cogn. Neurosci.* **26**, 1736–1747.

39. Turk-Browne, N. B., Scholl, B. J., Chun, M. M. & Johnson, M. K. 2009 Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *J. Cogn. Neurosci.* **21**, 1934–1945. (doi:10.1162/jocn.2009.21131)

40. Tremblay, P., Baroni, M. & Hasson, U. 2013 Processing of speech and non-speech sounds in the supratemporal plane: auditory input preference does not predict sensitivity to statistical structure. *Neuroimage* **66**, 318–32. (doi:10.1016/j.neuroimage.2012.10.055)

41. Meyer, T. & Olson, C. R. 2011 Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl. Acad. Sci.* **108**, 19401–6.

(doi:10.1073/pnas.1112895108)

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

43. Bogaerts, L., Siegelman, N. & Frost, R. 2016 Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychon. Bull. Rev.* (doi:10.3758/s13423-015-0996-z)

44. Maye, J., Werker, J. F. & Gerken, L. 2002 Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* **82**, 101–111. (doi:10.1016/S0010-0277(01)00157-3)

45. Daikhin, L. & Ahissar, M. 2012 Responses to deviants are modulated by subthreshold variability of the standard. *Psychophysiology* **49**, 31–42. (doi:10.1111/j.1469-8986.2011.01274.x)

46. Goujon, A., Didierjean, A. & Thorpe, S. 2015 Investigating implicit statistical learning mechanisms through contextual cueing. *Trends Cogn. Sci.* **19**, 524–533. (doi:10.1016/j.tics.2015.07.009)

47. Curtin, S., Mintz, T. H. & Christiansen, M. H. 2005 Stress changes the representational landscape: Evidence from word segmentation. *Cognition* **96**, 233–262. (doi:10.1016/j.cognition.2004.08.005)

48. van den Bos, E., Christiansen, M. H. & Misyak, J. B. 2012 Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *J. Mem. Lang.* (doi:10.1016/j.jml.2012.07.008)

49. Romberg, A. R. & Saffran, J. R. 2013 All together now: Concurrent learning of multiple structures in an artificial language. *Cogn. Sci.* **37**, 1290–1320.

(doi:10.1111/cogs.12050)

50. Shye, S., Elizur, D. & Hoffman, M. 1994 *Introduction to facet theory: Content design and intrinsic data analysis in behavioral research.* Thousand Oaks, CA: Sage Publications.

51. Guttman, R. & Greenbaum, C. 1998 Facet Theory: Its Development and Current Status. *Eur. Psycholoist* **3**, 13–36. (doi:10.1027//1016-9040.3.1.13)

52. Eichenbaum, H. 2013 Memory on time. *Trends Cogn. Sci.* **17**, 81–88. (doi:10.1016/j.tics.2012.12.007)

53. Schvaneveldt, R. W. & Gomez, R. L. 1998 Attention and probabilistic sequence learning. *Psychol. Res.* **61**, 175–190. (doi:10.1007/s004260050023)

54. Robertson, E. M. 2007 The serial reaction time task: implicit motor skill learning? *J. Neurosci.* **27**, 10073–10075. (doi:10.1523/JNEUROSCI.2747-07.2007)

55. Arciuli, J., von Koss Torkildsen, J., Stevens, D. J. & Simpson, I. C. 2014 Statistical learning under incidental versus intentional conditions. *Front. Psychol.* **5**. (doi:10.3389/fpsyg.2014.00747)

56. Maddox, W. T., Love, B. C., Glass, B. D. & Filoteo, J. V. 2008 When more is less: Feedback effects in perceptual category learning. *Cognition* **108**, 578–589. (doi:10.1016/j.cognition.2008.03.010)

57. Palmer, S. D. & Mattys, S. L. 2016 Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Q. J. Exp. Psychol. (Hove).* , 1–12. (doi:10.1080/17470218.2015.1112825)

58. Adriaans, F. & Kager, R. 2010 Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *J. Mem. Lang.* **62**, 311–331. (doi:10.1016/j.jml.2009.11.007)

59. Frost, R. 2012 Towards a universal model of reading. *Behav. Brain Sci.* **35**, 263–279. (doi:10.1017/S0140525X11001841)

60. Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J. & Mackintosh, N. 2010 Implicit learning as an ability. *Cognition* **116**, 321–340. (doi:10.1016/j.cognition.2010.05.011)

61. Goujon, A. & Fagot, J. 2013 Learning of spatial statistics in nonhuman primates: Contextual cueing in baboons (Papio papio). *Behav. Brain Res.* **247**, 101–109. (doi:10.1016/j.bbr.2013.03.004)

62. Ahissar, M. 2007 Dyslexia and the anchoring-deficit hypothesis. *Trends Cogn. Sci.* **11**, 458–465. (doi:10.1016/j.tics.2007.08.015)

63. Szmalec, A., Page, M. P. A. & Duyck, W. 2012 The development of long-term lexical representations through Hebb repetition learning. *J. Mem. Lang.* **67**, 342–354. (doi:10.1016/j.jml.2012.07.001)

64. Poletiek, F. H. & Wolters, G. 2009 What is learned about fragments in artificial grammar learning? A transitional probabilities approach. *Q. J. Exp. Psychol.* **62**, 868–76. (doi:10.1080/17470210802511188)

65. Wechsler, D. 2008 Wechsler adult intelligence scale - Fourth Edition (WAIS-IV). *San Antonio*.

66. Glicksohn, A. & Cohen, A. 2013 The role of cross-modal associations in statistical learning. *Psychon. Bull. Rev.* **20**, 1161–1169. (doi:10.3758/s13423-013-0458-4)

67. Turk-Browne, N. B., Junge, J. A. & Scholl, B. J. 2005 The automaticity of visual statistical learning. *J. Exp. Psychol.* **134**, 552–564. (doi:10.1037/0096-3445.134.4.552)

68. Karuza, E. A., Emberson, L. L. & Aslin, R. N. 2014 Combining fMRI and

behavioral measures to examine the process of human learning. *Neurobiol. Learn. Mem.* **109**, 193–206. (doi:10.1016/j.nlm.2013.09.012)

69. Unsworth, N. & Engle, R. W. 2007 The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **114**, 104–132. (doi:10.1037/0033-295X.114.1.104)

70. Malmberg, K. J., Criss, A. H., Gangwani, T. H. & Shiffrin, R. M. 2012 Overcoming the Negative Consequences of Interference From Recognition Memory Testing. *Psychol. Sci.* **23**, 115–119. (doi:10.1177/0956797611430692)

71. Hunt, R. H. & Aslin, R. N. 2001 Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *J. Exp. Psychol. Gen.* **130**, 658–680. (doi:10.1037/0096-3445.130.4.658)

72. Baldwin, D., Andersson, A., Saffran, J. & Meyer, M. 2008 Segmenting dynamic human action via statistical structure. *Cognition* **106**, 1382–1407. (doi:10.1016/j.cognition.2007.07.005)

73. 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.

74. Just, M. A., Carpenter, P. A. & Woolley, J. D. 1982 Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* **111**, 228–238. (doi:10.1037/0096-3445.111.2.228)

75. Gómez, D. M., Bion, R. H. & Mehler, J. 2011 The word segmentation process as revealed by click detection. *Lang. Cogn. Process.* **26**, 212–223. (doi:10.1080/01690965.2010.482451)

76. Vuong, L. C., Meyer, A. S. & Christiansen, M. H. 2016 Concurrent learning of adjacent and nonadjacent dependencies. *Lang. Learn.*

77. Franco, A., Gaillard, V., Cleeremans, A. & Destrebecqz, A. 2014 Assessing segmentation processes by click detection: online measure of statistical learning, or simple interference? *Behav. Res. Methods* (doi:10.3758/s13428-014-0548-x)

78. Bertels, J., Franco, A. & Destrebecqz, A. 2012 How implicit is visual statistical learning? *J. Exp. Psychol. Learn. Mem. Cogn.* **38**, 1425–1431. (doi:10.1037/a0027210)

79. Perruchet, P. & Vinter, A. 1998 PARSER: A Model for Word Segmentation. *J. Mem. Lang.* **39**, 246–263. (doi:10.1006/jmla.1998.2576)

80. Giroux, I. & Rey, A. 2009 Lexical and sublexical units in speech perception. *Cogn. Sci.* **33**, 260–272. (doi:10.1111/j.1551-6709.2009.01012.x)

81. Orbán, G., Fiser, J., Aslin, R. N. & Lengyel, M. 2008 Bayesian learning of visual chunks by human observers. *Proc. Natl. Acad. Sci.* **105**, 2745–2750. (doi:10.1073/pnas.0708424105)

82. Thiessen, E. D. submitted. What's statistical about learning? Insights from modeling statistical learning as a set of memory processes.

83. Creel, S. C., Newport, E. L. & Aslin, R. N. 2004 Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* **30**, 1119–1130. (doi:10.1037/0278-7393.30.5.1119)

84. Zhao, J., Ngo, N., McKendrick, R. & Turk-Browne, N. B. 2011 Mutual interference between statistical summary perception and statistical learning. *Psychol. Sci.* **22**, 1212–9. (doi:10.1177/0956797611419304)

85. Konkle, T., Wang, Q., Hayward, V. & Moore, C. I. 2009 Motion Aftereffects

Transfer between Touch and Vision. *Curr. Biol.* **19**, 745–750.

(doi:10.1016/j.cub.2009.03.035)