Exploring and exploiting uncertainty

Divjak, Dagmar; Milin, Petar

DOI:
10.1111/cogs.12835

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Divjak, D & Milin, P 2020, 'Exploring and exploiting uncertainty: Statistical learning ability affects how we learn to process language along multiple dimensions of experience', *Cognitive Science*, vol. 44, no. 5, e12835. https://doi.org/10.1111/cogs.12835

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12835

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12835

Exploring and Exploiting Uncertainty: Statistical Learning Ability Affects How We Learn to Process Language Along Multiple Dimensions of Experience

Dagmar Divjak,[a] Petar Milin[b]

[a]*Department of Modern Languages & Department of English Language and Linguistics, The University of Birmingham*
[b]*Department of Modern Languages, The University of Birmingham*

Received 3 May 2019; received in revised form 15 December 2019; accepted 5 March 2020

Abstract

While the effects of pattern learning on language processing are well known, the way in which pattern learning shapes exploratory behavior has long gone unnoticed. We report on the way in which individual differences in statistical pattern learning affect performance in the domain of language along multiple dimensions. Analyzing data from healthy monolingual adults' performance on a serial reaction time task and a self-paced reading task, we show how individual differences in statistical pattern learning are reflected in readers' knowledge of linguistic co-occurrence patterns and in their exploration and exploitation of content-specific and task-general information. First, we investigated the extent to which an individual's pattern learning correlates with his or her sensitivity to systematic morphological and syntactic co-occurrences, as evidenced while reading authentic sentences. We found that the stream of morphological and syntactic information has a more pronounced effect on the reading speed of, as we will label them, content-sensitive learners in that the more probable the co-occurrence pattern, the faster their reading of that pattern will be. Next, we investigated how differences in pattern learning are reflected in the ways in which individuals approach the reading task itself and adapt to it. Casting this relation in terms of exploration/exploitation strategies, known from Reinforcement Learning, we conclude that content-sensitive learners are also more likely to initially probe (explore) a wider range of directly relevant patterns, which they can later use (exploit) to optimize their reading performance further. By affecting exploratory behavior, pattern learning influences the information that is gathered and becomes available for exploitation, thereby increasing the effect pattern learning has on language cognition.

Correspondence should be sent to Dagmar Divjak, Department of Modern Languages & Department of English Language and Linguistics, The University of Birmingham, Edgbaston, B15 2TT Birmingham, UK. E-mail: d.divjak@bham.ac.uk

Keywords: Exploration-exploitation; Individual differences; Morphology; Statistical learning; Syntax

1. Introduction

Over the past two decades, evidence has amassed documenting the extent to which language cognition depends on learning spatiotemporal regularities in a probabilistic manner (Jost & Christensen, 2017). Recent studies highlight the existence of individual differences within this ability (Kidd, Donnelly, & Christiansen, 2018) and discuss the possibility that statistical learning might not be a unitary learning mechanism but a set of domain-general computational principles that operate in different modalities (Frost, Armstrong, Siegelman, & Christiansen, 2015). While the effects of pattern learning on language processing are well described, the way in which pattern learning shapes exploratory behavior more generally has long gone unnoticed.

In this paper, we link research on the effects of pattern learning on language processing to research on the alignment of desirable learning strategies across tasks. We build on previous work by Milin, Divjak, and Baayen (2017) to demonstrate how individual differences in pattern learning correlate with skilled readers' handling of morphological and syntactic co-occurrence patterns and how these differences are reflected in their exploration and exploitation of content- and task-related patterns. More specifically, we test three hypotheses: learning non-linguistic and linguistic sequences relies on the same mechanism (Section 3.1); individuals differ in their pattern learning and this is reflected in how they read morphological and syntactic patterns (Section 3.2); depending on their pattern learning, individuals differ in their readiness to explore and exploit the environment in order to adapt to the task and improve their performance (Section 3.3). By affecting exploratory behavior, pattern learning influences the information that is gathered and becomes available to be exploited, thereby doubling the effect it has on performance within the domain of language.

1.1. Statistical learning, learning theory, and usage-based linguistics

Statistical learning[1] (SL) starts from the observation that organisms appear to extract probabilistic information from their environment automatically. Probabilistic information can reveal properties of events co-occurring in an environment, or of events occurring one after the other in time. Learning these patterns is assumed to happen "non-intentionally, without explicit hypothesis formation or testing, and certainly without conscious access to the kinds of computations that appear to produce the learning" (Williams & Rebuschat, 2012, p. 237). Indeed, the theoretical assumption shared by many (if not all) models of learning is that learners approximate probabilistic (stochastic) relationships in the environment through iterative tuning of associative links between objects and/or events, that is, cues and outcomes. One way or another, all formal learning traditions share this assumption, even though learning models differ in their views on how cues compete to gain their predictive potential, yielding an adaptive advantage (cf. Johnston,

1982; Rescorla, 1988). Statistical learning has been highly successful in predicting and explaining performance on a range of tasks requiring the learning of regularities across modalities. Implicit recognition and unintentional learning of structured patterns are considered to be general abilities in both animals and humans (for an overview, see Milne, Wilson, & Christiansen, 2018). For a recent and comprehensive overview of how accumulated evidence regarding the (implicit or statistical) processing of regularities in the environment is starting to shape and constrain theories of cognitive systems, we refer to the *Philosophical Transactions of the Royal Society B* (2017, volume 372, issue 1711).

Language cognition is no exception. Statistical learning captures the ability to encode and represent sequential input, using preceding context to draw inferences about upcoming units (cf. Conway, Bauernschmidt, Huang, & Pisoni, 2010). Language abounds with patterns. Detailed explorations of the distributional richness of language that could serve as input for learning have acted as impetus for usage-based "emergentist" approaches to language (Tomasello, 2003). Rather than coming to the language learning task with knowledge about the possible structures of human languages, learners might be endowed with an ability that would enable them to discover structure and probabilistic abstractions (i.e., grammar), based on their experience with language and guided by the probabilistic properties of linguistic input. Support for this position stems from the observation that adults' knowledge and usage of grammatical patterns are graded and probabilistic, and reflect the statistics of the language they experience (Bybee, 2010). For a summary discussion of the successes and challenges that this account of language cognition faces, see Wonnacott (2013).

What has made statistical learning particularly attractive to language researchers is the assumed domain-generality of the statistical learning ability. If experience-based probabilistic learning can be shown to be driven by pattern recognition abilities that are not unique to language, it can be invoked to argue against nativist, domain-specific accounts of language and cognition. In other words, a domain-general statistical learning mechanism obviates the need to posit innate mechanisms that specifically account for the human ability to acquire the basics of a highly complex communicative system in a relatively short time span. Theories of learning, too, have invested considerable effort to argue for one general principle of learning (cf. Bouton, 2007). Recent studies examining commonalities in learning distributional information have, however, revealed consistent modality and stimulus specificity. Frost et al. (2015) have therefore proposed that statistical learning might not be a unitary learning mechanism but rather a set of domain-general computational principles that operate in different modalities. Despite its apparent contradiction, the view that combines domain-generality with modal specificity finds support in the neurological literature on statistical learning. Schapiro and Turk-Browne (2015) discuss the repeated observation of activation of the same brain regions across tasks and interpret this as a sign that SL is at least partly a domain-general process. Hasson (2017) concludes, in a similar vein, that different sensory systems implement similar computations over inputs with different characteristics. While domain-generality explains why SL in one domain influences performance in another as we see when visual segmentation predicts language abilities, the differences in inputs trigger modality and stimulus specificity.

What remains in limited supply in the literature, yet what is crucial for usage-based theories of language cognition, is empirical evidence supporting an association between individual differences in implicit learning abilities on a non-language task and processing authentic language in context in healthy populations. Among the work that addresses this question directly, the following papers stand out for their care to avoid language in the pattern learning task, and their concern for the contextual embedding of the linguistic patterns in the language processing task. Conway et al. (2010) made participants view and reproduce a sequence of colored squares and listen to audio recordings of poor quality to find out whether they would find it easier to infer that "mop" was missing in the expected statement "wash the floor with a mop" than to infer that "tent" was missing in the unexpected combination "we rode off in our tent." They found that implicit learning abilities are essential for acquiring long-term knowledge of the sequential structure of language, that is, knowledge of word predictability. Williams and Rebuschat (2012) adapted the self-paced reading (SPR) task to present both non-linguistic and linguistic stimuli which focused on word order regularities in German in L2 learners, for example, *Yesterday scribbled David a letter to his brother* versus **Scribbled yesterday David a letter to his brother*. They found that response times were longer overall in the ungrammatical than the grammatical sequences at the points at which the sequences were violated. Kidd (2012) presented children aged 4 to 7 with a 10-sequence pattern involving a cartoon character, the position of which they had to learn to predict. The linguistic task consisted of a priming experiment with full be-passives in English. He found implicit statistical learning ability to be directly associated with the long-term maintenance of the primed structure. Frost, Siegelman, Narkiss, and Afek (2013), finally, provided evidence that (visual) pattern recognition abilities correlate with attainment in L2 assimilation of the Semitic structure of Hebrew words. They employed a visual statistical learning task, monitoring their participants' implicit learning of the transitional probabilities of visual shapes in a continuous stream. Their linguistic task used a cross-modal morphological priming paradigm from auditory presentation to written recognition to test subjects' sensitivity to Hebrew word structure. Native English speakers who were better able to learn conditional relations between shapes generally performed better at picking up Hebrew word morphology.

On other linguistic tasks, however, evidence supporting a link between (visual) pattern learning and language processing is less equivocal. Reading is a case in point. Arciuli and Simpson (2012) found a relation between pattern learning as measured by a visual embedded triplet paradigm and reading single words out loud in English in children and adults. Yet, Schmalz, Moll, Mulatti, and Schulte-Körne (2019) found no evidence of a relationship between pattern learning as measured by a serial reaction time task or an artificial grammar learning task and reading single words out loud in German.[2] Our study will analyze the relationship between performance on a visual serial reaction time task and on a SPR task, which requires word-by-word reading of naturalistic sentences. The rationale for these tasks is provided in Section 2.2. below.

1.2. Statistical learning and individual differences

Older studies on statistical learning tended to focus on the ability of the average participant (Siegelman & Frost, 2015). Two factors played a role in this methodological or, rather, epistemological decision. On the one hand, there was Reber's (1993) explicit statement that profound individual differences in statistical learning ability are unlikely to be detected, due to the ancient nature of this type of learning. This would imply that any observed differences between individual language users would be negligible or purely random. On the other hand, and less specific to statistical learning, there was the general tendency of experimental approaches to treat individual differences as secondary to coarser but typically more robust group-based differences in experimental designs. Yet it has been argued that individual differences would usefully constrain cognitive theories (Vogel & Awh, 2008) by acknowledging that meaningful (i.e., non-random, systematic) individual variation exists in cognitive processes. Statistical learning is no exception. Hunt and Aslin (2001) were the first to report that, although all 10 participants showed improvement on a sequence learning task, seven showed consistent learning of bigram structure, while three seemed to have learned trigrams. Sensitivity to particular kinds of statistical regularities (i.e., adjacent or non-adjacent regularities) in artificial grammars is also predictive of processing ability for different types of sentence constructions. The better a participant was at learning adjacent dependencies in a statistical learning task, the more interference he or she experienced when processing information that conflicted locally (Misyak & Christiansen, 2007). Conversely, those participants who were better at learning non-adjacent dependencies in the statistical learning task were also better at processing linguistic long-distance dependencies (Misyak, Christiansen, & Tomblin, 2010). Kidd et al. (2018) present an extensive overview of individual differences in language acquisition and processing, showing that such differences are pervasive across the linguistic spectrum.

To rule out alternative explanations, for example, that both statistical learning ability and performance on an (artificial) grammar learning task would be correlated with a third variable, studies typically measure general cognitive abilities such as intelligence or working memory and partial out the effects of these abilities statistically. The results on possible confounds have not been equivocal. Siegelman and Frost (2015) were among the first to run a large-scale study involving a battery of statistical learning tasks in the visual and auditory modalities, using verbal and non-verbal stimuli, with adjacent and non-adjacent contingencies. Statistical learning tasks were administered along with general cognitive tasks in a within-subject design at two points in time. They found that statistical learning, as measured by some tasks,[3] is a stable and reliable capacity of an individual. Yet it is not a unified capacity; hence, individual sensitivity to conditional probabilities is not uniform across the visual and auditory modalities, nor across verbal and non-verbal stimuli. Moreover, statistical learning was found to be independent of general cognitive abilities. For this reason, general cognitive abilities are not included in the present study. Instead, we focus on the effect that an individual's statistical learning has on aspects of learned linguistic performance. Specifically, our study builds on previous work (Milin

et al., 2017) that links linguistic performance to the exploration/exploitation hypothesis from Reinforcement Learning (RL).

1.3. Probabilistic reinforcement of explorative and exploitative behavior

Reinforcement Learning emerged as a highly interdisciplinary field intersecting machine learning, psychology, and neuroscience. Yet it originated in experimental psychology, in the early work of Thorndike (1911/1965) and Skinner (1938/1990) on instrumental (operant) conditioning. In RL, individuals can show differences in their respective learning rates or, somewhat less obviously, in their "commitment" to an outcome (target) of learning, in terms of its rewarding characteristics (cf. Gallistel & Gibbon, 2000; Luzardo, Alonso, & Mondragón, 2017). The discrimination of rewarding behavior emerges over time and, eventually, yields an optimal set of actions (Gureckis & Love, 2015; Niv, 2009). An important determinant of such long-term adaptive behavior relates to explorative and exploitative actions; at first, a system (be it an organism or computer) needs to explore and exhaust the possibilities to discover those that are desirable, so as to be able to exploit them and benefit later on.

The observation that behavior might have a longer-term desirable effect forms the basis of a key explanatory framework for animal and human learning and adaptation (cf. the "law of effect" in Thorndike, 1911 and later work by Hull, 1935). For learnable aspects of language behavior, specifically, one question seems to be of crucial importance: Do particularly successful learning strategies show up in different guises across learning contexts? Previous work (Milin et al., 2017) has provided indirect evidence for the alignment of desirable learning strategies that encourage pattern learning across tasks. Participants' engagement with a (reading) task correlated with the mental speed with which they approached a sequence learning task and the information they had at their disposal. The change in participants' behavior across experimental trials involving random and patterned sequences was predictive of participants' language processing as indexed by word reading latencies. Those readers who were more challenged by the sequential responses task (interpreted as evidence of a lower mental speed) were the most reliant on visual orthographic input in the self-paced sentence reading task. Those who were less challenged (interpreted as having a higher mental speed) did not need such orthographic support at all. At the same time, learners with higher mental speeds showed evidence of an exploration-exploitation strategy. Their reading behavior was more variable at the beginning (exploration) of the task and less variable later on (exploitation).

Results from across a divergent set of experimental tasks have been successfully interpreted in terms of earlier exploration and later exploitation. Gureckis and Love (2009) relied on explorative-exploitative behavior to account for the appreciation of delayed (but overall higher) rewards in a decision-making task with conflicting short- and long-term rewards. Stafford and Dewar (2014) reported that game players who exhibit greater initial variability in performance achieved a higher overall score. Hayes and Petrov (2016) showed evidence of exploration-exploitation in eye pupil dilation during an analogical reasoning task. Gameiro, Kaspar, König, Nordholt, and König (2017) viewed the trade-

off between the number and length of fixations in eye-tracking data as, essentially, a trade-off between explorative and exploitative viewing behavior. Finally, applying the ideas of Stafford and Dewar (2014) and Milin et al. (2017) to the study of English compounds, Schmidtke, Gagné, Kuperman, and Spalding (2018) argued that the diversity of relational knowledge might be reflecting the "explored" space of conceptual knowledge, to enrich the semantic possibilities and "exploit" them by achieving greater semantic precision.

These results suggest that the exploration-exploitation hypothesis from RL is attractive as a framework for discussing the individual differences in adaptive behavior observed in experimental settings, including settings that focus on language processing (for further discussion see Gureckis & Love, 2015; Mehlhorn et al., 2015). Here, in particular, we propose a parallelism between the specifics of the exploration/exploitation hypothesis and, more generally, the dyad of learning and performance. Learning and performance do not exist without one another: We gradually learn to perform a task in an (presumably) ever more optimal way, and each performance creates new opportunities to learn and adapt further to our environment. This constitutes the core of our understanding of the exploration/exploitation hypothesis (see also Milin et al., 2017), which we expand here to highlight the synergy between the learning (exploration) and performance (exploitation) components of the full adaptation cycle.

Specifically, we use RL to model the incongruity between the computational demands of tracking the statistics of a more stationary system like language, represented by the experimental stimuli, and those of a continuously changing world, in this case, the experiment. This incongruity needs to be integrated and turned into coherent information with a certain adaptive value, such that the performance on the task at hand can become sufficiently efficient. The diverging demands are modeled in RL as two types of uncertainty: the *known uncertainty* of a set of potential outcomes given the current stimulus and the *unknown uncertainty* which reflects the lack of veridical knowledge about what the contingency structure of the environment actually is (Yu & Dayan, 2005). We link these two systems and the associated levels of uncertainty back to individuals' pattern learning ability to develop understanding of the many dimensions along which this ability affects performance.

1.4. The current study

We test three interrelated hypotheses that, taken together, show how individual differences in pattern learning ability surface as differences in processing language at the local sentence level and morph into radically different learning styles in the global learning context.

First, statistical learning is a general learning mechanism that supports the learning of both non-linguistic and linguistic sequences (cf. Milne et al., 2018). This pattern learning mechanism is different from task adaptation, which is attested even for random sequences (Section 3.1).

Second, individuals differ in their pattern learning, and this is reflected in their language processing. For learners focusing on the information content (hereafter: content-sensitive learners), more than for learners focusing on the task (hereafter: task-sensitive learners), the stream of morphological and syntactic information will have a more pronounced effect on reading speed in that the more probable the co-occurrence pattern, the faster their reading of that pattern will be (Section 3.2).

Third, individuals differ in their pattern learning, and this is reflected in their readiness to explore and exploit the environment in order to adapt. In that respect, content-sensitive learners will be more likely to initially probe (explore) a wide range of morpho-syntactic possibilities, for later use (exploitation) in optimizing their reading (Section 3.3).

2. Methods

In this paper, we focus specifically on testing whether individual differences in pattern learning are directly related to individual differences in linguistic processing and whether pattern learning affects performance on multiple dimensions of a given language-related task. To this end, we ran a sequence of two experimental tasks.[4] To measure pattern learning in the non-linguistic domain, we employed a serial reaction task (SRT, described in Section 2.2.1). Participants' knowledge of linguistic patterns was measured using a SPR task (described in Section 2.2.2).

2.1. Participants

The data stem from a homogenous group of highly educated native speakers of Russian who, crucially, did not significantly differ in self-reported reading habits. We recruited 39 (17 males, 22 females) adult native speakers of Russian, aged between 18 and 31 (M = 23.6, s = 3.3), and living in St. Petersburg, Russia. Participants appeared healthy, did not report any reading disabilities or cognitive impairments, and had normal or corrected-to-normal vision. They were not linguists, philologists, or language students, and except for three (one female, two males) who had left school aged 18, they were educated to degree level. Personal data on age, gender, place of birth and current residence, location and type of schooling, occupation, knowledge of foreign languages, and reading habits were collected. Participants' identities were anonymized, and a unique numeric code was used throughout the analyses.

The sociodemographic, educational, and reading differences did not play any role in explaining reading time latencies and, hence, were not considered in the final statistical models; this is unsurprising given the size and homogeneity of the sample, and no further conclusions should be drawn from it.

2.2. Data collection and preprocessing

The experiments were run in a quiet room at the Institute for Linguistic Studies of the St. Petersburg branch of the Russian Academy of Sciences. Participants provided personal

information using Google forms prior to attending. They had also been sent information sheets and consent forms and were offered the opportunity to read those again and ask any questions at the testing location where the documents were signed and handed over to the experimenter. Participants were informed that participation was voluntary and that they could quit at any time. Those who completed the experiment were paid the equivalent of 5 GBP for participating. Ethical approval was granted by the School of Language & Cultures, University of Sheffield, UK.

Both tasks were administered using the same equipment, which reduces the risk of differential task effects affecting our findings. The tasks were programed and executed in PsychoPy (v. 1.78, released August 2013: Peirce, 2007, 2009), and a Cedrus response pad RB-540 (designed for spatial orientation experiments) was used to collect participants' responses. All experiments were run on the same laptop, an Intel i5 core Windows machine with Nvidia graphics card, 15" screen, and Windows 7. All participants completed the SRT task before the SPR task, in individual sessions.

2.2.1. Serial reaction time task

To measure pattern learning, we used a variant of the multi-choice, disjunctive serial reaction time task (Willingham & Goedert-Eschmann, 1999), which assesses improvements to immediate memory span for statistically consistent, structured sequences. The serial reaction time task has been used extensively in research on sequential pattern learning ability (Kaufman et al., 2010).

Stimuli and procedure: The task unfolded as follows. A star appeared on the screen in one of four positions (Up = U, Right = R, Down = D, Left = L), and subjects were asked to press the corresponding position on the response pad as quickly as possible. After a practice session of 72 random trials, subjects entered a training block of in total 288 trials in which a 12-trial sequence was repeated. The 12-trial repeating sequence consisted of a star appearing U, D, L, R, U, L, R, D, R, U, D, L. Overall, the training block contained four blocks with six repeats of the 12-trial sequence. After the training block, subjects entered a test block with 72 trials: twice, 24 random trials were followed by 12 test trials. Subjects were not alerted when they moved from one phase into the next. At debrief, some participants seemed aware that they had been exposed to a pattern, but no one was able to recall it explicitly nor use linguistic labels to name the positions (this is in line with recent findings from Medimorec, Milin, & Divjak, in press).

Data preprocessing: An analysis of the response accuracy showed that participants were very accurate overall, yielding correct responses in 98.7% of trials; 35 out of 40 achieved an accuracy of 98% or more, three achieved 97% accuracy, and one achieved 96% accuracy. The one participant who achieved only 94% accuracy was excluded from the analysis for performance on the SPR task (for details see Section 3.2). A density plot of response time latencies revealed the presence of some outliers (both short and long). We removed 0.5% from both extremes (173 data points in total) and log-transformed and scaled the remaining latencies to obtain a symmetric, Gaussian-like distribution.

2.2.2. Self-paced reading task

To capture language processing, we used data from a SPR paradigm without placeholders, whereby one word at the time appears in the center of the screen and is replaced by the next word at the press of a button. The SPR task has been used extensively in research on predictive processing in language. Both the SRT and the SPR tasks are in essence visual, but they differ in two crucial respects. Whereas the SRT task measures "meaningless" pattern recognition in space, the SPR task captures the processing of "meaningful" pattern recognition in language. The SRT task thus aims to measure how well participants pick up on patterns they encounter for the first time during the experiment, while the SPR task measures how well participants deal with patterns they will have been exposed to prior to the experiment. Both tasks do, however, allow for learning to take place over the course of the experiment, in the sense that participants can get used to the task itself. Here, participants can explore and exploit which behavior, considering their inclinations (a unique combination of abilities and experience), ought to yield the highest reward.

Stimuli and procedure: Existing corpus research had provided a rich knowledge base concerning verbs of perception, and this was used to select stimuli in which there were no known confounds (for details of the data we refer to Divjak, 2015). From the Divjak (2015) perception dataset, we selected 24 authentic corpus sentences, eight for each of the three active perception verbs, *smotret'* (look), *slušat'* (listen), and *trogat'* (touch). The (Russian) stimuli and their translation are provided in our data repository, these sentences were vetted by native speakers of Russian during the pilot phase.

Each verb was used in a typical morphological form in half of these sentences, and in an atypical morphological form in the other half. Form typicality was established on the basis of probabilities from a multinomial regression model run using the **polytomous** package (Arppe, 2013) in **R** with the one-versus-all heuristic. The variables aspect, mood, number, person, and negation were used to predict the choice of one of the three lexemes over the other two. This yielded probabilities for lexeme selection dependent on morphological form; variation along other dimensions was controlled for statistically. Form (surface) frequencies for each of the perception verb forms were extracted from a subcorpus of texts from the Russian National Corpus, consisting of material created between 1990 and 2013, which approximated the age of the participants; the subcorpus contained 101,678,022 words in total.

In addition to controlled variation of the verb form, we made use of naturally occurring variation in verb position. To achieve precision in accounting for this source of variation, we differentiate between the position of the verb in the construction, and for the position of the verb in the sentence. Perception verb constructions are typically three-element constructions with a subject, a verb, and sometimes also an object. The relative word order freedom that is typical for Russian allows the verb to occupy first, second, or third position within the construction; we will call these positions Early, Middle, and Late. The second source of positional variation—the place that the verb occupies in the sentence—shows that perception verbs in our sample of sentences occur in positions 1, 2,

3, 4, 6, 8, 9, and 13. Relatively speaking, these two sources of variation in verb position are related "hierarchically": the verb is positioned in the construction and in the sentence. In other words, verb is nested in construction which is nested in sentence. Both sources of information about verb position are relevant for reading (and comprehension) as the verb receives more or less contextual support depending on its position in the construction and depending on the position of the construction in the sentence.

Every participant saw all sentences containing perception verbs. These were interspersed with 24 filler sentences containing verbs expressing attempt. Implementing standard precautions for keeping participants attentive throughout the experimental session, a third of all sentences was followed by a yes/no question that the participants had to answer. Prior to the presentation of the experimental trials, the participants were familiarized with the procedure in five practice trials. The presentation of trials was randomized for each participant.

Data preprocessing: To facilitate statistical modeling, we log-transformed the position of the perception verb in the sentence (VerbPosition) and then scaled the transformed values to standard units of deviation (i.e., z-scores). The order of the experimental trials in a given session (TrialOrder) was also scaled but did not require log-transformation as it appeared symmetrically bell-shaped. As suggested by the Box-Cox test (**R** implementation in the **car** package by Fox et al., 2017), the reading time (RT) latency distribution for the perception verb was also log-transformed. Two sentences with *touch* were removed because RT data were lost due to experimenter error and software malfunction. Following Baayen and Milin (2010), we applied a minimal a priori trimming strategy, removing only unambiguously discontinuous data points, that is, those that are visibly extremes, leaving a solid gap between themselves and the data mass.

2.3. Extraction of learning indicators

In a final preparatory step, indicators of learning were extracted from both the SRT task and the SPR task.

2.3.1. Indicators of learning and adaptation from the serial reaction time task

Non-linguistic structured sequence learning was measured using a SRT task, which differentiates between random and structured response sequences. We fit mixed effects linear regression models to the response time latencies on the random and the sequenced trials of the SRT task separately, using the **lme4** package (Bates, Maechler, Bolker, & Walker, 2015) in the **R** software environment (R Core Team, 2017). For both models, the main covariate was TrialOrder (log-transformed and scaled) while the random effects part of both models consisted of by-participant intercept and slope adjustments. These intercept and slope adjustments were extracted to be used as our indicators of individual differences in pattern learning, yielding four indicators of individual differences per participant: srtRandomIntercept, srtRandomSlope, srtPatternIntercept, and srtPatternSlope. Each by-participant adjustment provides a single scalar value for every individual participant.

The adjustments from the sequenced parts served as indicators of individual differences in pattern learning, while the adjustments from the random parts were taken to be indicative of an individual's adaptation to the task at hand, that is, the increase in speed in pressing a button upon seeing a star appear on screen due to improved manual dexterity, independent of any (possibility for) pattern learning (given the randomness of these trials). Such separate treatment allows for a straightforward prediction: for random sequences, we expect a decrease in response latencies over experimental trials, which reflects the experience participants gained with the task; for structured sequences, however, we expect both a decrease in response latencies as a consequence of experience and a further decrease in latencies (i.e., a steeper change) that would reflect the effect of learning the structured sequences. In other words, some adaptation to the task is expected even when the patterns are random, but a more pronounced adaptation is predicted when pattern learning takes place.

2.3.2. Indicators of learning and adaptation from the self-paced reading task

Language-specific structured sequence learning is evidenced by knowledge of linguistic patterns as measured through word-by-word reading of naturalistic sentences in a SPR task. In this type of task, the position that the target word occupies in a sentence (i.e., VerbPosition) plays a pivotal role, with readers tending to speed up as they make their way through the sentence (cf. Kuperman, Dambacher, Nuthmann, & Kliegl, 2010). At the same time, as readers work their way through a sentence, they accumulate more information. This implies that the target verb's position in a sentence is a potential source of collinearity with other variables designed to capture aspects of information release. To disentangle the effects of the task's setup from those of its contents, we ran a two-step statistical procedure that gives us a better grip on the effects each dimension has on (self-paced) reading.

First, we fit a mixed effects linear regression model to the reading latencies, using the **lme4** package (Bates et al., 2015) in the **R** software environment (R Core Team, 2017). The dependent variable was reaction time latency, which had been log-transformed to facilitate statistical analysis (as suggested by the Box-Cox test). VerbPosition in the sentence (log-transformed and scaled) was the main covariate of the statistical model, which additionally included by-participant random intercepts and by-participant slope adjustments for VerbPosition. This model was thus kept maximally similar to the one run on the SRT data, and it differed only in terms of indicator (covariate) of learning: TrialOrder in SRT and VerbPosition in SPR. From the SPR model we extracted the random slopes (sprSlope) for use as individual indicators of linguistic sequence learning. Intercept adjustments were excluded from further analyses because those are indicative only of an individual's baseline speed for a given task (this measure did not contribute to the model fit in Milin et al., 2017, either).

3. Results

In this section, we present statistical models that address the three hypotheses presented in Section 1.5. Section 3.1 presents evidence for the claim that learning non-linguistic and linguistic sequences relies on the same mechanism. Next, we pursue an in-depth understanding of the effects that an individual's learning abilities as well as explorative/exploitive learning strategies have on reading performance. In Section 3.2, we analyze the effects of pattern learning, as measured with an SRT task, on the processing of linguistic patterns as evidenced in reading. In Section 3.3, we proceed to modeling how different types of pattern learners approach the SPR task itself to facilitate their individual differences in performance.

3.1. Components of pattern learning: Distinguishing learning from adaptation

One of the challenges posed by experimentally collected data on learning concerns the apparent conflation of the effects of practice with the task and of learning of new information on performance. Within the SRT task, this problem is usually avoided by contrasting overall performance on random and structured parts of the task. To get a better grip, and to track actual learning (vs. mere practice) across tasks, we operationalized individual pattern learning ability using the by-participant adjustments for the slope from three linear mixed effects models (as described in Section 2.3).

Three indicators of individual differences (srtRandomSlope, srtPatternSlope, sprSlope) were subjected to a Principal Component Analysis (cf. Cattell, 1966; Harris, 1985) alongside VerbPosition, which captures information flow. PCA captures the structure of the relationships between variables. It allows indicators with a common latent source (e.g., specific task-related effects such as the position of the verb in the sentence or physical experience handling the button box) to be grouped and to be kept separate from indicators that have different sources such as pattern learning effects and information effects (e.g., by the time a verb is encountered that occurs later in the sentence, more information will have been revealed). The resulting four principal components are summarized in Table 1.

In decreasing order of importance (or explained variance), the four principal components capture the increase in speed due to pattern learning (PC1 = Learning), the increase in speed solely due to the accrual of experience with random SRT trials (PC2 = Mechanics), the position of the verb in the sentence (PC3 = Position), and the differentiating component between the otherwise similar aspects of two types of structured pattern tasks —sequenced non-linguistic SRT trials and SPR sentences, respectively (PC4 = Pattern Differentiation).[5]

Crucially, the slope adjustments from both SRT and SPR tasks load onto PC1, the first and strongest component, which highlights their similarity. The fact that learning is shared across the SRT and SPR tasks confirms that learning non-linguistic and linguistic sequences is related. This relationship signals the domain-generality of statistical learning.

Table 1
Principal component loadings and proportion of variance explained by a principal components analysis on four variables. The table lists only the high, significant loadings (i.e., >0.3)

Variable	PC 1	PC 2	PC 3	PC 4
VerbPosition			**0.994**	
srtPatternSlope	**0.710**			**0.694**
srtRandomSlope		**−0.971**		
sprSlope	**0.703**			**−0.688**
Explained variance (%)	0.355	0.255	0.250	0.140

PC2 shows that learning any type of structured sequences is not related to mere task adaptation nor to word position.

3.2. Individual differences in pattern learning and their effect on language processing

The full dataset, consisting of participant information, measures of pattern learning, SPR latencies, traditional lexical predictors such as word frequency and length, and the linguistic predictors mentioned above, was submitted to generalized additive mixed modeling (GAMM; for a short introduction in the context of modeling behavioral data, see Baayen, Vasishth, Kliegl, & Bates, 2017), using the **mgcv** (Wood, 2006, 2011) and **itsadug** (van Rij, Wieling, Baayen, & van Rijn, 2016) packages in the **R** software environment (R Core Team, 2017). Fitting an additive model that allows for non-linear trends and non-linear interactions takes justification in neuroscientific findings where evidence has accrued to show that the effect of uncertainty should not be assumed to follow a simple linear trend (cf. Hasson, 2017, p. 6).

The mean reading latency for the perception verb was 804.32 ms (*Mdn* = 674.26 ms, *s* = 384.51 ms). To avoid having overly influential values distort the results, we removed one participant with an extremely steep decelerating trend in the SRT task (i.e., their slope adjustment was discontinuous with the rest of the sample). This resulted in the loss of 19 data points, leaving 794 for analysis (*M* = 799.42; *Mdn* = 668.03; *s* = 379.02). Participant information and traditional lexical predictors appeared non-significant for all models and were not considered further. This was also the case for all linguistic predictors, except the position of VerbInConstruction.

As shown in Table 1 above, of the four principal components only the first (PC.Learning) and the fourth (PC.PatternDifferentiation) showed more than one significant loading. The second (PC.Mechanics) and the third (PC.Position) component appeared as singletons and we reverted to using the original underlying variables in further analyses. Ultimately, only PC.Learning and VerbPosition turned out to be significant predictors of reading latencies.

We consistently applied model criticism (cf. Baayen & Milin, 2010) by removing data points with absolute standardized residuals exceeding 2.5. After removal of these data points, the models were re-fitted. All effects remained robustly significant. In fact, they

became more pronounced, which confirms that the findings were not the result of distortion by overly influential outliers.

Table 2 summarizes the final model over the full (i.e., untrimmed) dataset, for reading latencies on the perception verb. The GAMM fitted to the reading latencies for the perception verb contained a three-way non-linear interaction (i.e., a tensor product) of the Principal Component of Learning (the increase in speed due to learning in the SPR task and on the structured trials in the SRT task), VerbPosition, and the additional three-level factor which captured the position of the VerbInConstruction (with levels: Early, Middle, Late). In addition to this tensor product, there was a significant by-participant factor smooth for Trial. This factor smooth captures differences between participants and introduces the intertrial dependencies in the reading latency time-series over the sequence of experimental trials to which a participant is exposed (cf. Baayen et al., 2017; Milin et al., 2017). Because of the inclusion of by-participant factor smooths over Trials, the random effects for sentence or verb appeared superfluous ($p = .34$). SupMat B contains details of a simulation that corroborates the findings.

The effect of the tensor product is visualized in Fig. 1 that consists of three panels. The left panel depicts the interaction of Learning and VerbPosition for verbs that occur early in the perception construction (i.e., the verb is in first position and thus precedes the subject and object). The middle panel shows the interaction for verbs that occur in the middle of the construction (i.e., the verb occurs as second word, between subject and object). The right panel visualizes the interaction for verbs that occur as third and last word in the construction (i.e., the verb occurs after the subject and object). The Principal Component of Learning is depicted on the horizontal axis, with the most content-sensitive learners on the left. The vertical axis represents the position the verb occupies in the sentence (VerbPosition), with verbs that occur early on in the sentence at the bottom and verbs that occur late in the sentence at the top. The colors on the plot encode reading

Table 2
Generalized additive mixed model fitted to the reading latencies for the perception verb, reporting parametric coefficients (Part A) and effective degrees of freedom (edf), reference degrees of freedom (*Ref.df*), *F*- and *p*-values for the non-linear terms, tensor products, and random effects (Part B; AIC = −18.422, fREML = 47.09, adjusted $R^2 = .697$)

A. Parametric coefficients	Estimate	*SE*	*t*-value	*p*-value
Intercept	−0.316	0.059	−5.403	<.001
VerbInConstruction: Middle	−0.030	0.027	−1.089	.277
VerbInConstruction: Late	−0.007	0.020	−0.281	.779

B. Smooth terms	*edf*	*Ref.df*	*F*-value	*p*-value
te(PC.Learning, VerbPosition) VerbInConstruction: Early	3.995	4.629	8.337	<.001
te(PC.Learning, VerbPosition) VerbInConstruction: Middle	4.339	5.113	1.797	.105
te(PC.Learning, VerbPosition) VerbInConstruction: Late	3.808	4.380	4.261	.001
s(Scaled Trial, Participant)	35.264	340.000	5.117	<.001

latencies, with ochre/yellow signaling longer reading latencies, that is slower reading, green average reading latencies, and blue short latencies.

In the left panel, picturing perception verbs that precede their subject and object, that is perception verbs that occur early in the construction, we see those pattern learners who show the most dramatic change in reading times across experimental trials: they are slow when reading these early perception verbs if the construction occurs early in the sentence, but they speed up as the early perception verb appears in a construction toward the end of the sentence, and hence more context precedes. We will call these readers content-sensitive pattern learners.

The middle panel sees a new pattern emerge (although the interaction is barely marginally significant: $p = .105$), which stabilizes and becomes strong in the right panel ($p = .001$). In other words, this new pattern occurs when the perception verb occurs in between the subject and the object, and it stabilizes and receives strong statistical support when the perception verb follows its subject and object.

For the content-sensitive learners, the overall pattern of effects remains the same: the later the verb occurs in the sentence, the shorter the reading latency (roughly, *X*-axis values between −2 and −1 show changes from peach over green to blue along the *Y*-axis). This downward trend, however, becomes steeper as the position of the verb in its construction also becomes later (visible from the left to the right panel in Fig. 1). As the position of the verb in the construction transits from Early over Middle to Late, the most and the least content-sensitive learners (represented by the leftmost vs. rightmost *X*-values on each of the three panels) turn into each other's mirror image. The most content-sensitive learners are slower on verbs that occur late in a construction that in itself occurs early on in the sentence (late verbs early in the sentence) and faster on verbs that occur late in a construction that in itself occurs late in the sentence (late verbs late in the sentence); the least content-sensitive learners, conversely, appear to be faster on late verbs early in the sentence than on late verbs late in the sentence.

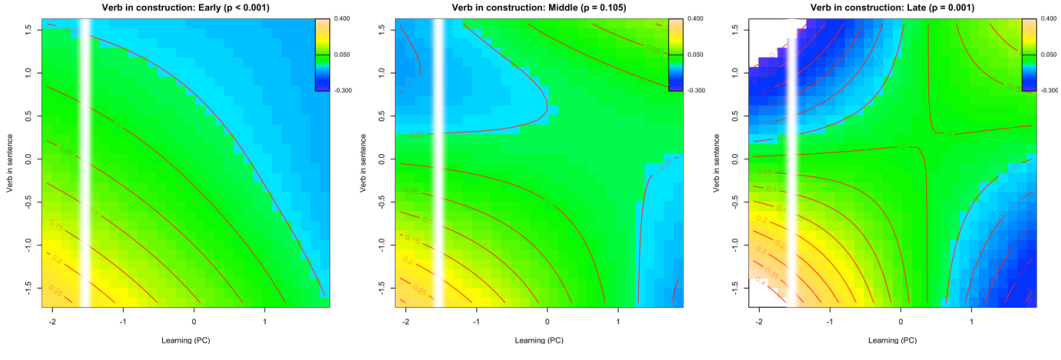

Fig. 1. Tensor product in the GAMM for reading latencies on the perception verb. Left panel: interaction of PC.Learning and VerbPosition for verbs that occur early in the construction. Middle panel: interaction of PC.Learning and VerbPosition for verbs that occur in the middle of the construction. Right panel: interaction of PC.Learning and VerbPosition for verbs that occur late in the construction.

Overall, the latter type of learners seems to be focusing on the task itself, rather than attempting to efficiently use linguistic information as it unfolds across the sentence that they are reading. They speed up when the information is limited (e.g., when the verb is presented early in the sentence and, thus, comes without much contextual support), and slow down when information accumulates (e.g., when the verb is late in the construction and the construction is late in the sentence). In other words, differences in pattern learning reveal radically different language processing signatures. Some learners may be trying to achieve efficiency given the task rather than given the content of the sentence. This interpretation is tested further in the following section.

3.3. Individual differences in pattern learning and their effect on task adaptation

Following the work by van Ravenzwaaij, Brown, and Wagenmakers (2011), Stafford and Dewar (2014), and Milin et al. (2017), we hypothesize that participants' deviations from the average reading latencies would be a valid and reliable indicator of individual differences in task adaptation. According to these studies, more *explorative* behavior (i.e., behavior marked by greater variability in performance) is expected at the beginning of an experiment and leads to knowledge that can be *exploited* to yield more rewarding behavior in later phases. The SPR task is interesting in this respect: there is the meaningful dimension of reading authentic sentences and the less directly meaningful dimension of coping with the fact that the sentences are randomized and do not form a coherent text. We are particularly interested in establishing whether pattern learning ability applies equally to these two different challenges within the same task, that is, to their performance in terms of "receiving the message" (the meaningful dimension) and their performance in adapting to the trial randomization (the meaningless dimension), as both characterize the learning/performing setup.

Stafford and Dewar (2014) suggested that the exploration-exploitation hypothesis could be tested by examining the variability (rather than the central tendency) of participants' behavior within an experiment.[6] Milin et al. (2017) made use of estimates of moving (or rolling) standard deviations using five consecutive data points (verb reading latencies over successive trials) as step size. Here, we use all words as they appear within a sentence and calculate the moving standard deviation over three consecutive data points. This procedure was repeated for all experimental sentences in the order in which they were presented to each individual participant given their randomization. We opted for the narrower window of three consecutive data points, trading off a reduction in the accuracy of the estimate of deviation for (a) maximization of the number of estimated deviations given the relative brevity of the sentences (i.e., there are not many data points in a sequence), and (b) minimization of the loss of data due to the fact that the rolling deviation is only available upon accumulation of the required number of data points (i.e., for the first two words of each sentence, a rolling deviation is not available). After removing the data for the one participant with an extremely steep decelerating trend in the SRT task (the same participant was also removed from the GAMM analysis in 3.2), we were left with 8,177 data points for analysis.

We fitted a quantile generalized additive model using the **qgam** package (QGAM: Fasiolo, Goude, Nedellec, & Wood, 2020) in the **R** software environment (R Core Team,

2017). A QGAM was preferred over a standard GAMM as it does not assume any particular distribution of the error term. This seemed more suitable given the fact that the dependent variable is the moving standard deviation, which should not be taken to have Gaussian properties. Furthermore, QGAM is particularly well-suited to reveal "snapshots" or "slices" of effects at given quantile values of the dependent variable. This approach allows us to emphasize the process of learning and brings us closer to a real-time model of the change that takes place (cf. Frost et al., 2015).

The model was fitted to the logarithm of the moving standard deviations of the reading latencies from the SPR task. We evaluated the model at the midpoint of the first quartile (quantile = 0.125), at the median (quantile = 0.5), and at the midpoint of the fourth quartile (quantile = 0.875) of the dependent variable (for more details on selecting the criterion quantile points, see Schmidtke, Matsuki, & Kuperman, 2017; Tomaschek, Tucker, Fasiolo, & Baayen, 2018).

As in the previous analysis on reading latencies (Section 3.2), our main predictor of participants' knowledge of language patterns was the Principal Component score for Learning (PC.Learning, discussed in Section 3.1). We defined three groups of participants according to their PC.Learning score, to match the model evaluation points: one for the lowest quartile (0-25%); 25% of scores around the median point (37.5%-62.5%); and the top quartile (75%-100%). Two further predictors are time-related and capture the position of the word in the sentence (WordInTrial) and the position of the trial in the experiment (Trial). Because the words in a sentence form a coherent whole but the sentences in the experiment do not, the former captures performance on structured, meaningful patterns, while the latter captures performance on random patterns; this is parallel with the SRT task that consists of structured and random parts. Both WordInTrial and Trial were scaled (z-transformed) prior to modeling. The model revealed significant interactions between these two temporal predictors and three groups of learners: task-sensitive (first quartile), balanced (median-centered quartile), and content-sensitive (fourth quartile).

First, the by-participant and by-item (i.e., word) random effect adjustments for the intercept were statistically highly significant. At the median values of the two temporal predictors, test results were, respectively: $edf = 28.172$, $\chi^2 = 1098.362$, $p < .001$ for participants, and $edf = 15.470$, $\chi^2 = 112.722$, $p < .001$ for items. Figs. 2a and b present the details of the interaction between types of learners (task-sensitive, balanced, content-sensitive), on the one hand, and WordInTrial (Fig. 2a) and Trial (Fig. 2b), on the other hand. The dependent variable—the 3-step moving standard deviation of reading latencies—was inspected, as explained above, at the three quantile points (low SD, median SD, high SD). Overall, the patterns for three groups of pattern learners are very different: the more content-sensitive the learner (balanced to high), the more pronounced their engagement at the sentence level (Fig. 2a). The task-sensitive learners, conversely, engage only at the most general level, that is, across experimental trials (Fig. 2b).

As Fig. 2a shows, balanced and content-sensitive pattern learners deviate further from the mean reading latencies. This is more pronounced at the start and then again at the end of the sentence but the sentence-final effect is less indicative given the width of the

confidence intervals. For both median and high SDs, the most content-sensitive learners show a stronger, more pronounced U-shaped effect than balanced learners. Both groups, however, show a similar trend of maximally exploitative behavior around the middle of the sentence, while explorations do occur both at the beginning and toward the end of the sentence. Task-oriented pattern learners, on the other hand, show a clear downward trend across experimental trials (Fig. 2b). For high SDs, this trend is even more pronounced, revealing a mild non-linear decrease of the indicator of exploration (i.e., the moving

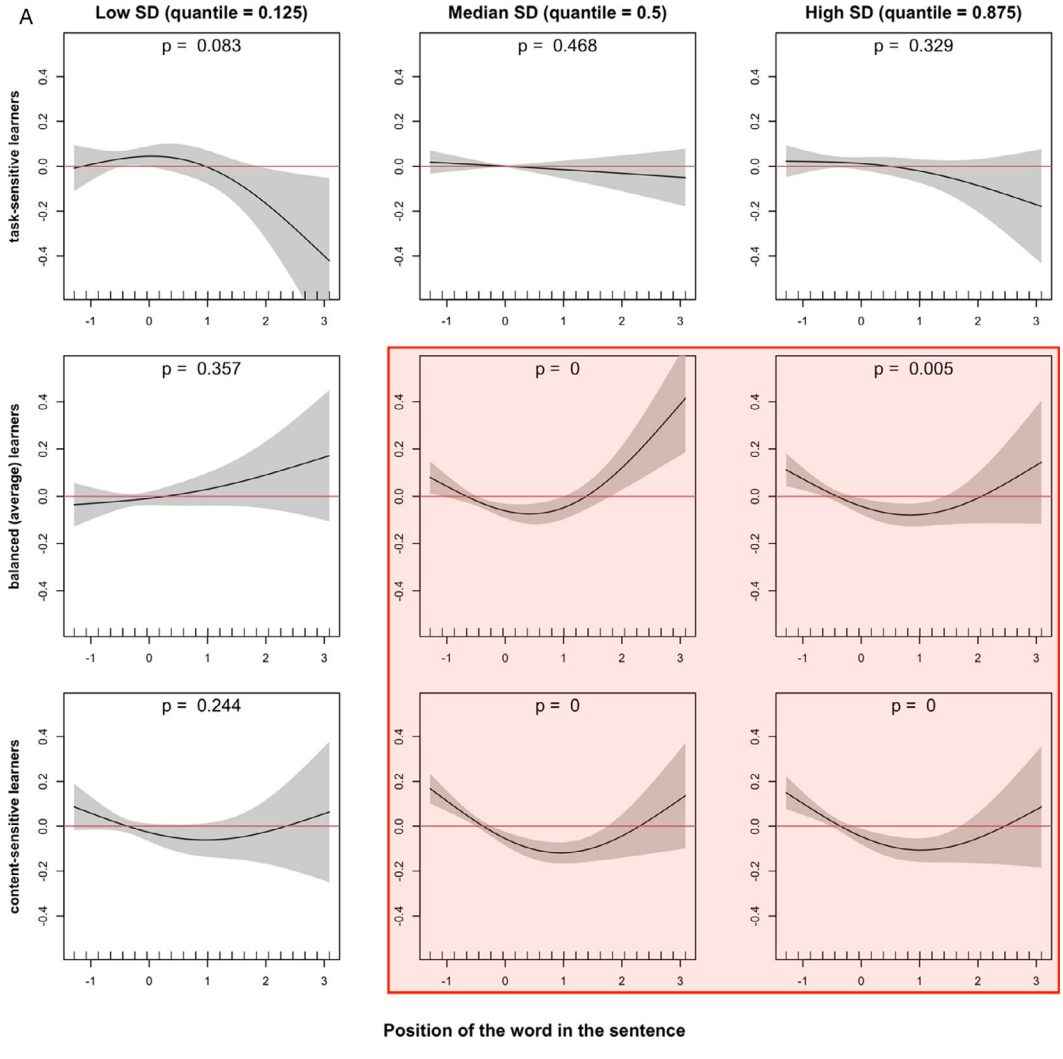

Fig. 2. (A) Partial effect plots of the rolling standard deviations on the reading latencies for all words in the SPR task. Top to bottom, the rows present panels for the three groups of learners. Left to right, the columns contain panels for low, mid, and high standard deviations (dependent variable). The *X*-axis represents the scaled position of the word in the sentence (WordInTrial). The red-highlighted panels mark statistically significant interactions of readers by WordInTrial.

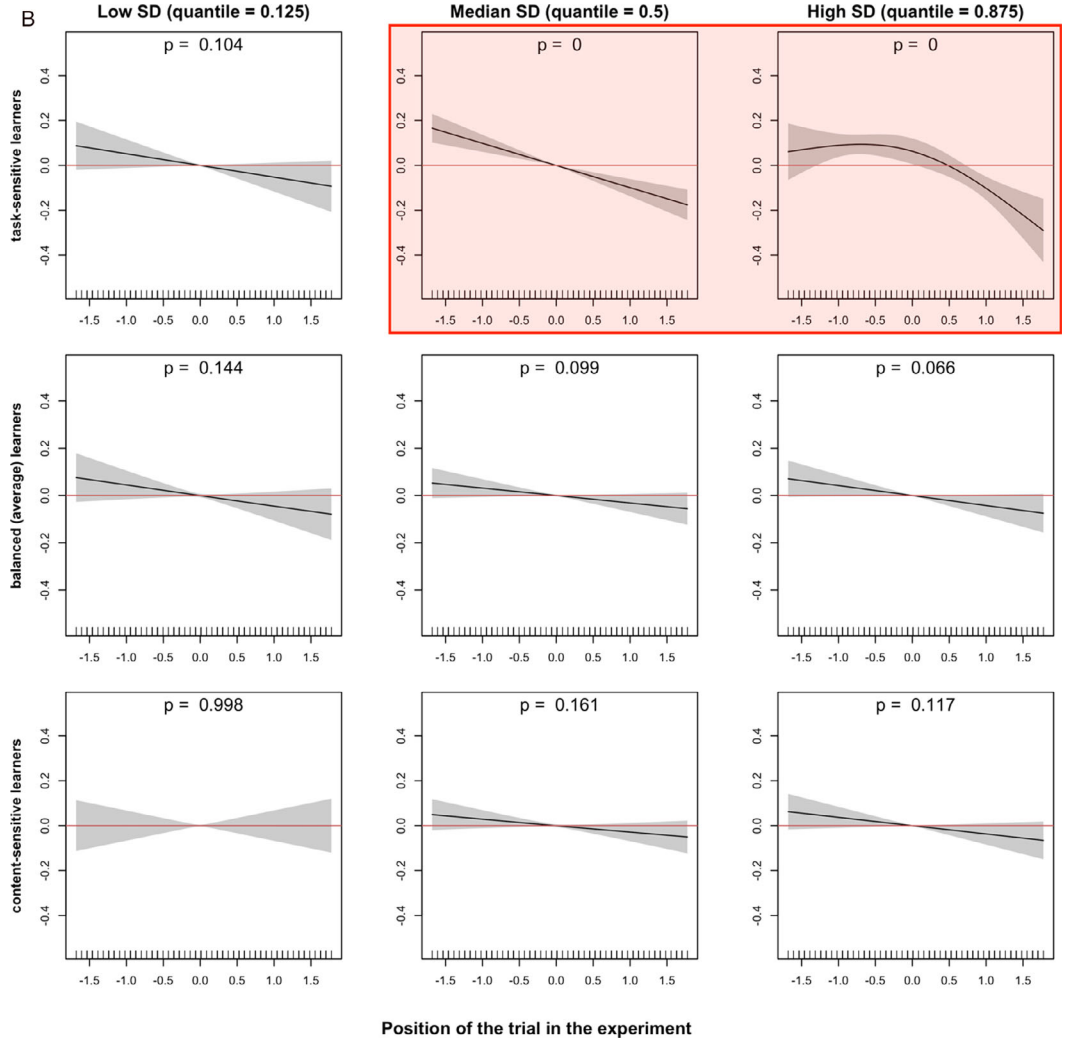

Fig. 2. (B) Partial effect plots of the rolling standard deviations on the reading latencies for all words in the SPR task. Top to bottom, the rows present panels for the three groups of learners. Left to right, the columns present panels for low, mid, and high standard deviation. The *X*-axis represents the scaled position of trial in the experiment (Trial). The red-highlighted panels mark statistically significant interactions of learners by Trial.

standard deviation). This type of exploration, however, remains restricted to the level of the task and does not indicate true engagement with the sentence. Balanced and content-sensitive learners, conversely, perform rather consistently across trials, at the level of the experiment; the changes are non-significant.

Thus, in sum, task-sensitive learners do not engage with the sentence content and only show significantly deviating reading times at the level of the experiment. Both balanced and content-sensitive learners engage with the words in the experimental sentences, showing the most exploitative reading performance at the sentence midpoint.

4. Discussion

In this paper, we have focused specifically on testing whether individual differences in pattern learning link directly to individual differences in linguistic processing and whether pattern learning affects performance along multiple dimensions of a task. In the following sections, we discuss how our analyses of data from an SPR and an SRT task answer our three research questions: learning non-linguistic and linguistic sequences relies on the same mechanism (Section 4.1); individuals differ in their pattern learning and this is reflected in how they read morphological and syntactic patterns (Section 4.2); depending on their pattern learning, individuals differ in their readiness to explore and exploit the environment in order to adapt to the task and improve their performance (Section 4.3).

4.1. Components of pattern learning: Distinguishing learning from adaptation

One of the challenges posed by experimentally collected data on learning concerns the apparent conflation of the effects of practice with the task and of learning new information on performance. Using principal component analysis, we obtained a fine-grained view on the different components that affect performance. We established that the increase in speed due to pattern learning (PC1 = Learning) and the increase in speed solely due to practice with the SRT task (PC2 = Mechanics) are independent from each other. Learning any type of structured sequences is different from mere task adaptation. Furthermore, and crucial to our argument, our results showed that the slope adjustments from both SRT and SPR tasks load onto the same and strongest component (PC1 = Learning). This highlights the similarity in learning structured patterns across tasks: While non-linguistic pattern learning is evidenced by an increased speed in recognizing the location at which the star appears in the SRT task, linguistic pattern learning becomes apparent through an increase in speed when reading morphologically or syntactically typical chunks in the SPR task.

The type of pattern learning shared across the SRT and SPR tasks and captured by PC1 confirms our hypothesis that one general mechanism supports pattern learning across domains: Learning non-linguistic and linguistic sequences is (strongly) related and this relationship implies the domain-generality of statistical learning. We argued earlier that experience-based probabilistic learning, driven by pattern recognition abilities that are not unique to language, can be invoked to argue against nativist, domain-specific accounts of language and cognition. The ability to detect patterns in usage makes it possible to build a grammar from exposure to the ambient language without having to posit specific cognitive abilities or modules that, ultimately, separate language cognition from other areas of cognition. Over the past three decades, usage-based theories of language have made impressive advances in explaining how knowledge of language could be built up with the support of general cognitive capacities only (Ambridge & Lieven, 2011). Among the basic cognitive capacities that are generally listed are perception, attention, categorization, abstraction, and memory. Although pattern detection is not typically mentioned explicitly, it deserves a place on this list, too: It separates the initially unseparated experience into

manageable chunks on which attention, categorization, abstraction, and memory operate. Thus, grammar can be viewed as the cognitive organization of one's experience with language (Bybee, 2010). It is, however, recognized that languages are constantly changing, and usage depends on the individual and the situation (Hopper, 1987). The cognitive organization of language knowledge is, therefore, not a static structure, expressible as a fixed set of representations.

In other words, usage-based linguists recognize that both the experience and the organization of the experience are in a perpetual state of flux. This aligns with the starting point for theories of learning. In its most basic form, learning can be presented as a simple *discriminative principle* that emerges from "exposure to relations among events in the environment" to help an organism acquire knowledge about "the structure of its world" (Rescorla, 1988, p. 152). Initially, all is one, and wholes are only separated into their component parts if this is indicated by adaptive pressures (cf. James, 1890). Given that the world is in a constant state of flux, the mind—as part of that world—is, too, and (mental) units only emerge if needed.

Theories of learning, too, have invested considerable effort to argue for one general principle of learning. Surprisingly, the statistical learning literature does not appear to engage with learning theory more generally. This raises the question of whether there is anything about statistical learning that sets it apart from other types of learning? The prevalent view on statistical learning posits that statistical learning proceeds through the computation of transitional probabilities (Aslin, Saffran, & Newport, 1998), but the nature of the statistical computations remains underdetermined (Perruchet & Poulin-Charronnat, 2012, p. 121). Some assume that learners perform the same computations as statisticians, albeit unconsciously, while others work from the assumption that learners approximate these probabilistic "guestimates" through progressive tuning of associative links between cues and outcomes. The latter interpretation is in line with the general learning theoretic view. In fact, all formal learning traditions assume that what is learned are probabilistic regularities present in the environment. Competition learning (i.e., discriminative learning, like Rescorla & Wagner, 1972), attention learning (e.g., Mackintosh, 1975; Pearce & Hall, 1980), and even comparator-driven learning (Stout & Miller, 2007) naturally accommodate "statistical learning," although they would differ in their views on how cues compete to gain their predictive (i.e., discriminative) value for a given outcome.

4.2. Individual differences in pattern learning and their effect on language processing

Our statistical model captures accurately how subtle differences in sequence learning and contextual information co-determine the speed with which sentences are read and processed on an individual basis. There is a clear difference in how pattern learners process perception constructions, depending on whether the verb occurs first, second, or last in the construction, and depending further on where in the sentence the construction is positioned.

Content-sensitive learners are slow in reading perception verbs if the perception verb occurs early in the construction and the construction occurs early in the sentence; they

become faster as the perception verb moves to later positions, in the construction and in the sentence. In fact, content-sensitive learners show that, when verbs occur late in the construction, the position of the construction in the sentence will further modulate reading speed: If the construction appears early in the sentence, reading will be at its slowest, but if it occurs late in the sentence, reading will be at its fastest. This indicates that such learners are tuning their reading speed to the current level of information: Late in the construction, the verb completes the event structure, and late in the sentence, the verb wraps up the sentence structure, at which point the information is completed and uncertainty is at its lowest. Content-sensitive learners, who flexibly adjust to the accumulation of information as the sentence unfolds, make predictions based on the available evidence: They anticipate the verb and hence they do not need much time to read the verb in these "doubly" late positions, i.e., when the verb appears late in the construction and late in the sentence.

Conversely, task-sensitive learners' reading speed is less affected by the positional aspect of the linguistic information available. These learners, overall, implement the opposite strategy of content-sensitive learners: Verbs occurring late in the construction but early in the sentence are processed fastest, while verbs occurring late in the construction are processed slowest if that construction is also late in the sentence. This indicates that task-sensitive learners are unable to engage efficiently with the information collected over longer stretches of text. Their "guesses" are less well-grounded, which render their reading less efficient.[7]

Generally, these results can be taken to show that the amount of information that is available affects both types of learners differently. Content-sensitive learners benefit from the accumulation of information and use it to optimize their reading speed; hence, they suffer more from lack of information. Task-sensitive learners struggle to navigate information and uncertainty at the sentence level, and their reading slows down as more information accumulates over the course of the sentence; hence, they suffer more from accumulation of information. A similar interindividual difference in local versus global focus was recently reported by Siegelman, Bogaerts, Armstrong, and Frost (2019) for visual statistical learning.

Furthermore, our findings reveal that the information that readers are sensitive to is syntactic in nature. Information at a lower, morphemic level did not appear to have this effect in word-in-sentence reading, which is in line with the findings of Schmalz et al. (2019) on single-word reading. Although morphological (dis)preferences have been found to facilitate or impede reading (Milin et al., 2017) and sensitivity to (dis)preferences correlates with mental speed, it does not seem to be related to individual differences in sequence learning. The units that play a role in language processing, as measured by SPR, and that appear to be correlated with pattern learning more generally are *natural* units of experience: Verb argument constructions render the core of an event, and several such events can be concatenated to form a complex sentence. In case of perception verbs, typically, there is someone who perceives and something that is perceived. Both in the relevant experiential and in the corresponding linguistic context, the subject and object will occur together with the perception verb.

Sensitivity to the edges of short strings, such as our three-part (Subject Verb Object) event structures, appears to be a limitation typical of sequence learning in general (see, e.g., Williams & Rebuschat, 2012, p. 253). Why is this so? Although it has been customary to posit statistical learning as a separate learning mechanism, it may well turn out to be more parsimonious to assume that the effect of probabilistic distributional patterns on learning follows from basic learning principles, and from associative (i.e., discriminative) learning principles in particular. Expressions of support for considering statistical learning as a re-branding of associative learning are few and far between. Perruchet and Poulin-Charronnat (2012) is one notable exception, but for them, attention plays a key role in learning as it can be seen as a crucial condition for the formation of an association between two events. In human beings, attentional coding of incoming information appears to naturally segment the material into chunks. Sensitivity to the edges of short strings is thus naturally accommodated and associations will not be learned between elements that straddle an intonational phrase boundary because these elements are not held within the same attentional chunk. Yet we would argue that, if we have learning, we do not need attention as a separate construct. Due to the distribution of events in the world as we experience it, relations between event structures and their corresponding linguistic expressions form naturally; the event components occur frequently together in experience, which ensures that they are often talked about in conjunction, too. This implies that, both in experience and in language, certain things are likely to be highly activated at the same time, resulting in a state that equals being in focal attention. Co-occurrence is a precondition for being considered a unit: What goes together must be often (highly) activated at the same time. This fits well with Rescorla's (1988) observation that discrimination emerges from "exposure to relations among events in the environment" (p. 152). Here, it is the linguistic, that is, syntactic, environment that is directly related to and thus relevant for relations in the material world. From a strictly discriminative perspective, effects that could be called "attentional" emerge naturally from the highly complex interplay between environment and organism and the resulting adaptive pressures. Perhaps specific to humans, that environment is also social and communicative and, thus, linguistic. What our results show, then, is that individual pattern learning affects the discriminative event structure for the purpose of processing written language.

4.3. Individual differences in pattern learning and their effect on task adaptation

Contrasting the top and bottom 25% of learners in terms of their pattern learning revealed that the task adaptation trends for content-sensitive versus task-sensitive learners are substantially different and, crucially, they appear significant for different dimensions of the experiment. More specifically, content-sensitive learners focus on the sentence and apply a strategy, whereby they deviate rather more from the mean, especially at the start and end of the sentence. This could indicate that these learners are more able to engage in parallel processing, too. Our results also show that what we observe is a trend that develops across learning styles: It shows for balanced learners and peaks for content-sensitive learners. Task-sensitive learners, on the other hand, focus on the experiment, and

overall they deviate rather less from the mean; in addition, for task-sensitive learners, there is a clear downward trend within the experiment, with deviations diving below average half-way through the experiment. In other words, although the general, task-related behavior of these learners becomes more consistent, their specific reading engagement does not.

From the perspective of RL, this indicates that content-sensitive learners end up not only with more information, but also with more relevant information that they can exploit in order to maximize their (reading) performance. To put it differently: not only do they explore more; they also explore in the most relevant way, that is, at the most relevant level. The sentences that make up the SPR task do not form a coherent text; hence, nothing can be learned across trials of disconnected sentences (i.e., at the level of the experiment). At the level of the sentence, however, much is to be gained: whether a verb occurs early or late in its construction and whether that construction occurs early or late in the sentence is a proxy for the amount of information that has been revealed before the verb is encountered. Tracking that information and making use of it aid sentence processing.

Another interesting difference between the two types of learners—content-sensitive and task-sensitive learners—is that, when they do show significant change (over position of the word in the sentence for the former and over position of trial in the experiment for the latter), they differ in their respective variability, as indicated by the confidence intervals (in Figs. 2a and b, red-highlighted panels). While the task-sensitive learners show consistently narrow confidence intervals (Fig. 2b), the content-sensitive learners exhibit substantial change in variation, with larger deviations at the end of the sentence (Fig. 2a). In other words, toward the end of the sentence, content-sensitive learners vary more (upward trend of the curves), and they show more individual differences in that variation (wider confidence intervals). At the same time, their reading times get shorter *on average* (visible in the right-hand panel in Fig. 1). One plausible explanation is that, through shorter but more variable reading times (given the average and the deviation), content-sensitive learners demonstrate higher overall reading efficiency (shorter mean reading time) and, simultaneously, content-specific and highly individualized integrative efforts (overall higher and more dispersed deviations) that, naturally, take place towards the end of the sentence (cf. Kuperman et al., 2010). To put it simply, content-sensitive learners are trying their best to *read to understand*. Task-sensitive learners, conversely, are just *keeping up with reading*.

In sum, content-sensitive and task-sensitive learners also exhibit differences in exploration focus and strategy, which affect the amount of relevant information they collect for later exploitation. Overall, content-sensitive learners benefitted more from information that became available at a more optimal point in time during reading. Pattern learning ability can thus be said to shape attempts to achieve optimal performance in different ways or at different levels of the task. The information that is gathered is influenced by such attempts and participants who differ in their respective pattern learning abilities also differ in how they engage with the task at hand and, ultimately, how they collect their rewards.

5. Conclusions

Our statistical pattern learning ability, that is, our ability to learn regularities in sequential input, has been invoked to explain how knowledge of a language can emerge from exposure to usage only. In this paper, we have investigated the strength of the correlation between statistical pattern learning and the processing of morphological and syntactic information, and we have shown that individual differences in pattern learning are reflected in how individuals approach a task and adapt to it. This allows to optimize future behavior and reap further rewards, thereby strengthening and widening the impact that pattern learning ability has on language-related performance.

Using data from a visual SRT task and a SPR task with naturalistic Russian sentences, we have provided further evidence that the learning of non-linguistic and of linguistic sequences is strongly related. Moreover, we have shown that learning any type of structured sequences is unrelated to mere task adaptation.

Next, we have focused on the question of whether pattern learning underpins sensitivity to morphological as well as syntactic co-occurrence patterns in native speakers. We have found that subtle differences in sequence learning and contextual information co-determine the speed with which sentences are read and processed on an individual basis, despite the fact that all participants were highly educated young adults who did not differ significantly in their reading habits. While task-sensitive learners struggle to navigate through sentence uncertainty—their reading slows down as more information accumulates over the course of the sentence—content-sensitive learners benefit from the accumulation of information and use it to optimize their reading.

Finally, we have demonstrated that pattern learning also correlates with how the learning/performing setup is approached. Content-sensitive and task-sensitive learners exhibit a difference in exploration focus and strategy. Not only do content-sensitive learners explore more, they also explore at the most relevant level. This affects the amount of relevant information they collect and have at their disposal for later exploitation. In other words, they identify those patterns that yield the biggest return in terms of performance improvement.

Wrapping up, we can say that pattern learning ability affects performance along multiple dimensions, as would be expected from a domain-general cognitive function or mechanism. Its effects on language processing are well known, but the way in which it shapes exploratory behavior has long gone unnoticed. Yet, by affecting exploratory behavior, it influences the information that is gathered and becomes available to be exploited. The effect it has on performance consequently doubles: (exogenous) information is processed by a (endogenous) general mechanism that is adaptively attuned to maximize efficiency across domains and to accommodate individuals' strengths or styles of learning.

Acknowledgments

The financial support of the British Academy, the Prokhorov Foundation, and The Leverhulme Trust (grant RL-2016-001) is gratefully acknowledged. The experiments received

ethical approval from the University of Sheffield (UK), School of Languages & Cultures. The Institute for Linguistic Studies of the St. Petersburg branch of the Russian Academy of Sciences kindly made facilities available for testing. The PsychoPy scripts were written by Lily FitzGibbon; participants were recruited and scheduled by Daria Satyukova. We thank Harald Baayen, Adnane Ez-Zizi, Victor Kuperman, Danielle Matthews, Srdan Medimorec, and two anonymous reviewers for comments on earlier drafts of this paper.

Conflict of interests

The authors declare that they have no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

Data and code used in this study can be found at the following links

Script files: https://github.com/ooominds/Exploring_and_exploiting_uncertainty
Data files: https://doi.org/10.25500/edata.bham.00000452

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://doi.org/10.25500/edata.bham.00000452.

Notes

1. For the purposes of our study, we will not differentiate between *statistical* vs. *implicit* learning. According to Perruchet (2019), these two prolific approaches to the study of learning discuss, in fact, the very same mechanisms (or functions) of domain-general and incidental learning. They do, however, focus on different aspects of that same mechanism: learning the "units" (implicit) or learning the transitional probabilities (statistical) in order to discover probable "units."
2. They discarded data from the visual embedded triplet paradigm due to the poor psychometric properties of the task, yielding insufficient variability in the learning performance.
3. Tasks involving sequences of auditory syllables or visual nonsense shapes seem more powerful predictors of learning outcomes in other domains than tasks involving nonadjacent dependencies, or nonverbal auditory stimuli.

4. We also ran a forward digit span task that turned out to be insignificant for modeling the data. Therefore, it is not considered further in our main analysis but reported only as SupMat (A).
5. This is a common outcome in PCA when the maximum number of components is exhausted (i.e., is equal to the number of input variables): The least significant component can indicate a differentiation between variables that are strongly loaded on the most significant components (see, e.g., PCA over measures of individual reading abilities in Andrews & Lo, 2012, 2013). Typically (and arguably), this is considered to be an artifact of the method itself. Simply put, PCA maximizes the "greediness" of the isolated components such that the first component takes as much as possible from the total variability of a given set of variables (in our particular case 4). The next component takes as much as possible from what remains, and so on. This, then, ensures that the newly isolated dimensions have the following two characteristics: (a) their explained variance decreases (based on the "leftovers" policy) and (b) they are mutually independent, uncorrelated, or orthogonal (exactly because they take only what is left unexplained or unaccounted for). The former characteristic is why PCA has many criteria for deciding the *true* number of components, which is usually much smaller than the number of variables in the original set. This is why we may consider PCA to be a method that will efficiently reduce the dimensionality of our initial variable space; the later components contain progressively less accounted variance and, thus, they are often odd and uninterpretable. In our particular case, for example, we would retain only the first two components under the criterion of Kaiser-Guttman (cf. Cattell, 1966; Yeomans & Golder, 1982) as these two components are the only ones that account for variance greater than one—greater than a single variable can account for ($0.355 \times 4 = 1.42$; $0.255 \times 4 = 1.02$; $0.250 \times 4 = 1.0$; $0.140 \times 4 = 0.56$). The latter characteristic, i.e., orthogonal, uncorrelated components, conversely, allows us to effectively reparametrize the original variable space into a new space of perfectly independent variables. This is why, in some cases, using PCA to devise new variables that are mutually independent and, hence, can cure the problem of multicollinearity is a sound statistical approach (as it is the case in the studies of Andrews & Lo, 2012, 2013).
6. This idea links influential work of Hick (1952) on the *rate of gain of information* with the *drift diffusion model* of Ratcliff (1978) who argued that the speed of information processing shows a strong correlation with the deviation in processing latencies (see also Ratcliff, Schmiedek, & McKoon, 2008).
7. Recent work has found, indeed, that the length of successive correct predictions in an oculomotor SRT task is correlated to working memory capacity, with lower working memory capacity yielding shorter chunks of successive correct predictions (cf. Medimorec et al., in press).

References

Ambridge, B., & Lieven, E. V. M. (2011). *Child language acquisition: contrasting theoretical approaches*. Cambridge, UK: Cambridge University Press.

Andrews, S., & Lo, S. (2012). Not all skilled readers have cracked the code: Individual differences in masked form priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 152–163. https://doi.org/10.1037/a0024953

Andrews, S., & Lo, S. (2013). Is morphological priming stronger for transparent than opaque words? It depends on individual differences in spelling and vocabulary. *Journal of Memory and Language*, *68*(3), 279–296.

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*(2), 286–304.

Arppe, A. (2013). Package "polytomous": Polytomous logistic regression for fixed and mixed effects (version 0.1. 4). The R Project for Statistical Computing.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Sunderland, MA: Sinauer Associates.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge, UK: Cambridge University Press.

Cattell, R. B. (1966). *Handbook of multivariate experimental psychology*. Chicago, IL: Rand McNally.

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*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Divjak, D. (2015). Exploring the grammar of perception: A case study using data from Russian. *Functions of Language*, *22*(1), 44–68.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2020). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2020.1725521

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., & Heiberger, R. (2017). *Package "car" (Version 2.1-2)*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289–344. https://doi.org/10.1037/0033-295X.107.2.289

Gameiro, R. R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and exploitation in natural viewing behavior. *Scientific Reports*, *7*(1), 2311.

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313.

Gureckis, T. M., & Love, B. C. (2015). Computational reinforcement learning. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational mathematical psychology* (pp. 99–117). New York: Oxford University Press.

Harris, R. J. (1985). *A primer of multivariate statistics*, 2nd ed. London: Academic Press.

Hasson, U. (2017). The neurobiology of uncertainty: implications for statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160048.

Hayes, T. R., & Petrov, A. A. (2016). Pupil diameter tracks the exploration-exploitation trade-off during analogical reasoning and explains individual differences in fluid intelligence. *Journal of Cognitive Neuroscience*, *28*(2), 308–318.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26.

Hopper, P. (1987). Emergent grammar. *Annual Meeting of the Berkeley Linguistics Society*, *13*, 139.

Hull, C. L. (1935). Review of Thorndike's fundamentals of learning. *Psychological Bulletin*, *32*(10), 807–823.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*(4), 658–680. https://doi.org/10.1037/0096-3445.130.4.658

James, W. (1890). *The principles of psychology* (Vol. 1). New York: Henry Holt and Co.

Johnston, T. D. (1982). Selective costs and benefits in the evolution of learning. In J. S. Rosenblatt (Ed.), *Advances in the study of behavior* (Vol. *12*, pp. 65–106). New York: Academic Press.

Jost, E., & Christensen, M. H. (2017). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 227–244). Washington, DC: American Psychological Association.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340.

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*(1), 171–184. https://doi.org/10.1037/a0025405

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *The Quarterly Journal of Experimental Psychology*, *63*(9), 1838–1857.

Luzardo, A., Alonso, E., & Mondragón, E. (2017). A Rescorla-Wagner drift-diffusion model of conditioning and timing. *PLOS Computational Biology*, *13*(11), e1005796.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

Medimorec, S., Milin, P., & Divjak, D. (In press). Working memory affects anticipatory behavior during implicit pattern learning. *Psychological Research Psychologische Forschung*, https://doi.org/10.1007/s00426-019-01251-w

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1730–1751. https://doi.org/10.1037/xlm0000410

Milne, A. E., Wilson, B., & Christiansen, M. H. (2018). Structured sequence learning across sensory modalities in humans and nonhuman primates. *Current Opinion in Behavioral Sciences*, *21*, 39–48.

Misyak, J. B., & Christiansen, M. H. (2007). Extending statistical learning farther and further: Long-distance dependencies, and individual differences in statistical learning and language. In Proceedings of the Annual Meeting of the Cognitive Science Society (pp. 1307–1312). The 29th Annual Conference, Nashville, TN.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138–153. https://doi.org/10.1111/j.1756-8765.2009.01072.x

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*(10), 1–8.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Science*, *11*(3), 520–535. https://doi.org/10.1111/tops.12403

Perruchet, P., & Poulin-Charronnat, B. (2012). Word segmentation: Trading the (new, but poor) concept of statistical computation for the (old, but richer) associative approach. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 119–143). Berlin: De Gruyter.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, *36*(1), 10–17.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Rescorla, R. A., & Wagner, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In H. Black & W. F. Proksay (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Schapiro, A., & Turk-Browne, N. (2015). Statistical learning. In A. W. Toga (Ed.), *Brain mapping: An encyclopedic reference* (Vol. *3*, pp. 501–506). Amsterdam: Elsevier.

Schmalz, X., Moll, K., Mulatti, C., & Schulte-Körne, G. (2019). Is statistical learning ability related to reading ability, and if so, why? *Scientific Studies of Reading*, *23*(1), 64–76.

Schmidtke, D., Gagné, C. L., Kuperman, V., & Spalding, T. L. (2018). Language experience shapes relational knowledge of compound words. *Psychonomic Bulletin & Review*, *25*(4), 1468–1487.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793.

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.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*(2), 511–518. https://doi.org/10.1177/0956797613511466

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, *114*(3), 759–783. https://doi.org/10.1037/0033-295X.114.3.759

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: The Macmillan.

Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, *4*(s2), 20170018. https://doi.org/10.1515/lingvan-2017-0018

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*(3), 381–393.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2016). *itsadug: Interpreting time series and autocorrelated data using GAMMs (Version 2.3)*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Vogel, E., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176.

Williams, J. N., & Rebuschat, P. (2012). Statistical learning and syntax: What can be learned and what difference does meaning make? In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 237–264). Berlin: De Gruyter.

Willingham, D. B., & Goedert-Eschmann, K. (1999). The relation between implicit and explicit learning: Evidence for parallel development. *Psychological Science*, *10*(6), 531–534. https://doi.org/10.1111/1467-9280.00201

Wonnacott, E. (2013). Statistical mechanisms in language acquisition. In P. M. Binder & K. Smith (Eds.), *The language phenomenon: Human communication from milliseconds to millennia* (pp. 65–92). Berlin: Springer-Verlag.

Wood, S. N. (2006). *Generalized additive models*. Boca Raton, CA: Chapman & Hall.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Yeomans, K. A., & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *31*(3), 221–229. https://doi.org/10.2307/2987988

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692.

Supporting Information

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

Data S1. Supplementary Analyses.