Statistically-induced chunking recall:

A memory-based approach to statistical learning

Erin S. Isbilen[1], Stewart M. McCauley[2], Evan Kidd,[3,4,5] and Morten H. Christiansen[1,5,6,7]

Cornell University, Department of Psychology[1]

University of Iowa, Department of Communication Sciences and Disorders[2]

Max Planck Institute for Psycholinguistics, Nijmegen, Language Development Department[3]

The Australian National University, Canberra, Research School of Psychology[4]

ARC Centre of Excellence for the Dynamics of Language[5]

Aarhus University, School of Communication and Culture[6]

Haskins Laboratories[7]

Keywords: statistical learning; chunking; serial recall; nonword repetition; language acquisition; learning; memory; language

Please address correspondence to:
Erin S. Isbilen
Department of Psychology
Uris Hall
Cornell University
Ithaca, NY 14853
Phone: 607-255-3834
e-mail: esi6@cornell.edu

Abstract

The computations involved in statistical learning have long been debated. Here, we build on work suggesting that a basic memory process, *chunking,* may account for the processing of statistical regularities into larger units. Drawing on methods from the memory literature, we developed a novel paradigm to test statistical learning by leveraging a robust phenomenon observed in serial recall tasks: that short-term memory is fundamentally shaped by long-term distributional learning. In the statistically-induced chunking recall (SICR) task, participants are exposed to an artificial language, using a standard statistical learning exposure phase. Afterward, they recall strings of syllables that either follow the statistics of the artificial language, or comprise the same syllables presented in a random order. We hypothesized that if individuals had chunked the artificial language into word-like units, then the statistically-structured items would be more accurately recalled relative to the random controls. Our results demonstrate that SICR effectively captures learning in both the auditory and visual modalities, with participants displaying significantly improved recall of the statistically-structured items, and even recall specific trigram chunks from the input. SICR also exhibits greater test-retest reliability in the auditory modality and sensitivity to individual differences in both modalities than the standard two-alternative forced-choice task. These results thereby provide key empirical support to the chunking account of statistical learning, and contribute a valuable new tool to the literature.

1. Introduction

Statistical learning is recognized as a fundamental process by which humans and other animals learn the structure of their environment. Virtually all learning incorporates sensitivity to statistical regularities – from visual sequential and spatial learning (Turk-Browne, Jungé, & Scholl, 2005; Fiser & Aslin, 2001), to word and grammatical acquisition in infants and adults (Saffran, Aslin, & Newport, 1996; Saffran, 2003). Over the last two decades, a wealth of research has highlighted the ubiquity of statistical learning across domains, modalities, age groups and species, suggesting that statistical learning is a foundational cognitive process. However, despite its prevalence, a number of key outstanding questions about the nature of this phenomenon remain.

To date, studies of statistical learning have largely centered on demonstrating its versatility across a diverse array of contexts, often in isolation from other cognitive abilities (see Armstrong, Frost, & Christiansen, 2017; Frost, Armstrong, & Christiansen, 2019, for further discussion). As a result, the precise manner in which statistical learning interfaces with broader aspects of cognition remains a topic of debate. For instance, while some studies show that statistical learning occurs without the learner's overt attention (Evans, Saffran, & Robe-Torres, 2009), others report that learning is severely impaired when attentional resources are diverted to another task (Toro, Sinnett, & Soto-Faraco, 2005). Relatedly, the relative involvement of different memory systems in statistical learning has garnered increased speculation. While there appears to be distinct developmental differences in the amount of implicit and explicit knowledge gained during statistical learning tasks (Bertels, Boursain, Destrebecqz, & Gaillard, 2015; Batterink, Reber, Neville, & Paller, 2015), others have questioned whether the behavior observed is in fact

the result of long-term memory consolidation, or merely the product of short-term recognition (Kim, Seitz, Feenstra, & Shams, 2009).

The historical separation of statistical learning from other aspects of cognition underscores a gap in the field's current knowledge. Although we now know that humans attend to statistical regularities in order to learn the structure of their environment, comparatively little is understood about the precise computations and representations involved in the complex behaviors that it is argued to account for (Romberg & Saffran, 2010). For instance, statistical learning has frequently been discussed in the literature as a unitary mechanism, with the assumption that the various tests employed to probe this faculty do so with equal efficacy. However, recent theoretical proposals suggest that statistical learning may instead encompass a complex suite of computations, and that different tests of statistical learning may tap into different subcomponents of this process (Arciuli, 2017; Frost et al., 2019; Frost, Armstrong, Siegelman, & Christiansen, 2015; Misyak & Christiansen, 2012; Siegelman & Frost, 2015). What, then, do tests of statistical learning specifically measure? In the following section, we suggest that a key impediment to understanding the computations involved in statistical learning may arise from methodological limitations in how statistical learning is typically tested. In particular, we highlight the shortcomings of one of the field's most widely used tasks to test statistical learning in children and adults, the two-alternative forced-choice task (2AFC).

1.1. Methodological issues with testing statistical learning using 2AFC

Statistical learning typically occurs automatically, below the threshold of an individual's awareness. Yet often, the tests used to measure this ability require participants to translate

this passively-acquired knowledge into an overt, *reflection-based* response (Christiansen, 2019). For example, in the often-used 2AFC task, participants are asked to make an explicit choice between a target that follows the relevant statistical regularities and a foil that does not. While reflection-based measures have frequently been the standard for statistical learning research, they only provide an indirect measure of the effect of learning on underlying processing. Indeed, a growing body of literature questions whether reflection-based methods serve as an effective proxy for learning, given their disconnect from the type of behavior that they intend to measure (e.g., Franco, Eberlen, Destrebecqz, Cleermans, & Bertels, 2015; Siegelman, Bogaerts, Christiansen, & Frost, 2017; Siegelman, Bogaerts, & Frost, 2017). As forced-choice tasks require explicit decision making, the data they provide may therefore in part reflect participants' relative ability to speculate about learned material, rather than directly measuring the learning-based changes in processing (Christiansen, 2019).

In addition, the all-or-nothing scoring in 2AFC tasks may obscure subtle variations in individual performance (Siegelman et al., 2017). Participants either select the correct answer or not; this binary data provides limited insight into the specific information that the participant has learned (e.g., bigrams, whole words, or the relative positions of syllables within words). Furthermore, although 2AFC performance typically averages around 60% across an entire sample, up to one-third of these participants often perform below chance level. The task thus fails to capture reliable data about the statistical learning skills of a sizable proportion of the sample (Siegelman & Frost, 2015; Frost et al., 2015; Frost, Siegelman, Narkiss, & Affek, 2013; Misyak, Christiansen, & Tomblin, 2010).

It is worth noting that variations exist in the way 2AFC is implemented. For instance, there is considerable diversity in the kind of structures that participants are presented with for comparison at test. In many cases, participants see or hear trigrams from the input, which are then compared against performance on random foil items (e.g., Saffran, Newport, & Aslin, 1996; Turk-Browne et al., 2005), as is the focus of the current paper. Additionally, 2AFC stimuli also commonly comprise words from the input versus partwords that span word boundaries (Aslin, Saffran, & Newport, 1998; Saffran et al., 1996). This has been argued to be a more robust measure of learning as it makes the task more difficult: participants must distinguish a word from a part-word which differ only by a single syllable. Others have manipulated positional information while preserving legal transitional probabilities (Endress & Mehler, 2009), and 2AFC has even been used to test the generalization of structure by presenting novel items that either follow the statistics of the artificial language or do not (Frost & Monaghan, 2016; Peña, Bonatti, Nespor, & Mehler, 2002). Furthermore, some studies have taken a four-alternative forced choice approach (Bertels et al., 2015; Siegelman et al., 2017; Yu & Smith, 2007), which makes it more difficult for participants to arrive at the correct answer by chance. Although these variants of the 2AFC task may make it harder to choose the right response, they still require meta-cognitive processing (e.g., reflection over previously-learned material), and thus may still be subject to the limitations outlined above to varying degrees[1].

Taken together, tests of statistical learning in adults arguably rely on processes that are not actively employed during learning, and that may not reveal the full impact or extent

[1] We note here that the use of grammaticality or familiarity ratings for individual items also are subject to similar issues – although these ratings do allow for more detailed analyses of responses (e.g., it is possible to look for different response patterns for the target and foil stimuli).

of variation in learning. Here, we propose that an initial step towards better understanding the computations involved in statistical learning may lie in developing *processing-based* tasks to measure this ability (Christiansen, 2019; for examples, see Gómez, Bion, & Mehler, 2011; Karuza, Farmer, Fine, Smith, & Jaeger, 2014; Misyak et al., 2010; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018). Processing-based tasks aim to tap into the same computations that occur on-line during learning, and thereby circumvent many of the issues associated with reflection-based responses outlined above. Here, we propose that a basic memory process, *chunking* (e.g., Miller, 1956), or *Chunk-and-Pass* processing (Christiansen & Chater, 2016), may provide an on-line, computational account for how statistical regularities are used to form higher-level representations of an input. Chunking-based tasks may thereby serve as a viable processing-based method of testing statistical learning.

1.2. Statistical learning as chunking

Chunking has long been understood as a fundamental attribute of learning and memory (Miller, 1956; Simon, 1974). Some of the earliest investigations into chunking focused on its contribution to expertise within a specific domain (Chase & Simon, 1973) – language acquisition has also been likened to a form of skill learning involving similar processes (Chater & Christiansen, 2018; Christiansen & Chater, 2016; Gobet et al., 2001; Lieven, Pine, & Baldwin, 1997; Tomasello, 1992; 2003). By the time children reach adult-like proficiency, they become experts at processing their native language, a skill that is argued to be driven by their repeated experience with the regularities available in the input. Reframing statistical learning within this usage-based framework, we suggest that the

behavior observed in such experiments may be understood as the statistically-facilitated chunking of coherent structures over time and repeated exposure.

Chunk-and-Pass processing incrementally builds representations (or "chunks") at varying levels of linguistic abstraction as soon as the input is encountered (Christiansen & Chater, 2016). It utilizes all available information – including top-down expectations from previous context and real-world knowledge – to process the current input as rapidly as possible. For example, phonemes may be recoded into words, and words into multiword sequences, and so on until discourse representations are derived. Crucially, the process of repeated chunking across different levels of linguistic representation is thought to rely substantially on sensitivity to distributional regularities. Here, we propose that the same process may also be at play during statistical learning, where chunk formation is contingent upon the statistical structure of the encountered input. Through exposure to a continuous stream of information, learners use distributional information (e.g., the frequent co-occurrence of the syllables "A" and "B") to implicitly chunk the input into coherent units (the formation of the word "AB"). Our chunking framework thus provides concrete predictions about the mechanism that allows larger units to emerge from distributional regularities over smaller units (chunking as a process) and the outcome of these combined computations (chunked representations). Statistical sensitivity may thus be thought of as the cognitive system's primary method of learning distributional regularities, with chunking leading to the formation of concrete, itemized representations based on this statistical information, following the proposals of usage-based theories of language acquisition (Bannard, Lieven, & Tomasello, 2009; Bybee, 2003; Goldberg, 2006; Lieven, 2016; MacWhinney, 1998; Tomasello, 2000).

Computationally, the idea of statistical learning as chunking is supported by a growing body of research. Although the two literatures have historically remained separate (see Christiansen, 2019, for a review), chunk-based computational models can successfully capture word and phrase segmentation based on sensitivity to statistical information (Freudenthal, Pine, & Gobet, 2006; McCauley & Christiansen, 2011; 2014; 2019; Perruchet & Vinter, 1998). Furthermore, while the nature of what is being learned during statistical learning has long been debated, data from both infants and adults suggest that learners appear to represent concrete chunks of information (Slone & Johnson, 2015; Slone & Johnson, 2018; see also the related chunk-based approaches to artificial grammar learning, e.g., Knowlton & Squire, 1994, and Perruchet, 2019, for a review), rather than the statistical relations between elements alone (Elman, 1990; Endress & Mehler, 2009).

Notably, in their comprehensive review on the relationship between distributional sensitivity and chunking, Perruchet and Pacton (2006) outlined three possible scenarios for how the two may be related: 1) that the processes of statistical computation and chunk formation are independent of one another, 2) that they are successive steps in the learning process, with chunks being inferred from prior statistical computations, or, 3) that sensitivity to statistical structure is a byproduct of the chunking process. Although Perruchet and Pacton (2006) themselves support the third notion, and comparable claims have been made by other memory-based computational models (iMinerva; Thiessen & Pavlik, 2013), there is an additional possibility: that the computation of statistics and chunking are parallel processes, working together on-line during learning. For instance, the chunk-based learner model (CBL; McCauley & Christiansen, 2019), operates in this manner. Language acquisition, comprehension, and production are achieved by gradually

building an inventory of words and multi-word units, which are chunked based on the statistical properties of the input. The CBL model uses this inventory to help chunk future input on-line as it is encountered, and can successfully approximate children's comprehension and production performance across 29 languages from 9 language families across 15 genera. While the model has a number of limitations, such as working with pre-segmented input corpora, it nonetheless demonstrates how statistical computation and chunking may operate in parallel, and how such learning can closely approximate real-world linguistic performance.

Behaviorally, insights from the classic memory tasks of nonword repetition (NWR) and serial recall support the idea that chunking-based recall tasks may tap into long-term linguistic representations. While recall tasks have traditionally been viewed as a measure of working memory capacity (i.e., how well individuals can chunk items together and hold them in short term memory; Baddeley, Gathercole, & Papagano, 1998; Gathercole & Baddeley, 1989), it has become increasingly evident that immediate chunking abilities interface with long-term distributional learning, suggesting that long and short-term memory are highly interconnected (e.g., Christiansen, 2019; Jones & Macken, 2018). Through continued exposure to language, individuals learn to chunk recurring lexical and sub-lexical patterns into larger units, which facilitates short-term recall of items that contains such distributional patterns within them (Baddeley, 1964; Botvinick, 2005; Jones, Gobet, Freudenthal, Watson, & Pine, 2014; Jones, Gobet, & Pine, 2007; Jones & Macken, 2015). Prior studies have shown that more word-like nonwords are recalled more accurately (e.g., Archibald & Gathercole, 2006; Gathercole, 1995), which by our account suggests that such nonwords better reflect natural language statistics. Indeed, McCauley,

Isbilen, and Christiansen (2017) employed large-scale corpus analyses to show that nonwords comprising syllable combinations reflecting the co-occurrence statistics of natural language were better recalled than nonwords consisting of the very same syllables but in low-probability combinations. Of particular importance for the current study, Majerus, van der Linden, Mulder and Peters (2004) demonstrated that both children and adults exposed to an artificial grammar containing high- and low-frequency phonological patterns showed better performance on a subsequent NWR task for nonwords that followed the high-frequency phonological patterns. These ideas are in line with findings suggesting that NWR performance draws on pre-existing sub-lexical representations accrued over time (Jones, 2016; Szewczyk, Marecka, Chiat, & Wodniecka, 2018), in the form of high-frequency chunks. Building on this observation, we propose that chunking-based recall tasks may also serve as a valuable tool to gauge statistical learning in the lab.

In the current paper, we leverage the connection between recall and distributional sensitivity by adapting a classic chunking-based memory task to test statistical learning. To this end, we developed the statistically-induced chunking recall (SICR) task. In SICR, participants are exposed to an artificial language consisting of tri-syllabic nonsense words using the same training procedure as the seminal Saffran et al. (1996) study. Following exposure, they are then presented with strings of six-syllables – a number just beyond the threshold of typical working memory abilities (4 ± 1 items; Cowan, 2001) – which participants are asked to immediately recall. Critically, the strings in the SICR task follow one of two formats: they are either composed of two-word combinations from the artificial language, or consist of the same set of syllables presented in a random order. We hypothesized that if participants had chunked the input language into words based on its

distributional statistics, this would lead to significantly higher recall of the experimental items relative to the controls, similar to the results of the NWR and serial recall studies where long-term statistically-based chunking leads to improved memory. Notably, the scoring of SICR is performed syllable-by-syllable, which provides a window into the specific output representations derived from statistical learning tasks. This enables us to gauge the acquisition of specific chunk information by examining trigram scores (i.e., the number of words from the input language that participants correctly recall; Siegelman, Bogaerts, Armstrong, & Frost, 2019).

In two separate experiments, we tested the hypothesis of statistical learning as chunking by gauging the efficacy of SICR in capturing auditory and visual statistical learning in adults. As chunking is a domain-general process, SICR should be equally capable of accounting for both types of statistical learning, but with potential differences in performance due to modality-specific constraints (Frost et al., 2015). We further compared SICR performance with performance on 2AFC, and measured the test-retest reliability of both tasks. We predicted that in addition to providing a more fine-grained assessment of individual variation in statistical learning, that SICR – and thus statistical chunking ability – might also afford a more reliable measure of learning over time than the standard 2AFC task.

2. Experiment 1: Auditory Statistical Learning

Experiment 1 sought to determine whether chunking could account for the kind of statistical learning observed in Saffran et al.'s (1996) classic study. This study demonstrated that 8-month-old infants are sensitive to the statistical patterning present in

speech, and could successfully distinguish between items that followed these statistics from items that did not after a brief period of exposure. In addition to this theoretical proposal, we also test the methodological strengths and limitations of SICR relative to 2AFC, and assess the test-retest reliability of each task in measuring statistical learning in adults.

2.1. Method

2.2. Participants

The sample consisted of 43 Cornell University undergraduate students (32 females, 11 males), with a mean age of 19.75 years (*SD* = 1.43). All participants were native English speakers with no known auditory or language disorders, and were compensated with course credit. One participant was excluded due to a failure to provide responses for the SICR task in Session 1. The analyses reported below were performed on the remaining 42 participants.

2.3. Materials

The artificial language used in this experiment was adapted from Saffran et al. (1996) and consisted of 18 consonant-vowel syllables: *bi, bu, di, du, ga, ka, ki, la, lo, lu, ma, mo, pa, po, ri, ta, ti, to*. These were used to construct 6 tri-syllabic words that served as the input language: *kibudu, latibi, lomari, modipa, tagalu, topoka*. When constructing the words, the positions of the syllables were controlled such that no consonant or vowel always occurred in the same serial position within the words, and the positions of the vowels were further counterbalanced such that vowel repetitions occurred in either the first two syllables of the word, the last two syllables of the word, or in none of the syllables of each word.

Controlling for these factors ensured that no additional cues other than the transitional probabilities of the within-word syllables occurred with one hundred percent regularity.

For 2AFC, six additional nonword foils were created, using the same syllables as the input language. These foils were created by pseudo-randomizing the syllables from the input language in a manner that avoided reusing between-syllable transitional probabilities from the target words. The six foil words were: *dikabi, kigala, lopadu, mamoti, polubu, tatori*. For SICR, 24 recall items were created: 12 experimental items constructed from two-word combinations from the input (e.g., *latibitagalu*), and 12 random items (e.g., *tabigatilula*). Within the experimental items, each of the words from the input language appeared four times: twice as the first word in the test item, and twice as the second word. From these items, a set of 12 complementary foil items were constructed by pseudo-randomizing the syllables in the target items in a manner that avoided the reuse of transitional probabilities from all other test sequences (including the 2AFC foils). The foil strings were intended to serve as a baseline working memory measure, against which performance on the experimental items could later be compared to measure the degree of learning. Using the same method, 12 five-syllable practice items were constructed (6 target items, and 6 foils), following the methodology laid out in standard NWR tasks (Gathercole & Baddeley, 1996). This was to ensure that the amount of post-input exposure would be roughly equal across the counterbalanced conditions (whether participants completed 2AFC or SICR first). All of the SICR test and practice items are listed in Appendix 1.

All of the training and test stimuli were created using the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van der Vrecken, 1996), with each syllable lasting approximately 200 milliseconds, separated by 75 milliseconds of silence. Each

participant received one of four randomized lists for training and SICR. Presentation of the 2AFC items was fully randomized across participants. Both item presentation and data collection used the E-prime 2.0 experiment software (Schneider, Eschman, & Zuccolotto, 2002).

2.4. Procedure

To establish the reliability of SICR and its comparability to 2AFC, the test-retest reliability of each task was assessed within-subjects. Following the procedure used by Siegelman and Frost (2015), participants performed the same experiment twice, with a three-week interval between the two sessions. The same input language and test items were used in both sessions (as in Siegelman & Frost, 2015), but the items in each task were presented in a different randomized order between sessions. The test order condition for each participant was maintained across Sessions 1 and 2: participants who performed 2AFC first followed by SICR in Session 1 were given the same order of tests in Session 2, and vice versa. All other details of the sessions were identical and proceeded as described below.

Participants were first exposed to the artificial language, during which each word was presented 96 times. Additionally, a cover task adapted from Arciuli and Simpson (2012) was administered, in order to ensure active engagement during training. The cover task comprised a target detection task, where participants were asked to respond to syllable repetitions that occurred in immediate succession (e.g., *lolo*) by pressing the spacebar on the keyboard. To this end, three variants of each word were included during exposure, such that the first, second, or third syllable of each word repeated (e.g., *lomari* → *lolomari,*

lomamari, lomariri). Each repetition variant occurred four times during training, resulting in 72 repetitions in total (6 words x 3 variants x 4 exposures). Importantly, at no point was there any reference to language or structure in the instructions of the experiment, nor were participants informed that their knowledge of the training input would be tested. Participants were simply asked to respond to the syllable repetitions in the stream by hitting the space bar each time they heard a syllable repeat. In total, training lasted approximately 11.5 minutes.

Following exposure, both 2AFC and SICR were administered to test statistical learning, and were counterbalanced across participants to control for potential order effects. In 2AFC, each of the six target words were presented with each of the six foil words, yielding 36 trials in total. Each word and foil were presented over headphones one at a time, with a 1000-millisecond silence between the two. Participants were informed that certain triplets of syllables tended to co-occur during the exposure phase, and that they would be assessed on how well they had picked up on this structure. They were then prompted to identify which of the two items had appeared during training. The order of the targets and foils were counterbalanced, such that each foil and word appeared once as the first item of the pair, and once as the second item of the pair.

In SICR, 12 practice trials preceded the 24 experimental trials, resulting in 36 trials in total. This served to keep exposure to the input words approximately equal, regardless of the order in which participants received the two tests (i.e., 2AFC first or SICR first). In this task, participants were told that their ability to repeat the syllables presented in the experiment would be evaluated and were then instructed to repeat the syllables they heard in the correct order to the best of their ability. They were not informed of the strings'

underlying structure. For both the practice and experimental trials, the test items were presented over headphones, after which participants were prompted to repeat the syllables into a microphone. The oral responses were later transcribed by coders who were blind to the purpose of the study.

Prior to subsequent analyses of SICR, the inter-rater reliability of the coders who transcribed the task was assessed. To this end, the full data set (both Sessions 1 and 2) was independently coded using three different pairs of coders. Although the experience levels differed between each pair – one pair had a great amount of experience coding the SICR task, one pair had an intermediate amount of experience, and one pair had no experience coding – on average, significant inter-rater reliability was observed. Specifically, while coders with more experience tended to display higher inter-rater agreement (expert coders: $r(624) = .83$, $p < .0001$; intermediate-experience coders: $r(720) = .84$, $p < .0001$), even novice coders demonstrated a significant degree of inter-rater reliability, $r(672) = .69$, $p < .0001$. After the inter-rater reliability was assessed, each pair of coders was asked to re-visit the transcriptions that they and their coding partner differed on, and determine the appropriate transcription for each divergent item. All further SICR analyses reported below were performed on the corrected transcriptions on which the coders converged. The full details of the coding criteria are reported in the Supplementary Material file. All data and R code are available through the Open Science Framework (https://osf.io/mky4h/).

2.5. Results

2.5.1. 2AFC performance by session

We first assessed 2AFC performance (the proportion of correct target-word identifications), which was significantly above chance (50%) in both sessions (Session 1: $t(41) = 8.94$, $p < .0001$, $d = 1.38$); Session 2: ($t(41) = 10.94$, $p < .0001$, $d = 1.70$). Additionally, 2AFC performance significantly improved from Session 1 to Session 2, $t(41) = 2.72$, $p = .01$, $d = .54$. The mean values of 2AFC performance in both sessions is reported below in Table 1a.

2.5.2. SICR performance by session

The SICR data was scored for accuracy in two different ways: overall accuracy across the entire string (the total number of syllables recalled in the correct serial order; Cowan, Chen, & Rouder, 2004; Fallon, Groves, & Tehan, 1999) and accuracy at the trigram level (the number of syllable triplets or words recalled in the correct order) for both the experimental and random items. Difference scores between each measure (calculated as experimental item score minus random item score) were also assessed for test-retest reliability.

Participants' overall accuracy was significantly higher on the experimental items than on the random items, with participants recalling significantly more syllables in the correct serial order when the test items utilized the statistics of the input language in both Session 1, $t(41) = 6.57$, $p < .0001$, $d = 1.01$, and in Session 2, $t(41) = 9.32$, $p < .0001$, $d = 1.44$. Similarly, trigram recall was significantly higher for the experimental items than the random items in both sessions (Session 1: $t(41) = 7.99$, $p < .0001$, $d = 1.23$); Session 2: $t(41) = 8.12$, $p < .0001$, $d = 1.26$). While performance on the random items did not improve over time (both $p = .55$ or above), performance on the experimental items did significantly improve in Session 2 relative to Session 1 (overall accuracy: $t(41) = 2.27$, $p = .03$, $d = .31$;

trigram recall: $t(41) = 2.54$, $p = .02$, $d = .35$). The means of each SICR measure across both sessions can be found below in Table 1b. Additionally, the serial position curves of the SICR items in both sessions are depicted in Fig. 1.

[insert Tables 1a and 1b here]

[insert Figure 1 here]

2.5.3. SICR error regularization

Taking advantage of the rich data provided by SICR, performance was further assessed for error regularization, which examines the degree to which production errors on the random items reflect a bias towards higher-probability sequences (Botvinick & Bylsma, 2005). If participants have internalized the statistics of the artificial language, then presumably their mispronunciations of the random items, which contain the same syllables as the target words, may show a tendency toward statistically-legal patterns. To gauge whether this occurred, SICR random item recall was analyzed for whether participants' incorrect productions tended to incorporate the regularities present in the artificial language, including both legal bigram and trigram information. For example, for the random item *tabigatilula* (the foil for the target item: *latibitagalu*), if a participant's mispronunciation contained any legal bigrams from the target words that did not comprise a full trigram (e.g., *lati, tibi, taga, galu*) or any legal trigrams (e.g., *latibi, tagalu*), participants would be awarded one point for each statistically-legal error. These errors were calculated separately: if a participant produced a full legal trigram from the artificial language, this was counted only as a trigram, and not as two separate bigrams (e.g., a point was awarded

for the trigram *tagalu,* but no points were awarded for the bigrams therein: *taga* and *galu).* The total proportions of these errors were then calculated for the entire sample of participants. Each session was considered separately.

In Session 1, out of 516 total random trials across all participants, only 28 (5%) included a perfect recall response (e.g., every syllable in the random string was recalled exactly correctly). These trials were then discarded from the analyses, and the remaining 488 trials were analyzed for error regularization. Overall, error regularization comprised a small amount of production errors. Only 4 featured erroneous recall of a legal trigram (or a full word) from the artificial language (.8% of trials). An additional 61 trials (13%) featured a bigram combination corresponding to a legal sequence from the language (e.g., *tabigati**galu***, where *galu* is a legal bigram). A further 2 trials (.4%) included 2 (as opposed to just 1) legal syllable bigrams from the language which did not make up part of a larger legal trigram sequence (e.g., ***lati**gati**galu***, where *lati* and *galu* are legal bigrams that do not comprise a larger trigram).

In Session 2, out of 516 total trials in the random condition, only 15 (3%) included a perfect recall response. Out of the remaining 501 trials, the degree of error regularization was small, but slightly higher than in Session 1. In total, 7 trials featured erroneous recall of a full word from the artificial language (1.4% of trials), while an additional 74 trials (15%) featured bigram combinations corresponding to a legal sequence from the language. A further 4 trials (.8%) included 2 (as opposed to just 1) legal syllable bigrams from the language which did not make up part of a larger legal trigram.

2.5.4. Comparisons between 2AFC and SICR

Prior to measuring reliability, the data were first analyzed for test order effects (whether participants performed 2AFC followed by SICR in a given session, or performed SICR followed by 2AFC). For 2AFC, there was no significant effect of order in either Session 1 ($F(1,40) = .70$, $p = .41$), or in Session 2 ($F(1,40) = .87$, $p = .36$). For SICR, there was a significant effect of order on difference score measures in Session 1[2], with participants who performed 2AFC prior to SICR demonstrating larger difference scores than those who performed SICR first (overall accuracy: $F(1,40) = 7.65$, $p = .01$, $R^2 = .16$; trigram recall: $F(1,40) = 4.08$, $p = .05$, $R^2 = .09$). In Session 2, a one-way ANOVA replicated this effect with the overall accuracy difference score ($F(1,40) = 8.18$, $p = .01$, $R^2 = .17$), however, the trigram difference scores in the second session were not significantly impacted by order, $F(1,40) = 2.89$, $p = .10$, $R^2 = .07$.

Next, the degree to which performance on 2AFC and SICR were reliably correlated within sessions was assessed. The results revealed that in Session 1, 2AFC performance was significantly correlated with the SICR overall accuracy difference score, $r(40) = .31$, $p = .04$, but only marginally correlated with the trigram recall difference score, $r(40) = .30$, $p = .053$. These effects were slightly stronger in Session 2, with 2AFC correlating with both the overall accuracy difference score, $r(40) = .43$, $p = .004$), and the trigram recall difference score, $r(40) = .48$, $p = .001$).

Finally, the test-retest reliability of 2AFC and all SICR measures were assessed (Fig. 2). 2AFC performance was not reliable between Sessions 1 and 2, $r(41) = .19$, $p = .24$. However, all SICR measures demonstrated significant test-retest reliability. Overall

[2] In a pilot study comparing SICR and 2AFC (Isbilen, McCauley, Kidd, & Christiansen, 2017), the opposite order effect was found, with order significantly impacting 2AFC performance ($t(68) = 12.06$, $p < .0001$), but not SICR performance. This suggests that these results may be somewhat unreliable.

accuracy was highly reliable for the experimental items ($r(41) = .63$, $p < .0001$), the random items ($r(41) = .58$, $p < .0001$), and the difference score between the two ($r(41) = .40$, $p = .008$). The same results held for trigram recall on the experimental items ($r(41) = .62$, $p < .0001$), the random items ($r(41) = .54$, $p = .0003$), and the difference score between the two ($r(41) = .50$, $p = .001$). Additionally, a partial correlation run between the Session 1 and Session 2 experimental SICR scores reveals that the test-retest reliability of the measure remains strong, even when controlling for baseline working memory (Session 1 random item recall), $r(41) = .64$, $p < .0001$. Below, Fig. 2 depicts the correlations between these different measures.

[Insert Figure 2 here]

2.5.5. The impact of natural language statistics on 2AFC and SICR

We sought to gain a measure of whether, and to what extent, natural language statistics shaped performance on test items in the experiment. Our assumption was that if the tasks used to test in-lab statistical learning are indeed tapping into the same learning mechanisms involved in real-world language acquisition, then they should in theory also capture some degree of facilitation from natural language statistics. Additionally, if SICR is indeed sensitive to real-world distributional information, then this task may provide a useful tool for future statistical learning research that seek to further investigate this phenomenon.

For this purpose, we extracted data from two large corpora of spoken English: the Fisher (Cieri, Miller, & Walker, 2004) and Switchboard (Godfrey, Holliman, & McDaniel, 1992) corpora. The corpora were combined and each utterance was converted to a string

of phonemes using a speech synthesis engine (http://espeak.sourceforge.net). Statistics for phoneme pairs and triplets (bigrams and trigrams) were then extracted and used to evaluate the test items from Experiment 1. There were several reasons for conducting our analyses at the level of phoneme bigrams and trigrams. Firstly, there is no standardized corpus of English with coded syllable boundaries, which prompted us to focus on the more fine-grained level of phonemes. Secondly, as our stimuli were designed specifically to yield negligible overlap with natural language in terms of 1) syllable-to-syllable transitional probabilities, and 2) multi-syllable chunk strength, we focus here on phonemes in order to account for more fine-grained information which was not possible to control, such as that involving within-syllable transitions as well as within-syllable coherence.

The analyses focused, separately for bigrams and trigrams, on two measures: mean chunk strength (how frequently the phoneme sequences making up the test string occur in natural language) and mean transitional probability (how frequent the transitions between phonemes occur in natural language when measured using bigrams or trigrams of phonemes).

We constructed four linear mixed-effects models, two for each type of statistic (bigram vs. trigram; separate models were used due to collinearity between the two) and each measure of distributional strength (chunk strength vs. transitional probability; separate models used once more due to collinearity between measures). Each model sought to predict the total SICR score in each trial, featured subjects and items as random effects, with condition (2: Experimental vs. Random), the natural language statistic, and the interaction term as fixed effects.[3]

[3] Due to singular fit, the random effects terms were simplified to remove random slopes.

Average trigram chunk strength did not emerge as a significant predictor of SICR performance, while average bigram chunk strength was a significant predictor of SICR score (B = .16, t = 2.33, p < .05), with a significant bigram chunk strength by condition interaction (B = -.319, t = -2.30, p < .05), indicating a decreased effect of chunk strength for random as opposed to experimental items.

Mean transitional probability for trigrams did not significantly predict SICR score (B = .17, t = 1.91, p = .07), but there was a significant transitional probability by condition interaction (B = -.30, t = -2.43, p < .05), indicating a decreased effect of transitional probability for random as opposed to experimental items. Transitional probability for bigrams, however, did not reach significance as a predictor of SICR score.

We conducted a parallel set of analyses of the 2AFC scores on a trial-by-trial basis. We utilized mixed-effects logistic regression models, with 2AFC accuracy coded as a binary variable. All natural language predictors were the same as the analyses for the SICR scores, with one important modification: because 2AFC tasks involve simultaneous exposure to two different strings, we calculated the difference in the relevant statistic across the experimental and random items. For example, the bigram chunk strength model would include the difference in mean chunk strength for the experimental and random sequences in each given trial as a single predictor. Natural language predictors did not rise to any level of significance in any of the models. The same outcome resulted from a set of models in which only the natural language statistics of the experimental items was considered.

2.6. Discussion

In Experiment 1, we tested whether chunking as measured via serial recall might explain the learning effects found in classic embedded triplet statistical learning tasks (e.g., Saffran et al., 1996). We observed that participants' recall of the experimental items was significantly higher than that of the random items, both for overall accuracy across the entire string, and for the number of trigrams recalled, with the latter suggesting chunked representations of the input (Siegelman et al., 2019). Over a brief period of exposure, participants' experience with the statistics of the artificial language led to observable changes in short-term memory processing, such that items from the speech stream were better recalled than random combinations of the same syllables. Furthermore, learning also resulted in a small but observable degree of error regularization in the random items (Botnivick & Bylsma, 2005), with participants regularizing their productions of the control items to reflect the bigrams and trigrams present in the artificial language (with the effect being particularly prominent for bigrams). These findings mirror those that demonstrate how long-term distributional learning mediates short-term memory, extended here to the context of in-lab statistical learning.

It is noteworthy that the natural language statistics only affected SICR performance, and did so in both sessions. We take the sensitivity of the experimental SICR items to natural language distributional patterns as further evidence that the task may be capturing or reliant on the same processes involved in real-world statistical learning. That this effect is especially pronounced in recall of the experimental items suggests that the distributional patterns of phonological sequences in natural language influence how new statistics are acquired in the lab (and is something that could be manipulated explicitly in future studies). Recall arguably employs many of the same mechanisms – such as the rapid chunking of

input – that are leveraged for natural language processing, which may explain why the task was more readily affected by English phonological regularities. However, it is also possible that SICR may be a better measure of distributional sensitivity in general. NWR and serial recall tasks can reliably capture individuals' sensitivity to the statistics of their natural language and learning-based changes in memory (Jones, 2012; Jones et al., 2007; Szewczyk et al., 2018). In fact, decades of research show that recall is heavily influenced by the distributional patterns present in natural language (Baddeley, 1964; Botvinick, 2005); here, we extend these findings by showing that SICR captures both artificial and natural linguistic statistical patterns. SICR thus offers insights into different levels of linguistic entrenchment, the products of both long and short-term learning.

Our results confirm that while both 2AFC and SICR provide estimates of learning, SICR demonstrates a high degree of test-retest reliability on all sub-components of the measure. By contrast, 2AFC fails to demonstrate significant reliability. These findings diverge from those of Siegelman and Frost (2015), who report significant test-retest reliability of 2AFC in measuring auditory-linguistic statistical learning. However, their method for the construction of the 2AFC foil items differed slightly from those utilized here: while the structure of the foil items in the current experiment were fully pseudo-randomized (Saffran, Newport, & Aslin, 1996), the foils in Siegelman and Frost (2015) were part-words that spanned word boundaries. They also tested learning of only a subset of the input words (six out of the twelve words presented during training) whereas here, no input words were excluded from test. Additionally, the 2AFC task implemented here may have been easier than that employed by Siegelman and Frost (2015), as distinguishing words from random combinations may be easier than distinguishing words from part-

words. It is possible that these methodological differences may have contributed to our conflicting findings.

For SICR, the test-retest reliability was highest for recall of the experimental items, which suggests that statistical chunking abilities may be consistent over time. Notably, recall of the random items was also highly consistent between sessions, indicating that the random items are a reliable measure of base-line working memory abilities. The slightly lower reliability of the SICR difference scores may be explained by the additional noise that arises from the construction of the measure: while raw scores have only one source of noise present in the data, the difference scores combine the noise from both measurements (Caruso, 2004; Willet, 1988; Zimmerman & Williams, 1998; Zumbo, 1999). Difference scores also reduce the range of variability across participants, as the difference between experimental and random recall tends to be more similar between participants than the amount of variability within item type (experimental or random; Hedge, Powell, & Sumner, 2018). This in turn may contribute to the difference scores' somewhat diminished reliability relative to the raw SICR measures.

3. Experiment 2: Visual Statistical Learning

The processes of statistical learning and chunking are not specific to spoken language – rather, they extend to learning in a diverse variety of domains and modalities. The objective of Experiment 2 was thus to evaluate whether the chunking behavior observed in Experiment 1 might extend to statistical learning in other modalities. As an initial foray into this question, the same test-retest procedure as Experiment 1 was applied to the

statistical learning of visually-presented syllables. We expected to replicate the results of Experiment 1, with potential differences arising from modality-specific constraints.

3.1. Method

3.2. Participants

A separate sample of 40 Cornell University undergraduates was recruited, composed of 25 females, and 15 males, with a mean age of 19.75 (*SD* = 1.26). All participated in exchange for course credit and were native English speakers, with no known language or visual impairments.

3.3. Materials

Experiment 2 presented the same input language as Experiment 1, using written transcriptions of the syllables rather than auditory playback. Transcriptions of the same 2AFC foils and SICR practice and test items were also presented to assess word learning, in order to ensure that the two experiments were as similar as possible despite the differences in modality. All syllables in this experiment were presented one at a time in the center of the screen, in lowercase 72-point Arial font. The experiment was programmed in E-prime 2.0 (Schneider et al., 2002).

3.4. Procedure

In line with the procedure of the first experiment, Experiment 2 consisted of three separate tasks: exposure to the artificial language, 2AFC, and SICR. Exposure to the artificial language was self-paced, in order to maintain participant engagement during training.

Syllables were presented sequentially, one after another. Participants were instructed to press the space bar as soon as they had finished reading each syllable, after which the next syllable would immediately appear on the screen. There were no pauses or blank screens in between syllable presentations. A fixed minimum of 250 milliseconds was implemented, in order to provide the same baseline exposure time as the syllables in the auditory experiment. Similar to the auditory version of the task, training lasted approximately 11.5 minutes on average, with some individual variation depending on reading rate.

Following exposure, participants completed both 2AFC and SICR, using the same test items as Experiment 1. The order of each test was counterbalanced to control for task order effects. Unlike the exposure phase, the tests relied on a fixed presentation rate rather than a self-paced rate, to ensure that exposure to the test syllables was uniform across participants. For both the 2AFC and SICR tasks, each syllable was presented one at a time, with each appearing on the screen for 650 milliseconds with no pauses in between. In the 2AFC task, the foil and target words were presented sequentially, separated by a fixation cross that appeared on the screen for 1000 milliseconds. In the SICR task, participants typed their responses instead of saying them out loud. Both tasks utilized the same instructions from Experiment 1.

After the completion of Session 1, participants were asked to return to the lab and perform the experiment again, after a three-week delay. The same language and test items were used in each session, presented in different randomized orders. The test-order (2AFC first/SICR second or SICR first/2AFC second) was preserved within subjects across each session.

3.5. Results

3.5.1. 2AFC performance by session

2AFC performance (the proportion of correctly-identified target words) was significantly above chance (50%) in both Session 1 ($t(39) = 7.76$, $p < .0001$, $d = .70$) and Session 2 ($t(39) = 8.99$, $p < .0001$, $d = 1.42$). Additionally, performance on 2AFC significantly improved between sessions ($t(39) = 3.05$, $p = .004$, $d = .46$). The mean values of 2AFC performance in each session are reported in Table 2a.

3.5.2. SICR performance by session

The SICR analyses were performed on the participants' typed responses. Responses that had fewer than 6 syllables were amended using the same anchoring procedure that was implemented for the auditory experiment (Dollaghan & Campbell, 1998; Weismer, Tomblin, Zhang, Buckwalter, Chynoweth, & Jones, 2000), reported in Appendix 2. The results revealed that recall of the experimental SICR items was significantly higher than that of the random items. This pattern held for the total number of syllables recalled in Session 1 ($t(39) = 4.60$, $p < .0001$, $d = .73$) and in Session 2 ($t(39) = 5.44$, $p < .0001$, $d = .86$). Similarly, significantly more trigrams were recalled in the experimental than the random items in Session 1 ($t(39) = 4.84$, $p < .0001$, $d = .77$) and in Session 2 ($t(39) = 5.73$, $p < .0001$. $d = .91$). SICR performance on the experimental items improved between sessions ($t(39) = 3.99$, $p = .0003$, $d = .53$), although overall accuracy of the random items did not ($t(39) = 1.60$, $p = .12$, $d = .17$). Trigram recall on both the experimental items ($t(39) = 4.84$, $p < .0001$, $d = .65$) and the random items ($t(39) = 2.32$, $p = .026$, $d = .25$) improved

across sessions. The mean performance on all SICR measures are reported in Table 2b, and their corresponding serial position curves are depicted in Fig. 3.

[Insert Tables 2a and 2b here]

[Insert Figure 3 here]

3.5.3. SICR error regularization

Following the same procedure as Study 1, the degree of participant error regularization on the random SICR items was analyzed to assess whether participants' production errors reflected a bias toward the statistics of the artificial language. As before, each random item that contained any errors was analyzed for whether these mispronunciations contained legal bigram or trigram information. Like in Experiment 1, these errors were calculated separately: if a participant produced a full legal trigram, this was scored only as a trigram, and not as two separate bigrams. The total proportions of these errors were then calculated for the entire sample, and each test session was considered separately.

In Session 1, out of 480 total trials in the random condition, 77 (16%) included a perfect recall response. These trials were then excluded from the analyses, and the remaining 403 trials were considered. As in Experiment 1, only a small number of trials featured error regularization. In total, 4 featured erroneous recall of a legal syllable trigram (or a full word) from the artificial language (1% of trials). An additional 62 trials (15%) featured single bigram combinations corresponding to a legal sequence from the language (e.g., *tabigati**galu***, where *galu* is a legal bigram). One further trial (.2%) included 2 (as opposed to just 1) legal syllable bigrams from the language which did not make up part of

a larger legal trigram sequence (e.g., ***lati****gati****galu*** where *lati* and *galu* are legal bigrams that together do not comprise a larger legal trigram).

In Session 2, out of 480 total trials in the random condition, 92 (19%) included a perfect recall response. These trials were then excluded from the analyses. As with Experiment 1, the total proportion of error regularization was small, but slightly higher in Session 2 relative to Session 1. Out of the remaining 388 trials, 10 featured erroneous recall of a legal syllable trigram from the artificial language (3% of trials). An additional 63 trials (16%) featured a single bigram combination corresponding to a legal sequence from the language. A further 4 trials (1%) included 2 (as opposed to just 1) legal syllable bigrams from the language which did not make up part of a larger legal trigram word.

3.5.4. Comparisons between 2AFC and SICR

As the first measure of comparison, both tasks were analyzed for task order effects. Although in both sessions, 2AFC performance was slightly higher when it was performed after SICR, a one-way ANOVA revealed no significant order effects in either Session 1 ($F(1,38) = 1.35$, $p = .25$, $R^2 = .03$) or Session 2 ($F(1,38) = .57$, $p = .45$, $R^2 = .02$). There were no order effects on any of the SICR measures (all $p = .34$ or above).

Next, correlations between the tasks were analyzed. 2AFC was not significantly correlated with the SICR difference scores in Session 1. These results held for both the overall accuracy difference scores ($r(38) = .18$, $p = .30$), and the trigram recall difference scores ($r(38) = .16$, $p = .33$). However, by Session 2, 2AFC was significantly correlated with both overall accuracy ($r(38) = .44$, $p = .004$) and trigram recall ($r(38) = .55$, $p = .0002$).

To determine the test-retest reliability of each task, correlations were run between participants' scores on each measure at Session 1 and Session 2 (Fig. 4). 2AFC performance demonstrated significant test-retest reliability between sessions ($r(38) = .55$, $p = .0002$). For the SICR task, recall was highly reliable between sessions, both on the experimental items ($r(38) = .66$, $p < .0001$), and on the random items ($r(38) = .81$, $p < .0001$). This correlation remains significant, even when controlling for baseline working memory (random item recall in Session 1), $r(38) = .33$, $p = .04$. In addition, the same statistics were performed on the SICR overall difference scores (calculated as experimental minus random). The overall SICR difference score was only marginally reliable between sessions ($r(38) = .30$, $p = .059$). The trigram recall results were highly reliable for the experimental items ($r(38) = .66$, $p < .0001$), the random items ($r(38) = .78$, $p < .0001$), and the difference score between the two ($r(38) = .37$, $p = .02$).

[Insert Figure 4 here]

3.5.5. The impact of natural language statistics on 2AFC and SICR

As in Experiment 1, we once more sought to measure whether, and to what extent, natural language statistics shaped test performance. For this purpose, we extracted data from the written portion of the American National Corpus (ANC; Ide & Macleod, 2001). Statistics for letter pairs and triplets (bigrams and trigrams) were then extracted and used to evaluate the test items from Experiment 2.

The analyses focused, separately for bigrams and trigrams, on two measures: mean chunk strength (how frequent the letter sequences making up the test string are in natural

language) and mean transitional probability (how frequent the transitions between letters are in natural language when measured using bigrams or trigrams of letters).

We constructed four linear mixed-effects models, two for each type of statistic (bigram vs. trigram) and each measure of statistical strength (chunk strength vs. transitional probability). Each model sought to predict the total SICR score in each trial, featured subjects and items as random effects, with Condition (2: experimental vs. random), the natural language statistic, and the interaction term as fixed effects. Neither bigram nor trigram chunk strength emerged as a significant predictor of SICR score, nor did transitional probability.

We conducted a parallel set of analyses of the 2AFC scores on a trial-by-trial basis. Following the same procedure outlined in Experiment 1, we utilized mixed-effects logistic regression models, with 2AFC accuracy coded as a binary variable. Once more, natural language predictors did not rise to any level of significance in any of the models for the 2AFC task.

3.6. Discussion

The results of Experiment 2 replicate the key findings of Experiment 1 by demonstrating that SICR can, in fact, provide an effective measurement of statistical learning ability across modalities. The experimental SICR items were consistently recalled more accurately than the random items, and evidence of the acquisition of trigram chunks was once again observed. Similar levels of error regularization – where participants normalize their productions of the random sequences in a manner that resembles the statistics of the

artificial language – are also observed across the two experiments. However, the results of Experiment 2 diverge from Experiment 1 in a number of ways.

First, the order effect that was found for the overall SICR difference scores in Experiment 1 was not present in Experiment 2, suggesting that the order effect observed in the previous experiment may have been limited to that sample alone. Similarly, there was no correlation between 2AFC and SICR performance in Session 1 of this experiment, although a correlation was observed by Session 2. This may be due to participants' relying on long-term knowledge of the strings acquired in Session 1 to guide performance in Session 2. It is also likely that there is more reflection involved in the visual SICR task than in the auditory version. In the typed SICR task, participants are afforded the ability to revisit and revise their responses – while some degree of self-correction was observed in the auditory version of the task, participants' verbal responses were not available for playback during the experiment, and they could only self-correct from memory alone. Thus, while some studies show that different measures of statistical learning are not always correlated with one another, even when they intend to measure learning of the same material (Arnon, 2019; Erickson, Kaschak, Thiessen, & Berry, 2016; Siegelman & Frost, 2015; Misyak & Christiansen, 2012), the correlation observed between 2AFC and SICR may in part stem from this shared reflection-based component, as reflection-based tasks tend to correlate more highly with other reflection-based tasks than with processing-based tasks (Isbilen, Frost, Monaghan, & Christiansen, 2018).

With the exception of the overall difference score, which was only marginally reliable, the results from the visual statistical learning experiment revealed significant test-retest reliability for all SICR measures in a manner that is fairly comparable to those

observed in Experiment 1. As in the previous experiment, however, the difference scores were generally less reliable than the raw scores, suggesting that difference scores, on the whole, may be a less reliable measure of learning (Hedge, et al., 2018). This is further highlighted by the fact that the test-retest reliability of the experimental items remains significant when controlling for baseline working memory, suggesting that the task does in fact reliably capture individual differences in learning abilities. Thus, SICR is arguably a reliable task – all subcomponents of the measure, including both the raw scores and trigram difference score – demonstrate significant test-retest reliability, with the exception of the marginal finding for the overall difference score.

Unlike Experiment 1, 2AFC performance in the visual experiment demonstrated significant test-retest reliability. It is possible that the cognitive technology of reading enables participants to develop a more explicit awareness of the statistical structure than is possible in the auditory version of the task. This may have translated into the increased reliability of the measure in this experiment, as this version of 2AFC may have been more closely aligned with the kind of explicit knowledge participants had accrued during training. As far as we are aware, this is the first study to report the test-retest reliability of the statistical learning of visual-linguistic material, although Siegelman and Frost (2015) report significant test-retest reliability for visual-nonlinguistic statistical learning using 2AFC that is fairly comparable to ours (r = .58, as compared to our obtained findings of r = .55). Thus, 2AFC appears to be a reasonably reliable method of testing visual statistical learning in adults (for evidence from children, see Torkildsen, Arciuli, & Wie, 2019).

Unlike in the auditory experiment, natural language statistics had no effect on visual SICR performance. It may be the case that on the auditory SICR task, performance was

more heavily affected because individuals in general have considerably more experience with auditory-sequential than visual-sequential linguistic stimuli. That is, the auditory SICR task is arguably more similar to the constraints of natural spoken language processing, where one must rapidly process the auditory stimulus and produce a response – a latency that lasts only 200 milliseconds on average across languages (Levinson, 2016). By contrast, the presentation of stimuli in the visual SICR task resembles everyday reading less closely.

4. General Discussion

Understanding the computations involved in statistical learning – long assumed to play a pivotal role in language acquisition – has been a hotly debated topic in the language and cognitive sciences. In the current paper, we investigated the idea of statistical learning as chunking by employing a novel chunking-based recall task to measure learning in adults.

Our results demonstrate that SICR can successfully capture the statistical learning of both auditory-linguistic and visual-linguistic input using the standard Saffran et al. (1996) embedded triplet paradigm. Just as enhanced memory for real-world statistical regularities is reliably observed in nonword repetition and serial recall tasks (Archibald & Gathercole, 2006; Gathercole, 1995; Jones & Macken, 2015), after a brief period of exposure to an artificial language, we were able to simulate these same results in the lab. Through exposure to the language, participants' chunking of recurrent sub-patterns facilitates their retention in long-term memory. This is evident in their improved recall of the experimental items. Furthermore, we observe evidence of word-level chunking, with participants chunking recurrent syllable combinations into individual words on the basis of

transitional probability information, as evidenced by the trigram recall scores. This suggests that rather than representing transitional probability information alone, that individuals acquire specific, concrete items, similar to item-based theories of children's natural language acquisition (Bannard, et al., 2009; Bybee, 2003; Goldberg, 2006; Lieven, 2016; MacWhinney, 1998; Tomasello, 2000).

Our framework differs from previous accounts in that we view transitional probability sensitivity and chunk formation as interconnected processes, rather than as two points on a continuum, with chunking proceeding on the basis of statistical computations. It also diverges somewhat from other memory-based approaches to statistical learning, where statistical sensitivity is seen as a consequence of basic memory processing (Perruchet & Vinter, 1998; Thiessen & Pavlik, 2013), rather than existing as a computational process in its own right. In statistical learning experiments, participants appear to represent both transitional probability (bigram) and chunk information (c.f., Siegelman, et al., 2019), although transitional probability information may also be conceptualized as sub-lexical chunks (Jones, 2016; Jones et al., 2014). The statistical-chunking hypothesis thus provides a middle-ground between recognition-based and statistically-based models of language acquisition. Compared to statistically-based models that rely solely on the calculation and identification of transitional probabilities, recognition-based models provide a closer fit to human statistical learning data (French, Addyman, & Mareschal, 2011; Perruchet, Poulin-Charronnat, Tillman, & Peereman, 2014), through the recognition of familiar, learned chunks of information. However, one limitation of recognition-based models is that they historically have not included the kind of sensitivity to transitional probabilities that learners nonetheless exhibit.

To date, at least one model, CBL (McCauley & Christiansen, 2019), which employs statistical learning and chunking as parallel processes, can successfully approximate children's language acquisition, comprehension and production across multiple languages, via sensitivity to backward transitional probabilities. This model also allows for the active prediction of upcoming elements using previously-learned information, similar to natural language acquisition. Behaviorally, our results from using SICR to measure the statistical learning of non-adjacent dependencies demonstrate that chunking can also capture the learning and generalization of non-adjacent structures, and is thus not limited to the acquisition of adjacent distributional patterns alone (Isbilen et al., 2018; submitted). We thus suggest that rather than being limited to in-lab statistical learning studies, statistically-based chunking may also extend to natural language acquisition as well, with distributional sensitivity and chunking working as interactive processes.

Methodologically, SICR offers more specific insights into the output representations that arise from learning. It also lends itself to examining the impact of statistics in ways that are not possible with traditional 2AFC tasks, such as examining the degree of error regularization present in recall of the random items, as an additional measure of learning. In the auditory domain, SICR was more reliable at the individual level than the classical 2AFC task. This may be because 2AFC violates one of the central assumptions of statistical learning: that it is largely implicit. Asking participants to explicitly reflect on knowledge that may not be available to consciousness raises several problems. For instance, participants may exhibit certain preferences towards choosing one type of item over another (e.g., always choosing the second item; Siegelman et al., 2017), or may simply differ in their reflective abilities (Christiansen, 2019). Indeed, as discussed by Kidd,

Donnelly and Christiansen (2018), performance on many psycholinguistic tasks are influenced by multiple cognitive processes beyond the component of interest (e.g., as investigated via the Drift-Diffusion model; Ratcliff, 1978). It may thus be the case that the means of forced-choice tasks reflect a composite of different abilities, rather than the targeted cognitive process alone (see also Frost et al., 2015, for a similar discussion in the context of individual differences in statistical learning).

Recent years have seen an increasing use of more dynamic measures of statistical learning, such as the evaluation of reaction times during exposure (Franco et al., 2015; Karuza et al., 2014; Misyak et al., 2010; Siegelman et al., 2018; Qi, Sanchez, Georgan, Gabrieli, & Arciuli, 2019). Other studies have successfully gauged statistical sensitivity to novel phonotactic constraints utilizing sequence production tasks that analyze participants' production errors, and find that after extended training, error patterns reliably reflect the newly-acquired phonotactic structures as they become better learned (Dell, Reed, Adams, & Meyer, 2000; Warker & Dell, 2006; Warker, Dell, Whalen, & Gereg, 2008) – effects which have been replicated in the non-linguistic domain (Anderson & Dell, 2018). Here, we contribute another tool to this endeavor.

Although serial recall using pseudo-words does exist, to our knowledge, few studies of this nature have preceded recall with a statistical learning-style familiarization phase (with the notable exceptions of Botnivick & Bylsma, 2005; Conway, Bauernschmidt, Huang, & Pisoni, 2010; Majerus et al., 2004). In other words, while previous studies using recall have largely focused on the entrenchment of known statistics, fewer have made the connection to the acquisition of novel statistics. We thereby highlight how recall tasks can be seen as a proxy for statistical learning: just as nonword repetition and serial recall tap

into individuals' sensitivity to natural language distributional regularities, SICR can be seen as a processing-based task that taps into the same chunking processes using artificial language statistics (see also Christiansen, 2019). Recall tasks like SICR may thus also potentially better correlate with individual differences in language learning, much like how nonword repetition reliably predicts language skills and outcomes in children and adults (Dollaghan & Campbell, 1998; Gathercole, 2006; Gathercole, Willis, Baddeley, & Emslie, 1994; Gupta, 2003), including in second-language learning (Service, 1992; Service & Kohonen, 1995).

In the current paper, we sought to better specify the processes involved in statistical learning by more closely aligning the computations relied upon during learning and test. Specifically, we proposed chunking as the process by which the cognitive system uses statistical regularities to form higher-level representations, with chunked representations of the input as the outcome of learning. We thus suggest that a shift towards processing-based measures of learning, in comparison to reflection-based measures, may offer clearer insights into the extent and mechanisms of statistical learning, and lead to a more comprehensive understanding of the phenomenon as a whole.

Acknowledgements

We thank Dante Dahabreh, Phoebe Ilevebare, Eleni Kohilakis, Jake Kolenda, Farrah Mawani, Jeanne Powell, and Olivia Wang for their help collecting and coding data. We also thank two anonymous reviewers for their insightful comments on an earlier version of the manuscript. This research was in part supported by the NSF GRFP awarded to ESI (#DGE-1650441).

References

Anderson, N. D., & Dell, G. S. (2018). The role of consolidation in learning context-dependent phonotactic patterns in speech and digital sequence production. *Proceedings of the National Academy of Sciences*, *115*, 3617-3622.

Archibald, L. M., & Gathercole, S. E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research*, *49*, 970-983.

Arciuli, J. (2017). The multicomponent nature of statistical learning. *Philosophical Transactions of the Royal Society B, 372*, 1711.

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

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: Past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*, 20160047.

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality, *Behavior Research Methods.* https://doi.org/10.3758/s13428-019-01205-5

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.

Baddeley, A. D. (1964). Immediate memory and the "perception" of letter sequences. *Quarterly Journal of Experimental Psychology, 16*, 364–367.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158-173.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284-17289.

Batterink, L.J., Reber, P.J., Neville, H. J., & Paller, K.A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language, 83,* 65-78.

Bertels, J., Boursain, E., Destrebecqz, A., & Gaillard, V. (2015). Visual statistical learning in children and young adults: how implicit? *Frontiers in Psychology*, *5*, 1541.

Botvinick, M. M. (2005). Effects of domain-specific knowledge on memory for serial order. *Cognition*, *97*, 135-151.

Botvinick, M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 351-358.

Bybee, J. (2003). *Phonology and Language Use* (Vol. 94). Cambridge University Press, Cambridge.

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. European Journal of Psychological Assessment, 20(3), 166–171.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Chater, N., & Christiansen, M. H. (2018). Language acquisition as skill learning. *Current Opinion in Behavioral Sciences*, *21*, 205-208.

Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*, 468-481. https://doi.org/10.1111/tops.12332.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* e62.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: A resource for the next generations of speech-to-text. In *Proceedings of the Language Resources and Evaluation Conference*. Lisbon, Portugal: European Language Resources Association.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15*, 634-640.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1355-1367.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*, 1136-1146.

Dutoit, T., Pagel, N., Pierret, F., Bataille, O., & Van Der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the Fourth International Conference on Spoken Language Processing* (pp. 1393-1396). Philadelphia, USA: ICSLP.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science,* 14, 179-211.

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

Erickson, L., Kaschak, M., Thiessen, E., & Berry, C. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra: Psychology*, *2*, 1-17.

Evans J.L., Saffran J.R., Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*, 321–335.

Fallon, A. B., Groves, K., & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task: When CAT helps RAT, but not MAN. *International Journal of Psychology, 34*, 301-307.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499-504.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation. *Experimental Psychology, 62*, 346–351.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614-636.

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*, 277-310.

Frost, R., Armstrong, B. & Christiansen, M.H. (2019). Statistical learning research: A critical review and possible directions. *Psychological Bulletin, 145,* 1128–1153.

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*, 117-125.

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*, 1243-1252.

Frost, R.L.A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech, *Cognition, 147,* 70-74.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*, 83–94.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513-543.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*, 200-213.

Gathercole, S. E., & Baddeley, A. D. (1996). The Children's Test of Nonword Repetition. London: Psychological Corporation.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103-127.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*, 236-243.

Godfrey, J. J, Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 517–520). Washington, DC: IEEE Computer Society.

Goldberg, A.E. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford University Press, Oxford.

Gómez, D. M., Bion, R. A., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes*, *26*, 212-223.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology: Section A*, *56*, 1213-1236.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166-1186.

Ide, N., & Macleod, C. (2001). The American national corpus: A standardized resource of American English. In *Proceedings of Corpus Linguistics* (pp. 1-7). Lancaster, UK: Lancaster University Centre for Computer Corpus Research on Language.

Isbilen, E.S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018). Bridging artificial and natural language learning: Comparing processing-and reflection-based measures of learning. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.),

Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 1856-1861). Austin, TX: Cognitive Science Society.

Isbilen, E.S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (submitted). Statistically-based chunking of non-adjacent dependencies.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E.J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 564–569). Austin, TX: Cognitive Science Society.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology*, *3*, 167. doi: 10.3389/fpsyg.2012.00167.

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition*, *53*, 79-88.

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: the case of nonword repetition. *Developmental Science*, *17*, 298-310.

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: a computational model of the learning of new words. *Developmental Science*, *10*, 853-873.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition, 144*, 1–13.

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory & Cognition*, *46*, 216-229.

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., & Jaeger, T. F. (2014). On-line measures of prediction in a self-paced statistical learning task. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society (pp. 725–730). Austin: TX.

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

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters*, *461*, 145-149.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 79–91.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, *20*, 6-14.

Lieven, E. (2016). Usage-based approaches to language development: Where do we go from here? *Language and Cognition*, *8*, 346-368.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187-219.

MacWhinney, B. (1998). Models of the emergence of language. Annual Review of Psychology 49, 199–227.

Majerus, S., van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological

language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language, 51*, 297-306.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1619–1624). Austin, TX: Cognitive Science Society.

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9*, 419–436.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126,* 1-51. https://doi.org/10.1037/rev0000126

McCauley, S. M., Isbilen, E. S., & Christiansen, M. H. (2017). Chunking ability shapes sentence processing at multiple levels of abstraction. G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2681–2686). Austin, TX: Cognitive Science Society.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302-331.

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

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604-607.

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

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233-238.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica, 149*, 1-8.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246-263.

Qi, Z., Sanchez, Y., Georgan, W., Gabrieli, J., Arciuli, J. (2019). Hearing Matters More Than Seeing: A Cross- Modality Study of Statistical Learning and Reading Ability. *Scientific Studies of Reading*, 23, 101-115

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

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906-914.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools, Inc.

Service, E. (1992). Phonology, working memory and foreign-language learning. *Quarterly Journal of Experimental Psychology*, *45*, 21-50.

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*, 155-172.

Slone, L., & Johnson, S. P. (2015). Statistical and chunking processes in adults' visual sequence learning. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2218–2223). Austin, TX: Cognitive Science Society.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92-102.

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

Siegelman, N., Bogaerts, L., Christiansen, M. H. & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B*, *372*, 20160059.

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

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, *42*, 692-727.

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

Simon, H. A. (1974). How big is a chunk? By combining data from several experiments, a basic human memory unit can be identified and measured. *Science*, *183*, 482-488.

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition*, *179*, 23-36.

Thiessen, E. D., & Pavlik, P. I. (2013). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science*, *37*, 310-343.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*, 61-82.

Tomasello, M. (2003). Introduction: Some surprises for psychologists. In M. Tomasello (Ed.), New psychology of language: Cognitive and functional approaches to language structure (pp. 1–14). Mahwah, NJ: Lawrence Erlbaum.

Torkildsen, J., Arciuli, J., Wie, O. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences*, 69, 60-68.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*, B25-B34.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*, 552-564.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 387-398.

Warker, J. A., Dell, G. S., Whalen, C. A., & Gereg, S. (2008). Limits on learning phonotactic constraints from recent production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1289-1295.

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *43*, 865-878.

Willet, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414-420.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pretest and posttest scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343–351.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in Social Science Methodology* (pp. 269–304). Greenwich: JAI Press.

Appendix 1

Table of SICR test items

Target items	Corresponding foil items
kibudulatibi	bikatolapoti
kibudutopoka	bukapodukito
latibilomari	dibumokidupa
latibitagalu	gaditamolupa
lomarikibudu	kalutotapoga
lomarimodipa	lobukimaduri
modipakibudu	moripadimalo
modipatopoka	popamokadito
tagalulomari	rilobimatila
tagalumodipa	tabigatilula
topokalatibi	tarimalugalo
topokatagalu	tidubibulaki

Table of SICR practice items

Target items	Corresponding foil items
kibuduloma	dumabuloki
latibitaga	kipobutoka
lomaritopo	matopalori
modipalati	patilamoda
tagalumodi	taludigamo
topokakibu	tatigabila

Tables and Figures

Table 1a

Summary statistics for Auditory 2AFC by Session

Session	Mean	SD	Range
1	.68	.13	.42 – .97
2	.76	.15	.33 – 1

Table 1b

Summary statistics for Auditory SICR by Session

Experimental items

	Full sequence			Trigram		
Session	Mean	SD	Range	Mean	SD	Range
1	37.61	12.91	13 – 66	7.31	5.02	0 – 20
2	41.91	14.90	11 – 71	9.19	5.81	1 – 23

Random items

	Full sequence			Trigram		
Session	Mean	SD	Range	Mean	SD	Range
1	28.38	9.96	12 – 58	2.95	3.07	0 – 14
2	27.50	10.73	6 – 57	2.83	3.03	0 – 15

Table 2a

Summary statistics for Visual 2AFC by Session

Session	Mean	SD	Range
1	.70	.16	.33 – 1
2	.78	.20	.42 – 1

Table 2b

Summary statistics for Visual SICR by Session

	Experimental items					
	Full sequence			Trigram		
Session	Mean	SD	Range	Mean	SD	Range
1	45.50	13.44	24 – 71	10.13	5.86	2 – 23
2	53.35	16.19	20 – 72	14.58	7.70	1 – 24

	Random items					
	Full sequence			Trigram		
Session	Mean	SD	Range	Mean	SD	Range
1	39.85	11.67	18 – 63	7.73	4.67	1 – 17
2	42.03	14.52	13 – 70	9.03	5.69	0 – 22

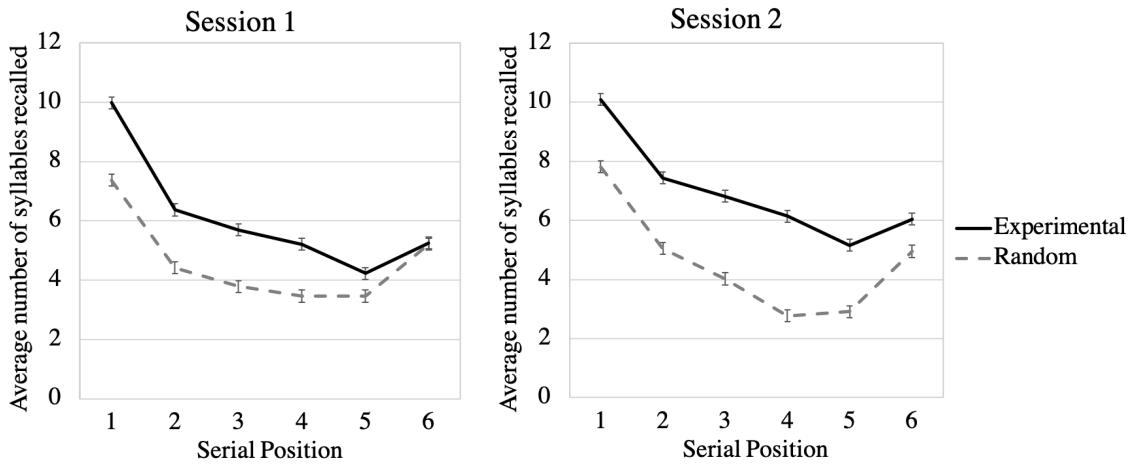

Fig. 1. Serial position curves for the SICR experimental and random items in auditory statistical learning. On average, participants recalled more syllables for the experimental items than the random items at every serial position in both sessions. This difference is especially pronounced in Session 2. Error bars depict standard error.

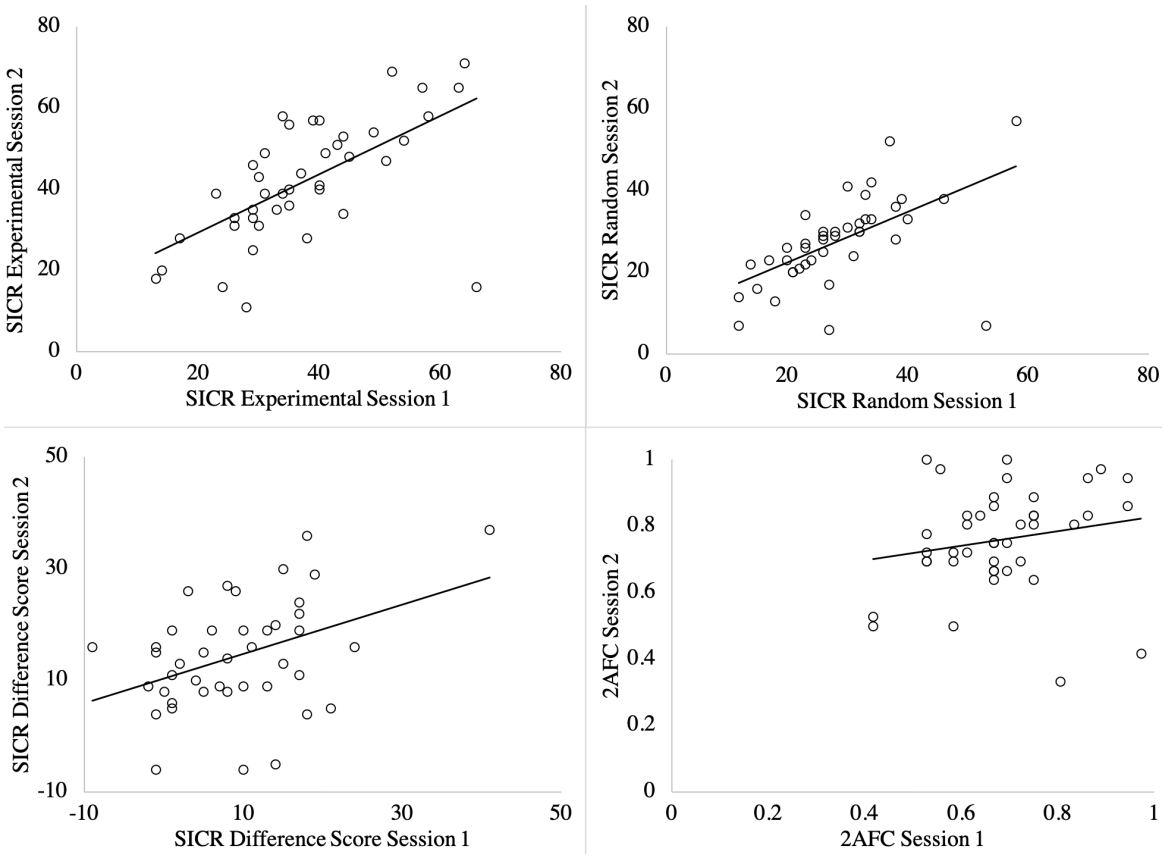

Fig. 2. The test-retest reliability of SICR and 2AFC in auditory statistical learning. SICR performance on the experimental items yielded the highest reliability, while 2AFC performance demonstrated the lowest test-retest reliability.

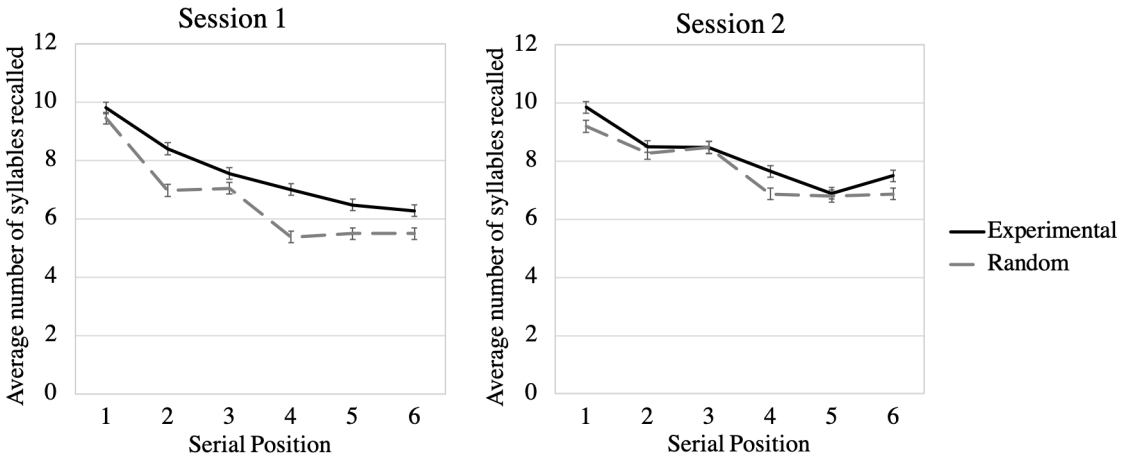

Fig. 3. SICR serial position curves by item type in visual statistical learning. In Session 1, participants on average recall a greater number of syllables in the experimental items than in the random items. However, this effect is less pronounced in Session 2. Error bars depict standard error.

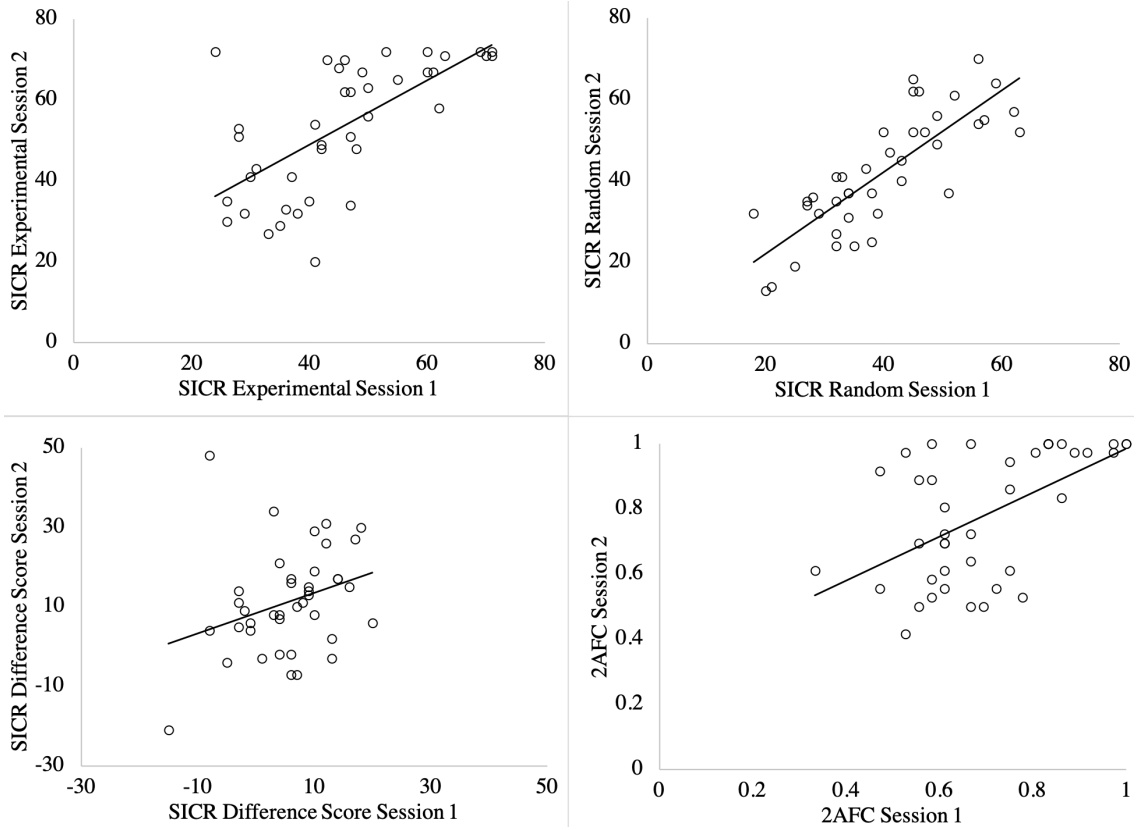

Fig. 4. The test-retest reliability of SICR and 2AFC in visual statistical learning. The reliability of SICR performance on the experimental items is higher than that of 2AFC.