Testing Statistical Learning Implicitly: A Novel Chunk-based Measure of Statistical Learning

Erin S. Isbilen (esi6@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA

Stewart M. McCauley (stewart.mccauley@liverpool.ac.uk)
University of Liverpool, Department of Psychological Sciences, Liverpool L69 7ZA UK

Evan Kidd (evan.kidd@anu.edu.au)
The Australian National University, Research School of Psychology, Canberra ACT 2601 AU

Morten H. Christiansen (christiansen@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA

Abstract

Attempts to connect individual differences in statistical learning with broader aspects of cognition have received considerable attention, but have yielded mixed results. A possible explanation is that statistical learning is typically tested using the two-alternative forced choice (2AFC) task. As a meta-cognitive task relying on explicit familiarity judgments, 2AFC may not accurately capture implicitly formed statistical computations. In this paper, we adapt the classic serial-recall memory paradigm to *implicitly* test statistical learning in a statistically-induced chunking recall (SICR) task. We hypothesized that artificial language exposure would lead subjects to chunk recurring statistical patterns, facilitating recall of words from the input. Experiment 1 demonstrates that SICR offers more fine-grained insights into individual differences in statistical learning than 2AFC. Experiment 2 shows that SICR has higher test-retest reliability than that reported for 2AFC. Thus, SICR offers a more sensitive measure of individual differences, suggesting that basic chunking abilities may explain statistical learning.

Keywords: statistical learning; chunking; language; language acquisition; implicit learning; learning; memory, serial recall; individual differences

Introduction

Statistical learning is understood as the process by which individuals implicitly track the distributional regularities in an input, leveraging recurring statistical patterns to facilitate cognitive processing (see Frost, Armstrong, Siegelman & Christiansen, 2015, for a review). In recent years, validating the theoretical link between the behavior observed in lab-based studies of statistical learning and broader aspects of cognition—such as working memory, language processing, and social learning—has garnered extensive interest. However, Romberg and Saffran (2010) noted that although typical tests of statistical learning demonstrate that individuals appear sensitive to statistical structure, such evidence on its own provides little insight into the *process* of learning, and the nature of the representations that consequently arise. The lack of a mechanistic understanding of statistical learning was further suggested to complicate attempts to tie this ability to other aspects of cognition, such as language acquisition.

Indeed, endeavors to relate individual variation in statistical learning to other facets of cognitive processing have yielded mixed results. For example, whereas some findings report that statistical learning abilities significantly correlate with verbal working memory and language comprehension (Misyak & Christiansen, 2012), others find no reliable relationship with language skills (Siegelman & Frost, 2015). These conflicting reports could suggest either that statistical learning is not meaningfully related to other aspects of cognition, or alternatively, that the measures used to assess statistical learning may not capture its full extent nor the scope of individual variation in this behavior.

In many studies, statistical learning is typically tested using a two-alternative forced-choice task (2AFC), in which learners are presented with pairs of stimuli and are asked to identify which of the two items were present during familiarization. As such, a possible limitation of the 2AFC task is that it is inherently meta-cognitive in nature, requiring the participant to make an explicit response (a button press) based on a “gut feeling” about implicitly acquired statistical regularities. Thus, as suggested by Franco, Eberlen, Destrebecqz, Cleeremans and Bertels. (2015), 2AFC may therefore more accurately reflect explicit decision-making processes rather than the actual underlying statistical learning mechanisms. Relatedly, although the 2AFC task is assumed to serve as an accurate proxy for the learning of statistical structure, the strategy for successful performance on this task may differ from that required for successfully detecting statistical regularities in the input stream (Siegelman, Bogaerts, Christiansen & Frost, 2017). Lastly, even though 2AFC may yield useful mean estimates of performance at the group level, the additional cognitive complexity associated with 2AFC performance is likely to introduce error variance such that individual scores may not optimally reflect individual differences in statistical learning ability (Siegelman & Frost, 2015).

Because of these limitations, a unified theoretical framework that situates statistical learning within broader cognitive processing has thus far remained out of reach. In the current paper, we propose a new measure that implicitly tests statistical learning. Our novel task aims to offer more direct insights into what is being learned in statistical learning-based experiments, while at the same time aligning such learning with the wider learning and memory literature.

Recent theoretical considerations suggest that basic abilities for chunking may subserve many aspects of learning and memory, particularly within the domain of language processing (Christiansen & Chater, 2016). Our perspective builds on classic memory studies demonstrating that the number of items that can be held in memory significantly increases when successfully chunked into larger units (Miller, 1956; Cowan, 2001). This underscores the potential contribution of chunking processes to the successful learning and retention of new material. For example, when tasked with remembering the novel sequence of letters *ailcpaphrtleca,* preserving the letters in memory poses a considerably greater challenge than successfully recalling the same set of letters chunked into larger coherent units, such as in the sequence *catapplechair*. Due to our extensive experience with language, the same set of letters can be more easily retained by exploiting our ability to chunk them into words (i.e. "cat", "apple", and "chair"), which in turn can subsequently be deconstructed to retrieve the individual letters. Our novel task takes advantage of similar chunking processes.

Here, we leverage the general capacity for chunking in a statistically-induced chunking recall task (SICR) as a novel implicit measure of statistical learning. We refashion a central tool in the chunking and memory literature—serial recall (e.g., Miller, 1956)—for use in statistical learning-based tasks. Subjects are exposed to six trisyllabic nonsense words using the classic Saffran, Newport and Aslin (1996) paradigm. After training, participants are aurally presented with syllables from the input and asked to recall them out loud. Critically, the experimental items in our task consisted of the concatenation of two words from the input language (Word A + Word B), and control items consisted of the exact same six syllables in a random configuration, like in the example above. Our hypothesis is that if subjects have statistically chunked the syllables in the input stream into words, then recalling a string consisting of two words should yield more accurate recall of the presented syllables than recalling the same set of syllables in a random order. Crucially, our task is scored on a syllable-by-syllable basis rather than assigning a binary 0 or 1 score as in the 2AFC task, enabling the calculation of subjects' sensitivity to trigrams and serial position. This yields a richer set of performance data than the 2AFC task, thus providing a more detailed picture of each subject's individual sensitivity to different kinds of information in the input.

In the current paper, we conducted two experiments to determine the efficacy of SICR in capturing statistical learning behavior, and the formation of the word-level representations from accrued statistics. In Experiment 1, we compare 2AFC performance to SICR, showing that the latter provides a useful, memory-based measure of implicit statistical learning. To be able to relate statistical learning to specific aspects of language and cognition through individual differences studies requires a performance measure that is stable across time. Because recent research has cast doubts on the reliability of the 2AFC task in the context of the classic Saffran-style paradigm (Siegelman, Bogaerts & Frost, 2016), we conducted a test-retest study of our SICR task in Experiment 2. We conclude with a discussion of the methodological and theoretical implications of SICR, and how future use of this task may help in establishing a definitive relationship between statistical learning and cognition more broadly.

Experiment 1: Comparing statistically-induced chunking recall (SICR) with 2AFC

Experiment 1 investigated whether chunking might account for the word-level representations gleaned in statistical learning experiments using the classic Saffran et al. (1996) paradigm. In addition to these theoretical considerations, we also sought to assess the methodological efficacy and sensitivity of both the established 2AFC task, and our novel SICR task in assessing statistical learning. Through exposure to the input, we predict that syllables that regularly co-occur in the input will be chunked into words, which should yield higher recall accuracy of the chunked words than the same syllables heard in a random order.

Method

Participants 69 native English-speaking undergraduates from Cornell University (34 females; age: *M*=19.78, *SD*=1.62) participated for course credit.

Materials The input language consisted of 18 syllables (*bi, bu, di, du, ga, ka, ki, la, lo, lu, ma, mo, pa, po, ri, ta, ti, to*), combined into six trisyllabic words: *kibudu, latibi, lomari, modipa, tagalu, topoka*. Seventy-two randomized blocks of the six words were concatenated into a continuous speech stream using the MBROLA speech synthesizing software (Dutoit et al., 1996). Each syllable was approximately 200 milliseconds long, separated by 75 milliseconds of silence.

For the 2AFC task, six additional foil words were pseudo-randomly generated, avoiding the reuse of transitional probabilities from the target words above: *dikabi, kigala, lopadu, mamoti, polubu, tatori*.

The stimuli for the SICR task consisted of 24 six-syllable items. The twelve experimental items were composed of two adjacent words from the input (e.g., *kibudulatibi*), and the twelve corresponding foil items consisted of the same set of syllables in pseudorandom order (e.g., *kibudulatibi* → *tidubibulaki*), avoiding preexisting transitional probabilities from all other syllable combinations in the experiment. Additionally, 12 5-syllable practice items were included, which were constructed in the same manner as the 24 items reported above, but using one full word and the first bigram of a second word.

Procedure The experiment consisted of three distinct tasks. First, subjects were familiarized with the artificial language. To ensure active engagement, a cover task based on Arciuli & Simpson (2012) was administered. In addition to each of the six words in the experiment, three variants of each word containing a syllable repetition was included in the training stream (e.g., *tagalu* → *tatagalu, tagagalu, tagalulu*). Participants were instructed to click the space bar when they noticed a repeated syllable. Each of the three variants of the words appeared 4 times, yielding 72 repetitions. In total, training lasted 11 minutes.

After training, participants' knowledge of the artificial language was tested using both the standard 2AFC task, and our SICR paradigm. The order of these two tasks was counterbalanced such that half of the subjects were given 2AFC first, and half were given SICR first. In the 2AFC task, each of the 6 target words were aurally presented with one of the 6 2AFC foil words, and subjects were asked to report which of the two trigrams had been present during training. There were 36 2AFC trials in all, in which each target word appeared alongside each foil once.

In the SICR paradigm, 12 five-syllable practice trials were administered prior to the 24 six-syllable items to familiarize subjects with the task, and to ensure that the amount of post-test exposure to the words would be the same regardless of whether subjects did 2AFC first, or SICR first. In this task, participants were told that we would be gauging their ability to recall the syllables from the experiment. Each item was aurally presented, after which subjects were prompted to recite back each syllable in the sequence to the best of their ability. Importantly, at no point in the experiment were subjects informed that they were partaking in a language experiment, nor was their attention directed to the presence of structure.

Results and Discussion

The mean accuracy of correctly choosing the word over the foil in the 2AFC task was 66% (*M*=.66, *SD*=0.13), which is significantly greater than chance, *t*(68)=11.11, *p*<.001. These results are comparable with other studies that utilize 2AFC to assess statistical learning, which typically report performance within the range of 60% (Frost et al., 2015).

Scoring for the SICR task was done on a syllable-by-syllable basis, enabling analysis of both the overall strings, and the individual words composing the strings. When comparing the number of syllables accurately recalled for the experimental items (*M*=42.7, *SD*=10.68) to the number of syllables recalled for random items (*M*=31.19, *SD*=10.29), participants accurately recalled significantly more syllables for the experimental items than the random items, *t*(68)=13.85, *p*<.0001. A similar pattern was observed for trigram performance: participants accurately recalled significantly more of the experimental trigrams (*M*=8.68, *SD*=4.25) than items consisting of random trigrams (*M*=3.58, *SD*=3.02), *t*(68)=13.72, *p*<.0001 (Figure 1a). Additionally, the serial position curves for the experimental and random items can be found in Figure 1b. These results confirm our hypothesis that through exposure to the distributional regularities in the input, individuals appear to have successfully chunked co-occurring syllables into larger units, and the formation of these word-level representations of the input leads to markedly better memory for experimental items.

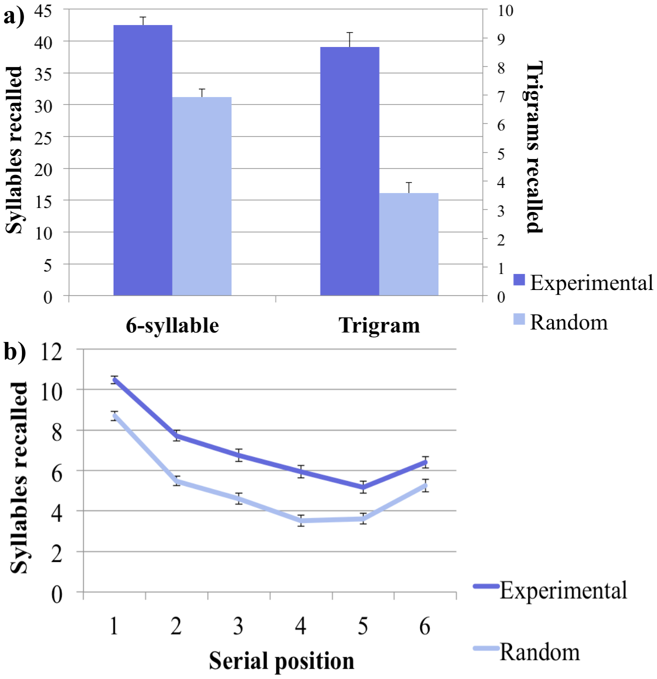

Figure 1: a) Average SICR performance. Participants recall significantly more syllables when the test items consist of two concatenated input words, and significantly more trigrams within the experimental six-syllable items. b) Serial position curves for experimental and random items.

Interestingly, our analyses revealed no significant correlations between 2AFC and any of our SICR measures (*r*(67)=0.21, *p*=.084 for experimental items, and *r*(67)=0.18, *p*=.4 for experimental trigrams. For the score distributions of the two tasks, see Figure 2). However, this finding mirrors recent results by Franco et al. (2015), who also found no correlation between 2AFC accuracy and their Rapid Serial Auditory Presentation task (RSAP), a detection task intended to serve as a more implicit measure of auditory statistical learning. Similar to SICR, RSAP works by exposing subjects to an artificial speech stream composed of trisyllabic words, after which subjects were tasked with detecting a target syllable embedded within strings of target words from the training corpus. Unlike explicit measures like 2AFC, RSAP and SICR are implicit measures in which no reference is made to a desired discrimination, and thus may be more sensitive to the acquired statistical regularities, including information about

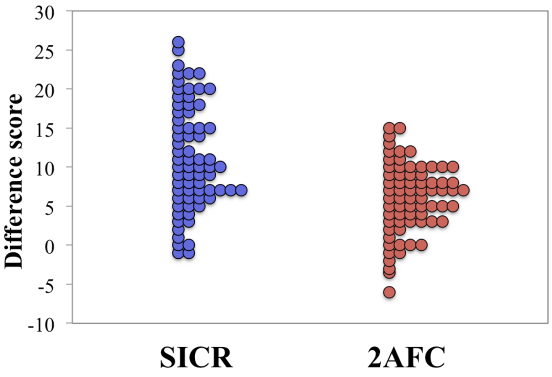

Figure 2: The distributions of SICR (experimental-random items), 2AFC scores as compared to chance, and syllable recall for experimental items.

which the participant lacks awareness. Thus, 2AFC and SICR may be picking up on different aspects of statistical learning – decision-making processes based on learned information and underlying mechanisms, respectively – which may contribute to the low correlation between the two measures.

Notably, our analyses revealed a strong order effect for 2AFC performance: individuals who performed SICR prior to 2AFC exhibited significantly higher 2AFC scores, $t(68) = 12.06$, $p<.0001$. Compared to the means of those who completed 2AFC first, a 7%-point increase in 2AFC performance was observed for participants who did SICR first. This may account for why our participants on average performed higher on 2AFC than the 60% typically reported for this type of statistical learning. By contrast, SICR was unaffected by the order in which it was performed ($t(68)=0.22$, $p=.59$ for experimental items, $t(68)=-0.22$, $p=.42$ for experimental trigrams). The robustness of SICR is notable given that in both conditions, the amount of post-input exposure was kept the same, ruling out exposure differences as an explanation for the order effects. That is, despite both tasks being granted the same opportunity for post-input learning, only 2AFC was affected by the additional exposure.

Taken together, several conclusions can be made from the results of Experiment 1. Firstly, our findings support the idea that chunking may serve as the mechanism by which exposure to statistical regularities lead to representational changes in memory. Secondly, our results affirm that SICR can serve as a valid means of testing the acquisition of sequential regularities, with the additional benefit of offering more fine-grained insight into the acquired representations. Finally, the lack of correlation between 2AFC and SICR may represent fundamental differences between explicit versus implicit measures of learning (Franco et al., 2015). Thirdly, the lack of order effects on SICR performance suggests that it may be a more stable measure of statistical learning ability than 2AFC. To further examine the stability of SICR across time, we assessed its test-retest reliability in Experiment 2.

Experiment 2: Establishing the test-retest reliability of SICR

To date, varying levels of test-retest reliability for different measures of statistical learning have been found. For instance, using 2AFC as the primary measure, Siegelman and Frost (2015) reported adequate test-retest reliability for auditory verbal adjacent (r=0.63), and visual nonverbal adjacent statistical learning (r=0.58), and relatively low reliability for auditory nonverbal adjacent (r=0.23) and auditory verbal non-adjacent statistical learning (r=0.31). The implications of this are twofold: a) that certain types of statistical learning capacities are not stable within individuals and/or b) that certain tasks may lack specificity as to the behavior they aim to capture (Siegelman et al., 2017). Thus, the goals of Experiment 2 were to determine whether SICR provides a reliable measure of individual statistical learning capabilities, and to establish whether the associated hypothesis—that chunking abilities can account for statistical word learning—would replicate.

Method

The same general method from Experiment 1 was employed, with a few notable exceptions. Subjects were exposed to the same input language, after which SICR was administered to measure word learning. Unlike the previous study, 2AFC was not included in Experiment 2, given existing studies assessing its test-retest reliability. Following the completion of Session 1, participants returned three weeks later and completed the same tasks again in Session 2, mirroring the timespan between test and retest in Siegelman and Frost (2015).

Participants 26 native English-speaking undergraduates from Cornell University (15 females; age: M=19.31, SD=1.32) participated for course credit.

Materials The same input language from Experiment 1 was used. The SICR stimuli consisted of the same 24 six-syllable items from Experiment 1, half composed of two concatenated words from the input, and the other half their complementary randomized foils.

Procedure The experiment consisted of two tasks. First, subjects were familiarized with the input language, including the same cover task as before. In total, training lasted 11 minutes. The SICR task was identical to Experiment 1, with the exception that participants were given a different randomized input and SICR item order in each session.

Results and Discussion

As in Experiment 1, participants performed significantly better on the experimental items than on the random items, both in Session 1, $t(25)$= 5.46, p<.0001, and in Session 2, t(25)=7.08, p<.0001. The same results were found for

Table 1: Means and standard deviations of SICR scores

	Session 1		Session 2	
	M	SD	M	SD
6-syllable experimental	36.42	12.48	40.15	12.73
6-syllable random	27.04	10.71	28.0	10.38
Trigrams experimental	6.89	4.41	8.31	4.46
Trigrams random	3.0	2.65	2.96	2.60

performance on the trigrams, with participants recalling significantly more experimental trigrams in both Session 1, *t*(25) =6.18, *p*<.0001, and in Session 2, *t*(25)=7.67, *p*<.0001. The mean performance on these measures can be found in Table 1. Thus, the results from both sessions replicated the results from Experiment 1.

Between the two sessions, the test-retest reliability of SICR proved to be very strong. SICR performance was highly correlated across the two sessions. Performance on the recall of six-syllable experimental items was *r*(24)=0.81, *p*<.0001 (Figure 2). This exceeds the correlation coefficient of 0.63 reported for 2AFC in an auditory statistical learning task by Siegelman and Frost (2015). Recall performance on the six-syllable random items was also highly stable, *r*(24)=0.85, *p*<.0001. Performance on experimental trigrams *r*(24)=0.73, *p*<.0001 and random trigrams *r*(24)=0.82, *p*<.0001 was also consistent across the two sessions. However, the correlations of the differences scores (performance on experimental minus random items) were slightly lower, yielding *r*(24)=0.46 *p*=.0192 for six-syllable recall, and *r*(24)=0.53 *p*=.0053 for trigram recall. These results suggest that performance on the experimental items may be a better measure of individual differences in statistical learning than the difference scores.

In all, the results of Experiment 2 corroborate our findings from Experiment 1, in which experimental items yield significantly better recall. Our results also confirm the stability of SICR. Taken together, these findings suggest that SICR proves to be both a theoretically valid and methodologically sound measure of statistical learning.

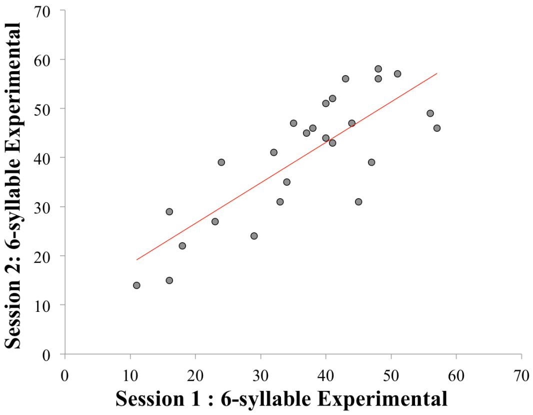

Figure 3: Correlation between Sessions 1 and 2 recall scores for statistically experimental items.

General discussion

In this paper, we introduced a novel chunk-based method to implicitly test statistical learning—the SICR task—as an alternative to the standard 2AFC task. The results of our experiments demonstrate that through exposure, subjects' implicit chunking of the distributional regularities in the input significantly amplified their baseline working memory abilities (as captured by performance on the random items), and that the formation of multi-syllabic chunked representations of the input markedly boosted recall. Furthermore, these results appear to be strikingly stable over time and are less subject to order effects than 2AFC, which underscores the promise of SICR as a reliable and multifaceted measure of statistical learning faculties.

SICR offers several methodological benefits that circumvent a variety of issues inherent to 2AFC. Because 2AFC relies on overt decision-making processes about the familiarity of stimuli, it is unclear as to whether 2AFC may thus only be reflective of the more explicit meta-cognitive aspects of statistical learning. 2AFC appears to provide more limited sensitivity to individual differences, as it tends to rely on a binary all-or-nothing score. This lack of granularity in the scoring also makes it more difficult to accurately assess the precise extent of learning.

One important difference between explicit tasks like 2AFC and implicit tasks such as SICR is that they may be respectively characterized as 'direct' versus 'indirect' measures of learning (Franco et al, 2015). Whereas direct measures steer participants' attention toward the relevant discriminations they are expected to make, indirect measures that circumvent the need for explicit instruction may be more sensitive to any knowledge the subject has acquired, including material below the threshold of conscious awareness. That is, although direct and indirect measures should exhibit equal sensitivity to consciously known information, direct measures may not be as adept at capturing the accretion of information of which the learner is not yet fully aware. Furthermore, unlike 2AFC and reaction time tasks, SICR requires both immediate comprehension and production on the part of the learner. The task thus provides the means to capture how exposure to statistical regularities can facilitate memory abilities via improved chunking abilities, which in turn may help the learner to overcome the processing pressures deriving from the Now-or-Never bottleneck (Christiansen & Chater, 2016). As such, SICR may be seen as an ecological measure of the impact of accrued statistics on the online memory processes used to track verbal input, without the need for participants to rely on explicit decision-making.

Whereas 2AFC relies on a binary scoring method, SICR offers a more granular approach by performing scoring on a syllable-by-syllable basis, allowing the evaluation of sensitivity to trigrams and serial position. The richness of this dataset may also lend itself to acoustic measurements of production durations and analysis of prosody. Because of the sensitivity of SICR to a number of different individual capacities, and findings suggesting that chunking ability serves as a strong predictor of online language processing skills (McCauley & Christiansen, 2015), SICR may also be employed compare how individual differences in statistical learning may predict other language learning abilities. Indeed, preliminary results from an ongoing study with 5-6-year-old children (*N*=73) indicate that performance on the experimental items in the SICR task correlates significantly with language skill (*r*=0.41, *p*<.001), whereas 2AFC performance does not (*r*=0.20, *p*=.096).

More generally, the basic recall methodology upon which SICR piggy-backs has a long pedigree in the domain-general memory literature, including serial recall (e.g., Miller, 1956). Of particular importance is the related work on nonword repetition, which has been established as one of the primary predictors of language ability (e.g., Gathercole et al., 1994). Our SICR measure may be seen as a statistical learning-based variation on a nonword repetition task, in which we manipulate the distributional support for the items to be recalled via artificial language exposure. This interpretation of the SICR task dovetails with evidence that nonwords constructed from phoneme sequences that occur frequently in natural language are repeated more accurately than nonwords based on infrequent phoneme strings (Majerus, van der Linden, Mulder & Peters, 2004). In a similar vein, recall of random digit sequences has also been shown to reflect natural language statistics (Jones & Macken, 2015).

In addition to the methodological advantages afforded by this novel method, SICR also points toward a theoretical answer to Romberg and Saffran’s (2010) concern about the lack of connection between measures of statistical learning and potential underlying processes and representation. Our proposition, given the efficacy of SICR in capturing statistical learning behavior, is that chunking may be seen as the process by which encountered statistics are used to form concrete, discrete units, thereby effectively segmenting a continuous stream into individual words. As such, the output of statistical learning may thus be seen as individual chunks of varying sizes. This notion is corroborated by previous research suggesting that chunking-based processes enable the recoding of incoming information into gradually higher levels of abstraction, from acoustic input, to words, to multiword units and beyond (Christiansen & Chater, 2016). Thus, SICR provides both a compelling tool to effectively and ecologically appraise statistical learning, and strives to bridge the statistical learning and chunking memory literatures.

Acknowledgments

We wish to thank Dante Dahabreh, Jess Flynn, Jake Kolenda, Jeanne Powell and Sam Reig for their assistance with collecting and coding data. This work was supported in part by NSF GRFP (#DGE-1650441) awarded to ESI.

References

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

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

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

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

Frost, R., Armstrong, B.C., Siegelman, N. & Christiansen, M.H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19*, 117-125.

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*(2), 103-127.

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

McCauley, S.M. & Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. 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.

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.

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

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606-621.

Siegelman, N., Bogaerts, L., Christiansen, M. H. & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Phil. Trans. R. Soc. B*, *372*(1711), 20160059.

Siegelman, N., Bogaerts, L. & Frost, R. (2016). The peculiar tale of ASL: What do we measure when we use the auditory statistical learning task? Talk presented at the conference on First vs. Second Language Learning: From Neurobiology to Cognition conference, Hebrew University, Israel.

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