UNIVERSITY of CALIFORNIA PRESS | Collabra

Erickson, L C et al 2016 Individual Differences in Statistical Learning: Conceptual and Measurement Issues. *Collabra,* 2(1): 14, pp. 1–17, DOI: http://dx.doi.org/10.1525/collabra.41

ORIGINAL RESEARCH REPORT

Individual Differences in Statistical Learning: Conceptual and Measurement Issues

Lucy C. Erickson*,‡, Michael P. Kaschak†, Erik D. Thiessen‡ and Cassie A. S. Berry†

The ability to adapt to statistical structure (often referred to as "statistical learning") has been proposed to play a major role in the acquisition and use of natural languages. Several recent studies have explored the relationship between individual differences in statistical learning and language outcomes. These studies have produced mixed results, with some studies finding a significant relationship between statistical learning and language outcomes, and others finding weak or null results. Furthermore, the few studies that have used multiple measures of statistical learning have reported that they are not correlated (e.g., [1]). The current study assesses the reliability of various measures of auditory statistical segmentation, and their consistency over time. That is, do the generally low correlations observed between measures of statistical learning stem from task demands, the psychometric properties of the measures, or the fact that statistical learning may be a highly fragmented construct? Our results confirm previous reports that individual measures of statistical learning tend not to correlate with each other, and suggest that the somewhat weak reliability of the measures may be an important factor in the low correlations. Our data also suggest that aggregating performance across tasks may be an avenue for improving the reliability of the measures.

Keywords: statistical learning; implicit learning; individual differences; reliability

Humans show a remarkable ability to adapt to the probabilistic structure of their environment (e.g., [2]). This adaptability plays a role in many domains of formal and informal learning, including language (e.g., [3]), mathematics [4], locomotion (e.g., [5]), and perception (e.g., [6]). Recent linguistic theories suggest that adaptation to probabilistic information characterizes both language acquisition in infancy [7, 8, 9], and language comprehension and production in adulthood (e.g., [10, 11, 12]). Within the domain of language, the acquisition of probabilistic structure has often been termed *statistical learning* (SL)[1]. SL has been demonstrated in learning and use of phonemic categories [13], lexical forms [14, 15], phonotactic and phonological regularities characterizing words [16, 17, 18], word meaning [19], and syntactic structure [20, 21, 22, 23, 24].

Statistical information plays a key role in recent theories of language use. From this perspective, a straightforward prediction is that individual differences in SL ability should be related to variation in language outcomes, with the specific expectation that individuals who are better at SL should show better language learning and processing (e.g., [1, 9, 25, 26]). Although this prediction flows naturally from a number of extant psycholinguistic theories (e.g., [10, 12]), it has received comparatively little attention in the literature. This state of affairs may be partly due to the fact that until recently (e.g., [27]), there has been some doubt as to whether substantive individual variation in SL ability (or, more generally, implicit learning of probabilistic structure) actually existed (e.g., [2]). A handful of studies on individual differences in SL ability have been reported over the past several years (e.g., [27, 28, 29]), with some studies reporting significant correlations between SL and language outcomes (e.g., [30]) and others reporting null or weak correlations (e.g., [31]).

One complicating factor in interpreting these conflicting findings is that SL ability has been studied with a wide array of tasks. In some of these tasks, the statistical information to be learned is about the likelihood with which events predict each other (e.g., [3]), while in other tasks the statistical information relates to the distribution (i.e., frequency and variability) of exemplars (e.g., [13]). Some of these tasks present events sequentially such that participants can learn transitional probabilities (TPs; e.g., [32]), and others present events simultaneously such that participants can learn co-occurrence probabilities (e.g., [33]). SL tasks differ in the modality in which probabilistic

* University of Maryland, College Park, US
† Florida State University, US
‡ Carnegie Mellon University, US
Corresponding author: Lucy C. Erickson (lce@umd.edu)

information is presented, including audio, visual, and tactile inputs and their combination (e.g., [34, 35, 36]). They differ in the response they ask from the participants, ranging from explicit judgments of probability or familiarity as in two-alternative forced-choice tasks (2AFC; e.g., [37]) to implicit behavioral responses in which participants show decreases in both errors and reaction times in pressing a series of keys as they adapt to the statistical structure of the input, as in Serial Response Time (SRT) tasks (e.g., [27]). Finally, although one of the hallmarks of traditional SL tasks is the absence of explicit feedback, a consequence of the response format used in the SRT task is that the motor response may provide the participant with an incidental source of feedback that may influence the learning systems that are engaged. As evidence of this, Lim, Fiez, and Holt [38] report that the basal ganglia are recruited by incidental feedback of the sort that is often found in these tasks.

Although successful learning in each of these tasks has been construed as evidence of SL, the tasks themselves vary so widely as to raise the question about whether they all tap into the same underlying process, or if the term "statistical learning" might be an umbrella term describing a set of independent processes (see [39]). This uncertainty can be seen at both a theoretical and a methodological level. On a theoretical level, different proposals carve up SL in different ways: some suggest that it is a unitary process (e.g.,[40]), whereas others suggest it is an umbrella term for multiple processes, but differ on the number and nature of these underlying processes (e.g., [9, 34]). On a methodological level, different measures of SL are quite different, and not necessarily correlated (e.g., [1, 41, 42]).

These theoretical and methodological uncertainties are deeply interrelated. When two different measures of SL are uncorrelated or differ in some obvious way, this may reflect the fact that they tap into different underlying processes, the fact that the measures themselves have low reliability, or the fact that the measures have very distinct task demands. Consider the lack of correlation evidenced by the two SL measures used by Misyak and Christiansen [1]. The two measures indexed adjacent and nonadjacent regularities, respectively, and range of performance on the nonadjacent SL task was truncated by a ceiling effect. Consequently, insufficient evidence exists to attribute the lack of correlation to psychometric properties, task demands, or to potentially different processes underlying SL as a function of adjacency. Similarly, consider demonstrations that visual SL shows different learning constraints than audio SL: the optimal rate of presentation across these modalities is quite distinct, such that the visual modality requires a slower rate of presentation than the auditory modality for best learning [43, 44]. This may be due to the fact that there are different processes underlying audio and visual SL. Alternatively, it may be because the same underlying process shows different patterns of learning as a function of the input's perceptual characteristics. For example, sequential information may be easier to perceive for audio input than simultaneous information, and the reverse true for visual input. These differences in perceptual ease may, in turn, lead to different patterns of learning across audio and video input.

In addition to modality, researchers have also proposed that different mechanisms are involved in the processing of different kinds of statistical information (e.g., frequency vs. transitional probabilities; [45]). Neuroimaging studies of SL have reported that multiple neural systems (e.g., fronto-striatal circuits; hippocampal regions) are activated during SL tasks (e.g., [46, 47]). Amso et al. [45] reported that the caudate was responsible for learning based on simple frequency, whereas the hippocampus was involved in associative learning. Turk-Browne et al. [47] suggest that the hippocampus may be implicated in learning that results in relatively more abstract representations, whereas the caudate is involved in the learning of more specific representations. Because most SL tasks involve multiple kinds of regularities, the possibility that different neural processes are involved in the learning of different kinds of regularities may also contribute to the challenge of understanding the nature of individual differences on these tasks.

The controversies we have described present obstacles to understanding how individual differences in SL relate to language outcomes (as well as other measures of cognitive performance; [42]). Overcoming these obstacles will require a detailed look at the demands that each SL task presents to the learner as well as an examination of the psychometric properties of the individual measures. One such example of this approach is the work reported in Siegelman and Frost [42]. Using five different SL tasks (verbal and nonverbal auditory tasks with adjacent and non-adjacent dependencies, a visual task with adjacent dependencies, and an SRT), Siegelman and Frost [42] show that the SL measures do not correlate highly with each other, and do not correlate highly with other measures of cognitive ability. They also demonstrate that the test-retest reliability of the SL measures is quite variable, ranging from just below 0.70 to around 0.20. Siegelman and Frost's [42] results are important both because they show that SL tasks tend not to correlate with each other using a wider range of tasks than is typically seen in studies of SL, and because they hint at the possibility that at least some of the reason why SL tasks may not correlate with each other (or, with other cognitive measures) is that the reliability of most of the SL tasks examined was not optimal.

Siegelman and Frost's [42] study represents an important first step in clarifying the issues involved both in measuring SL, and in using these measures to study the structure of SL and its relationship to other cognitive abilities. However, there are important limitations to what can be concluded from their research design. The SL tasks chosen varied on a number of dimensions – whether the tasks were auditory or visual, whether they had adjacent or non-adjacent dependencies, whether they used verbal or non-verbal stimuli, as well as on their observed reliability—and as such there remains a degree of uncertainty about the cause of the low correlations between tasks. Therefore, an important question that remains unaddressed is whether SL measures do not correlate with each other because SL is a much more fragmented construct than originally

thought, or because there is something intrinsic to the tasks themselves (e.g., task demands or psychometric properties) that limit the ability to find correlations across measures. The research reported here was intended to take a step toward addressing this issue.

Our study used a design similar to that reported by Siegelman and Frost [42]—we assessed SL performance on a range of tasks, and assessed test-retest reliability of our tasks—but our selection of tasks was less varied. Rather than assess performance on a set of tasks that differ on multiple dimensions (such that each task has a unique set of demands), we assessed performance on a set of very similar tasks. Specifically, we assessed performance on a set of auditory word segmentation tasks[2] such as the one introduced to the literature by Saffran, Aslin and Newport [3]. By using a set of tasks with essentially identical demands, we hoped to create an ideal case for observing high correlations between SL measures. Even if SL is a fragmentary construct, it is expected that performance on tasks with very similar demands should correlate with each other due to the fact that they are assessing the same "fragment" of the construct. If we failed to see correlations between tasks under such conditions, it would support the idea that there may be characteristics of the SL tasks themselves that do not lend themselves to good measurement of the construct. To preview our results, our findings are potentially consistent with the possibility that 1) SL is a construct that is so fragmented that it may be of limited theoretical use, or 2) null correlations between SL measures results from measurement rather than theoretical issues. We return to the discussion of these issues in the General Discussion.

Experiment 1

In Experiment 1, we familiarized and tested learners on four distinct artificial languages that can be segmented based on statistical structure (e.g., [48]). We focus on word segmentation tasks for several reasons: there are several different "languages" available for use (so that participants can be trained on multiple stimulus sets, each with a different set of training items), these "languages" and tasks are all quite similar in nature (short training sets, followed by an assessment), and these tasks all tapped into the same dimension of SL (which we describe as "chunking" or "extraction"; [9]), yet are distinct enough to minimize potential confusion or interference. Because these languages have been used previously, their selection allows for continuity with and comparison to prior research. Four languages were selected as a balance between allowing for comparison across multiple languages without causing participants undue fatigue. Finally, the nature of the word segmentation task allows us to assess learning in different ways. In this study, we asked participants to choose between words and foil items using a two alternative forced choice (2AFC) task, and also using Likert scale ratings (RS) of the goodness or familiarity of word and foil items. Assessing learning in two ways provides another opportunity to determine whether different response types are more or less sensitive to individual differences, and whether the different learning outcome measures used in prior research may also explain low correlations among SL tasks.

After listening to each word segmentation language, participants were given both a forced-choice assessment and a rating-scale assessment to measure their acquisition of the words in the segmentation language. This study used a test–retest design in which participants completed the four learning tasks twice, in two sessions separated by a week. This design afforded us the opportunity to address several issues. The similarity of the tasks that were used provides a test of the hypothesis that a lack of task similarity led to the low correlations between SL measures in previous studies. Because the four tasks in this experiment are very similar, low inter-correlations between them would suggest that low correlations between SL tasks found in prior experiments cannot be entirely explained by the dissimilarity in tasks used in those experiments. Testing participants on the same measures on two separate days allowed us to assess the reliability of the measures. If it were poor (as found for some of the tasks reported in Siegelman and Frost [42]), it would suggest an explanation for the lack of correlation between SL measures. Finally, the use of forced-choice and rating-scale assessments provides the opportunity to observe whether the nature of the assessment affects the relationships observed between measures, and an opportunity to determine whether these two kinds of responses differ in their sensitivity to individual variability in SL.

In addition to examining the reliability of the measures over time, using two testing sessions separated by a week allowed us to investigate an important question that has been largely unexplored. If SL is to play a meaningful role in the acquisition of natural language, the representations formed during SL must be long-lasting. To date, few studies have investigated duration of learning effects in the context of the statistical segmentation paradigm [42, 49, 50]. Two of the studies [49, 50] tested whether the representations formed during visual statistical segmentation last over a 24 hour delay, whereas Siegelman and Frost [42] had a longer test-retest delay. The results of these studies are equivocal, with long-term learning being demonstrated for some measures, but not others (e.g., [42]). The present study, by virtue of testing participants learning on two sessions separated by a week, will provide another opportunity to assess the longevity of auditory statistical segmentation representations. Because participants are tested twice, any improvement in performance on the second session suggests some degree of retention over the intervening week. If performance improves, this would suggest that SL yields representations that are stored in long-term memory rather than reflecting a more transient effect. Note that although it would be valuable to also explore this question using a design without an initial test or exposure period (thus minimizing test–retest effect), we believe the present design is a suitable test of the longevity of the representations. Although participants may remember aspects of the first test as well as the first familiarization period, remembering the test is unlikely to influence responding on the second session because there is no feedback. Even if participants remembered their

response, there is no principled reason (because they do not receive feedback) why this memory should improve performance; instead, it should serve to make it more similar from Time 1 to Time 2. Similarly, although it would be useful to test participants at a second time point without additional exposure, if learning from the material is the only factor that influences performance on the second test, performance would be predicted to be equivalent on both sessions. Thus, if performance is elevated after the second session, this would indicate retention over the intervening week.

Method

Participants

We tested 96 undergraduate students from Florida State University, who participated for research credit. Of those participants, 77 returned for the second session. Due to a combination of technical problems, experimenter error, and participant dropout, the number of participants with data from each of the four languages and the two different assessment types varies (Time 1 2AFC: Language 1 = 95; Language 2 = 95; Language 3 = 92; Language 4 = 95; RS: Language 1 = 96; Language 2 = 96; Language 3 = 93; Language 4 = 95; Time 2 2AFC: Language 1 = 76; Language 2 = 75; Language 3 = 74; Language 4 = 74; RS: Language 1 = 78; Language 2 = 77; Language 3 = 76; Language 4 = 77).

Auditory Stimuli

The auditory stimuli consisted of four distinct artificial languages (three syllabic and one tonal[3]) used in previous studies [3, 48, 51, 52]. Although the languages varied in their perceptual characteristics (e.g., word length and speaker identity), each language consisted of four words repeated in a pseudo-randomized order such that no words immediately repeated. Each was designed so that it could only be segmented on the basis of conditional structure. Within each language, transitional probabilities (TPs) were higher between elements that formed words (all word TPs = 1.0) than sequences of elements that occurred incidentally across word boundaries or part-words (part-word TPs = 0.2 to 0.4). Testing for each language consisted of two sets of 8 questions (8 RS questions and 8 2AFC questions).

One potential concern with the presentation of multiple languages within a single session is the possibility of interference between the languages. However, although there was overlap in the particular syllables used across languages, prior research indicates learners can learn multiple languages when they are distinguished by a perceptual cue such a distinct speaker [53], or are separated by a brief pause [54]. All of the languages were spoken in distinct voices, with the exception of Languages 1 and 3, which were produced with the same synthesizer but different word lengths. In addition, Language 3 contains only stop consonants, and is noticeably faster than Language 1. These features make it unlikely that participants were unable to discriminate the languages. However, to ensure that the presentation of multiple languages concurrently did not cause interference that prevented participants from learning the languages, order effects were examined via ANOVAs that tested whether performance on each language at each time point was influenced by its position in the order (e.g., whether performance on Language 1 at Time 1 assessed via 2AFC differed as a function of whether it was presented in position 1, 2, 3, or 4). However, of all the comparisons (16 in total), only one order effect was observed. Performance assessed by the RS measure for Language 1 at Time 1 showed differences as a function of position, $F(3,88) = 3.52$, $p = .02$). Tukey post-hoc follow up tests indicated this effect was driven by superior performance when Language 1 was presented in position 4 relative to position 3, $p = .016$, with no other significant pairwise comparisons. No position effects were observed for any of the other measures or time points, all $ps > 0.16$. Moreover, performance was not unusually low on any of the languages relative to what has been observed previously (e.g., [36, 37, 55, 51]). Thus, it is unlikely the presentation of multiple languages within a single session prevented participants from learning the statistical structure of the languages.

Language 1. The first language is the language used in Saffran et al. [3], which was produced using a synthesizer in monotone female voice at approximately 220 Hz, and consisted of four trisyllabic words: *bidaku, tupiro*, *golabu,* and *padoti*. The language was a portion of the original experimental exposure of approximately 1 minute that was concatenated 10 times for a total of 10.02 minutes exposure. It was spoken at a rate of 270 syllables per minute. The test items were recorded in isolation, and consisted of two of the words, *tupiro* and *bidaku* which were paired exhaustively with two part-words, *tigola* and *bupado*, which were formed by joining syllables from the words *golabu* and *padoti*.

Language 2. The second language was originally recorded by Thiessen et al. [52], and was produced in fluent, infant-directed sentences by a native English speaker who was naïve to the statistical structure of the language. The 1.02 minute language included 12 acoustically distinct sentences, each of which each started with the syllable "mo" and ended with the syllable "fa" so that the silences between sentences could not be used as a cue to word boundaries. The 1.02 minute language was concatenated 10 times for a total of 10.3 minutes. Each sentence contained a different order of four nonsense words (*dibo, kuda, lagoti, nifopa*). They were spoken at an average rate of 2.5 syllables per second (150 syllables per minute) and 1.3 sec of silence separated the sentences. The average pitch (measured by fundamental frequency, F0) of the speaker was 292 Hz, with a range of 140–480 Hz. This large range reflects the exaggerated pitch contours of infant-directed speech, and is consistent with previous work on the characteristics of infant-directed speech (e.g., [56]). The same speaker recorded the four test items in isolation. The test items consisted of the words *lagoti*, and *dibo* and the part-words *danifo*, and *paku,* which are composed of the final syllable of one of the words *kuda* and *nifopa,* and the first two syllables of the other word.

Language 3. Language 3 was identical to the monotone language used by Thiessen and Saffran [48], and was

produced in a monotone with a fundamental frequency of 200 Hz. It consisted of four disyllabic words: *diti, bugo, dapu,* and *dobi*. The language was 2 minutes long and spoken at a rate of 270 syllables per minute, concatenated together 5 times for a total duration of 10.05 minutes. Two of the words (*diti* and *bugo*) were twice as frequent as the other words (*dapu* and *dobi*), which allowed for the creation of test items that were frequency balanced. The test items were the two infrequent words (*dapu* and *dobi*) that were paired with two part-words (*godi* and *tibu*) created by splicing together the two more frequent words. In other words, although *dapu* and *godi* were equally frequent in the artificial language, the internal transitional probabilities between *dapu* (the word) were higher than the transitional probabilities between *godi* (the part-word; for a detailed discussion of frequency balancing, see [57]).

Language 4. The fourth language was a tonal language with the same structure as the Language 1, which was modeled on a tonal language used previously [51]. Each syllable from the linguistic version of the language was replaced with a unique tone that was 30 ms in duration, slightly longer than standard presentation of syllables in line with previous research [51]. Thus, the language consisted of four tone words (*ADE*, *BFG*, *CC#D#*, and *G#A#F#*). The language was 1.375 minutes long, which was repeated 10 times for a total duration of 13.75 minutes. The test items were two tone-words, *ADE*, and *BFG*, paired with two part-words: *D#G#A#* and *F#CC#*, formed from components of the words *CC#D#*, and *G#A#F#*.

Procedure

Participants were tested individually during two sessions approximately a week apart (mean delay = 7.15 days, *SD* = 1.30 days). They were tested on each of the four languages in a pseudo-randomized order that was identical for each participant in both sessions. Language order was counterbalanced across participants, as was the order of the two tests, the 2AFC assessment (the percentage of selections of statistical words over part-word foils) and the RS test (average difference in endorsement of words and part-words). In the 2AFC test, participants heard two items (a word and part-word) and were asked to identify which item sounded "more familiar" to the exposure stimulus. There were 8 test trials in which words and part-words were exhaustively paired. In the RS test, participants heard each test item individually, and were asked to rate how familiar (on a scale of 1–5) the item sounded to the language; participants rated each item (2 words, and 2 part-words) twice for a total of 8 test trials assessing ratings for 4 distinct words and 4 distinct part-words. The experiment was administered using E-Prime software [58], and the auditory stimuli were presented over headphones. Signed consent was obtained for all participants, and testing was conducted in accordance with the ethical standards established by the university's Institutional Review Board.

Results

Time 1

To confirm that participants were able to learn the languages as in previous research with these measures, a series of analyses were conducted on both the 2AFC and RS tests.

The results of the 2AFC tests from Time 1 are presented in **Table 1**. Participants performed above chance on all of the languages, indicating that they had successfully segmented words from the training set. The results of the RS tests from Time 1 are presented in **Table 2**. The difference in ratings for the words and part-words was significantly different from 0 for all languages (with words being rated more highly than part words), indicating that participants were successful in learning the language from the training set.

A series of correlations were performed to assess how performance on the different measures at Time 1 were related to each other, both in terms of the correlations between different response formats of a specific language (e.g,. Language 1 RS and Language 1 2AFC) as well as across languages within a response format (e.g., Language 1 RS

		Language 1	Language 2	Language 3	Language 4
Time 1	*M*	0.58	0.56	0.57	0.83
	SD	0.23	0.23	0.22	0.18
	One-sample *t*	3.15**	2.52*	3.23**	17.83***
	df	94	94	91	94
Time 2	*M*	0.61	0.63	0.64	0.82
	SD	0.23	0.21	0.22	0.18
	One-sample *t*	3.99***	5.44***	5.50***	15.79***
	df	75	74	73	73
Cross-time Comparisons	*r*	0.49***	0.097	0.28*	0.24*
	Paired-samples *t*	1.13	2.60*	1.285	0.19
	df	74	73	70	72

Table 1: Performance on 2-Alternative Forced Choice Assessment at Time 1 and Time 2 in Experiment 1.
Note: *$p < .05$, **$p < .01$, ***$p < .001$; 0.5 represents chance performance.

		Language 1	Language 2	Language 3	Language 4
Time 1	*M*	0.19	0.21	0.44	1.49
	SD	0.91	0.75	0.88	0.94
	One-sample *t*	2.085*	2.77**	4.89***	15.38***
	df	95	95	92	94
Time 2	*M*	0.22	0.33	0.41	1.43
	SD	0.82	0.81	0.99	1.08
	One-sample *t*	2.41*	3.57***	3.64***	11.57***
	df	77	76	75	76
Cross-time Comparisons	*r*	0.40***	0.23*	0.27*	0.26*
	Paired-samples *t*	.029	1.26	.087	0.31
	df	76	75	72	74

Table 2: Performance on Rating Scale Assessment at Time 1 and Time 2 in Experiment 1.
Note: *$p < .05$, **$p < .01$, ***$p < .001$; 0 represents chance performance (no difference in endorsement of words and part-words).

and Language 2 RS). The correlations between individual languages are presented in **Table 3** (2AFC tests) and **Table 4** (RS tests). For the 2AFC data, none of the languages were significantly correlated with each other, *p*s > .05. For the RS data, the only significant relationship was a positive correlation between rating performance on Language 1 and Language 3, $r = 0.24$, $p = .023$[4]. Although performance on the different languages was largely uncorrelated, RS and 2AFC performance was correlated for each language (Language 1: $r = 0.56$, $p < .001$; Language 2: $r = 0.35$, $p = .001$; Language 3: $r = 0.38$, $p < .001$; Language 4: $r = 0.29$, $p = .004$).

Time 2

As with Time 1, a series of analyses were completed to determine how individuals performed on the different languages and with the different response formats at Time 2. Performance on the 2AFC and RS tests in Time 2 are presented in **Tables 1** and **2**, respectively. As was found for Time 1, participants demonstrated learning of the languages for both test types. In addition, correlations were performed to assess for interrelations between performance across languages within response formats as well as across response formats within languages. The correlations between the individual languages are presented in **Table 5** (2AFC tests) and **Table 6** (RS tests). For the 2AFC data, none of the languages were significantly correlated with each other, *p*s > .05. For the RS data, there was a significant positive correlation between rating performance on Language 1 and Language 3, $r = 0.30$, $p = .009$. There was also a significant positive correlation between performance on Language 1 and Language 4, $r = 0.23$, $p = .045$. There were no other significant correlations, *p*s > .05. Finally, 2AFC and RS performance was correlated for each language (Language 1: $r = 0.52$, $p < .001$; Language 2: $r = 0.50$, $p < .001$; Language 3: $r = 0.51$, $p < .001$; Language 4: $r = 0.47$, $p < .001$).

Comparing Time 1 and Time 2

There were two main issues to address regarding the comparison of performance in Time 1 and Time 2. The first was determining whether performance on any of the SL tasks show test–rest reliability, as demonstrated by correlations between performance at Time 1 and Time 2. The Time 1–Time 2 correlation for each language is presented in **Table 1** (2AFC) and **Table 2** (RS). For the 2AFC tests, three of the four languages were significantly correlated between Time 1 and Time 2. For the RS tests, all of the languages showed significant correlations between Time 1 and Time 2.

The second issue we wished to address in comparing performance across Time 1 and Time 2 is whether participants performed better at the second time point, providing evidence that the learning from Time 1 aided performance at Time 2. Although accuracy was generally numerically higher at Time 2 than at Time 1, only the 2AFC assessment for Language 2 showed significant improvement between Time 1 ($M = 0.54$; $SD = 0.24$[5]) and Time 2 ($M = 0.63$, $SD = 0.21$; (see **Tables 1** and **2**; see also **Fig. 1** for scatterplots of performance).

Because each individual measure necessarily involved a small number of unique test items, a compositing approach was used to investigate whether averaging across measures would result in a measure with better psychometric properties than observed for any individual tasks. Although we acknowledge that caution should be observed in averaging across measures that are not correlated, we argue that this approach is warranted based on the high similarity of the materials and the identical testing procedures used for each task. The lack of correlation between the individual measures may be the result of the low number of unique items, or repeated items, or because successful performance on different languages taps into different abilities. Regardless, compositing is a useful strategy that may provide a more complete summary of a

		Language 1	Language 2	Language 3	Language 4
Language 1	*r*	1			
	n	95			
Language 2	*r*	–0.17	1		
	n	94	95		
Language 3	*r*	0.14	–0.10	1	
	n	91	91	92	
Language 4	r	–.048	–0.11	–.011	1
	n	94	94	91	95

Table 3: Correlations Between Individual Languages Assessed Via 2-Alternative Forced-Choice Test at Time 1 in Experiment 1.
Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

		Language 1	Language 2	Language 3	Language 4
Language 1	*r*	1			
	n	96			
Language 2	*r*	–0.16	1		
	n	95	96		
Language 3	*r*	0.24*	.074	1	
	n	92	92	93	
Language 4	r	–.062	.036	.079	1
	n	94	94	91	95

Table 4: Correlations Between Individual Languages Assessed Via Rating Scale Test at Time 1 in Experiment 1.
Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

		Language 1	Language 2	Language 3	Language 4
Language 1	*r*	1			
	n	76			
Language 2	*r*	.068	1		
	n	75	75		
Language 3	*r*	0.11	.22	1	
	n	74	73	74	
Language 4	r	.001	.010	–.10	1
	n	74	73	72	74

Table 5: Correlations Between Individual Languages Assessed Via 2-Alternative Forced-Choice Test at Time 2 in Experiment 1.
Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

given individual's statistical learning abilities. Composite scores for the 2AFC and RS performance were made by averaging performance on all four languages together at each of the two time points. A comparison of performance over time for the composite measures can be seen in **Fig. 2**. For the 2AFC composite, performance at the two time points was significantly correlated, $r = 0.45$, $p < .001$. This correlation was larger in magnitude than three of the four individual correlations, as well as an average of the four individual *r* values. In addition, learning effects from Time 1 ($M = 0.63$, $SD = 0.10$) to Time 2 ($M = 0.67$, $SD = 0.11$) were observed, $t(76) = 2.85$, $p = .006$. These results indicate that at least for the 2AFC assessments, composite measures provide an advantage not seen with any of the individual measures, namely the ability to detect learning effects over delay of a week.

For the RS composite, performance at the two time points was also significantly correlated, $r = 0.36$, $p = .001$. Unlike

		Language 1	Language 2	Language 3	Language 4
Language 1	*r*	1			
	n	78			
Language 2	*r*	–.028	1		
	n	77	77		
Language 3	*r*	0.30**	0.14	1	
	n	76	75	76	
Language 4	r	0.23*	0.12	0.12	1
	n	76	75	74	77

Table 6: Correlations Between Individual Languages Assessed Via Rating Scale Test at Time 2 in Experiment 1. Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

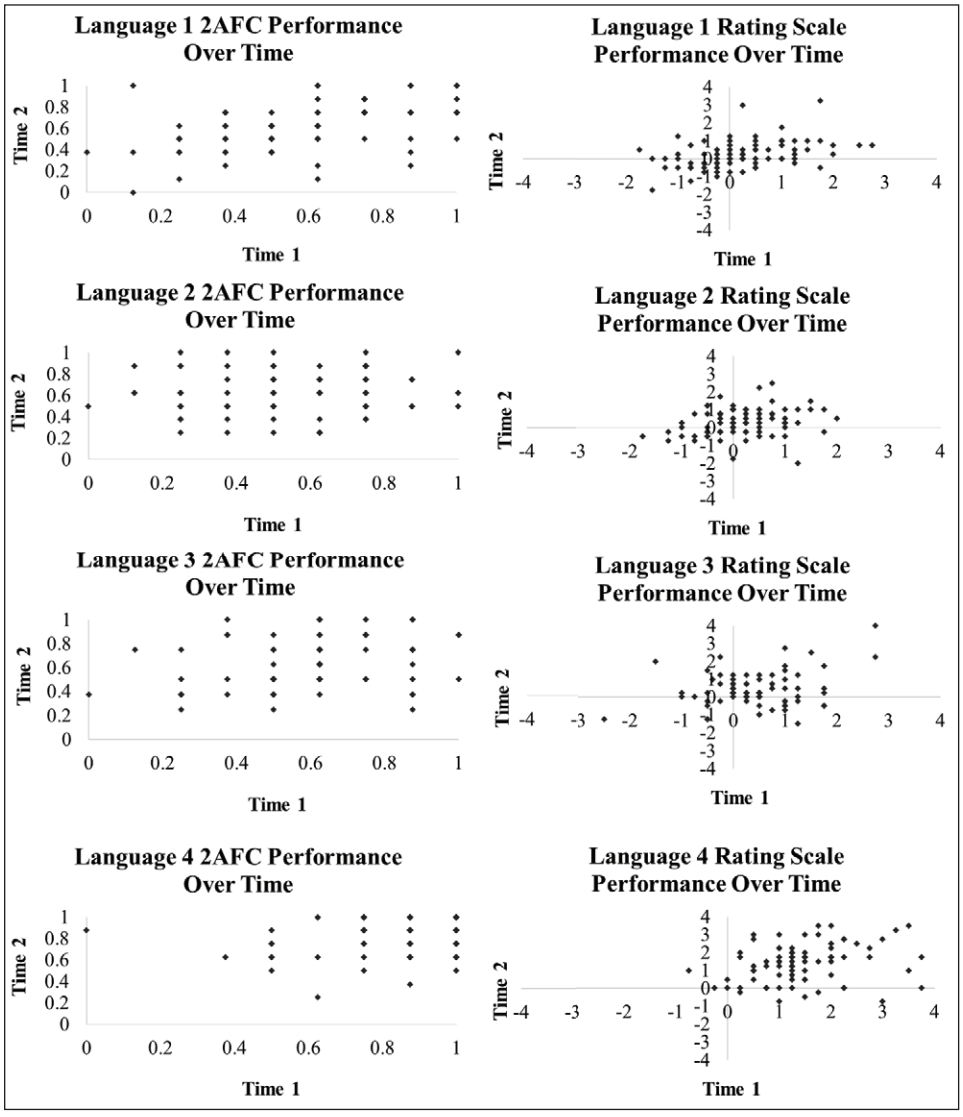

Figure 1: Relationship between performance at Time 1 and Time 2 in Experiment 1 for the individual measures, assessed by both the 2-Alternative Forced Choice and Rating Scale tests.

with the accuracy composites, no learning effects were found between Time 1 ($M = 0.57$, $SD = 0.47$) and Time 2 ($M = 0.62$, $SD = 0.61$), $t(78) = 0.71$, $p = 0.50$. To determine whether the lack of learning effects was due to changes in endorsement of words or part-words, composites of the average endorsement for each item were compared from Time 1 to Time 2. There were no differences in endorsement across time for words, (Time 1: $M = 4.0$; $SD = 0.43$; Time 2: $M = 4.0$; $SD = 0.45$; $t(74) = 0.39$, $p = 0.70$), or for part-words, (Time 1: $M = 3.43$; $SD = 0.46$; Time 2: $M = 3.39$;

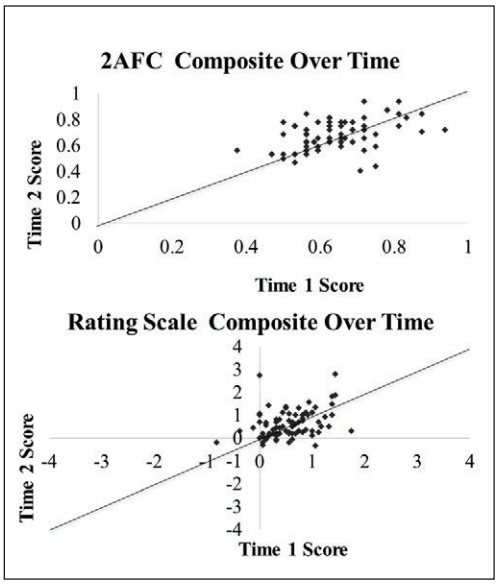

Figure 2: Relationship between performance at Time 1 and Time 2 in Experiment 1 for the composite measures, assessed by both the 2-Alternative Forced Choice and Rating Scale tests.

SD = 0.57; $t(74) = 0.45$, $p = 0.65$). Unlike the 2AFC measure, a composite measure of the RS did not detect changes in learning over time. This is largely due to the fact that even though the increase in performance from Time 1 to Time 2 is of the same magnitude in the RS composite as in the 2AFC composite, the RS composite shows much greater variability. These results raise the possibility that the RS measure is less sensitive than the 2AFC measure, at least when these scores are composited.

Similar results were observed when 2AFC-RS composites were created for each language. The cross-time comparisons significant for all of languages, but the correlations were weaker than those between the composites collapsed across all the languages (Language 1 $r = 0.45$; Language 2 $r = 0.26$; Language 3 $r = 0.33$; Language 4 $r = 0.28$). One potential explanation for the lower reliability is that only two measures were used to compute the composites. Another possibility is that success with different languages draws on distinct abilities and the composite that incorporates more measures provides a more stable estimate of generalized SL ability.

Discussion

There are several results of interest from Experiment 1. Performance on the four word segmentation tasks was largely uncorrelated, although there were a few weak but significant correlations between specific languages. Performance on the 2AFC and RS assessments of individual languages were correlated (and comparable). Three of the four tasks showed significant correlations between the first and second time points for both of the assessment types. Consistent with Siegelman and Frost [42], we found that the reliability for individual tasks was variable and generally low, ranging from around 0.45 to around 0.20. When performance for all tasks was averaged at Time 1 and Time 2, the reliability of the SL measure (both 2AFC and RS measures) improved markedly over what was seen for most of the individual measures, although the composite reliability was still somewhat low (~0.45). The improvement in reliability for the composite is likely attributable to the increase in the number of test items, as auditory statistical segmentation languages are designed such that few distinct items are possible.

When performance on the individual measures was composited, another effect that was not present in the analysis of individual languages (with the exception of the 2AFC assessment for Language 2) emerged. For the 2AFC assessment, an improvement in performance from Time 1 to Time 2 emerged. This facilitation suggests that SL results in long-lasting representations, which is a necessary feature of any mechanism proposed to play a major role in language acquisition. To our knowledge, this represents one of the first demonstrations that learning in a putatively linguistic SL segmentation task is preserved over lengthy intervals (see [49, 50] for demonstrations with visual SL segmentation over a shorter delay; [42], for a demonstration of long-lasting learning in an auditory task with non-verbal stimuli; and [59, 60], for linguistic adaptations persisting for a week). Although this discovery is not surprising, the fact that it was only observed when multiple measures are taken in aggregate highlights the usefulness of using multiple convergent measures of SL.

We take a few things away from these results. First, although the individual SL measures showed significant test–retest correlations, the correlations were generally weak. This suggests that poor reliability may be a factor in the low-to-non-existent correlations seen between the individual SL measures. Note that this claim does not rule out the possibility that the differences across SL tasks results in low correlations between measures in other circumstances, as we only explored this issue in the context of the statistical word segmentation paradigm. Second, the largely comparable results observed across the 2AFC and RS assessments of SL suggest that the nature of the assessment used in the tasks does not play a major role in shaping performance, with the caveat that the learning effects from Time 1 to Time 2 were only evident with the 2AFC measure. Finally, the increased reliability that resulted from creating a composite index of performance from all the SL tasks suggests the possibility that the use of such composite measures may be a promising avenue for improving the measurement of individual variation in SL. Many questions remain, such as whether the advantages of the composite stem from simply having more items, and whether similar results would be obtained with other measures of SL (e.g., the SRT). To explore whether the advantages of compositing are contingent on the presence of multiple distinct test items, Experiment 2 explores the effects of test item repetition on test–retest reliability on a subset of the languages tested in Experiment 1.

Experiment 2

In Experiment 1, we observed that pooling performance across SL tasks lead to generally higher reliability than what was observed when looking at the individual

measures. One explanation of this finding is that having more test items results in a more stable measure of the participants' performance. Experiment 2 was designed to test this hypothesis on individual measures of SL from Experiment 1. We selected two of the SL tasks from Experiment 1, and lengthened the forced-choice test that we gave participants to assess their learning. Our choice to pick two of the word segmentation tasks for this study was driven in part by practical concerns. Limiting the number of word segmentation tasks we employed in the experiment allowed us to assess whether increasing test length would improve SL task reliability and result in correlations between similar SL tasks, while also providing room for us to collect additional measures of interest for an exploratory analysis of the relationship between our SL measures and other cognitive constructs[6]. Because word segmentation tasks typically use a small number of words in their training set, our options for creating a larger set of unique test items were limited. As such, we lengthened the tests by repeating the original test items.

In addition to assessing questions about the reliability and inter-correlations among the word segmentation tasks in Experiment 1, Experiment 2 also explored the relationship between word segmentation measures of SL and other SL measures. We included a different measure of SL in this experiment, namely an artificial grammar syntax-learning task that was used by Kaschak & Saffran [61, 62]. We included this task for two reasons. First, on the assumption that increasing test length for our word segmentation tasks would improve task reliability and potentially boost the correlations between the tasks, we thought it would be useful to include an entirely different kind of statistical learning task to assess whether the segmentation tasks would be correlated with it. Based on the results of Siegelman and Frost [42], among others, we expected that the artificial grammar learning task would not be strongly correlated with the word segmentation tasks. A second reason for including this task was that the test for this measure is much more extensive (that is, there are more test items) than the test for the word segmentation tasks. If having reliable SL measures is (at least in part) a function of having an assessment of performance with a number of items, we expect that the grammar learning task should produce reasonably strong test–retest reliability (but see [63], for evidence that suggests that this may not be the case).

Method

Participants

We tested 80 undergraduate students from Florida State University, who participated for research credit. Of those participants, 65 returned for the second session. Due to a combination of technical problems, experimenter error, and participant dropout, the number of participants with data from each of the three SL measures at each time point varies (Time 1 2AFC: Language 1 = 80; Language 3 = 78; Artificial Grammar Learning = 79; Time 2 2AFC: Language 1 = 63; Language 3 = 65; Artificial Grammar Learning = 63).

Auditory Stimuli

The auditory stimuli consisted of two of the four languages from Experiment 1 (Languages 1 and 3). These languages were selected because they were the most similar (both syllabic; produced in a monotone by a synthesizer) and because they showed the highest test-retest reliability (Language 1 $r = 0.49$; Language 2 $r = 0.28$). Consequently, they should be most likely to correlate with each other. Participants were only tested on the 2AFC assessments, which were increased from 8 to 16 by repeating trials. In addition, an artificial grammar learning task was included to assess how its psychometric properties compared to those of the speech segmentation tasks.

Artificial Grammar Learning Task. The artificial grammar learning task was an auditory phrase learning task, which was very similar to the task used in Kaschak and Saffran (2006). It consisted of auditory strings that that confirmed to the rules of a core grammar. In addition, 14% of the strings followed a different pattern designed to approximate idiomatic phrases (e.g., "*Him be a doctor?*"), which were presented in a different intonational contour (core sentence followed descending prosody, whereas idiomatic sentences followed an ascending prosody) and always contained the word, "*wug*". The pattern of the phrases was as follows:

A) S = A-Phase (AP) + C-Phase (CP) + E-word
B) AP = A-word + optional D-word
C) CP = C-word + optional G-word
D) Idiomatic: E-word + CP + *wug* + C-word

The A-words consisted of: *bif, hep, mib, rud*; the C-words consisted of: *cav, lum, neb, sig*; the D-words consisted of: *klor, pell*; the G-words consisted of: *tiz, pilk*; and the E-words consisted of: *jux, vot, loke, dupp*. As detailed in Kaschak & Saffran (2006), the grammar could also be described using the following set of rules:

Core Grammar:
Rule 1: All sentences must have an A-phrase.
Rule 2: In an A-phrase, A-words must precede D-words
Rule 3: In a C-phrase, C-words must precede G-words.
Rule 4: Sentence must have a G-word.
Rule 5: C-phrases must precede E-words.
Rule 6: If there is a G-word, there must be a C-word.

Idiomatic Phrase Rules:
Rule 7: An E-word and a C-phrase must precede the "*wug*" word.
Rule 8: The final C phrase must only be a C-word (no G-words in the final position).

Thus, there were several acceptable patterns of legal sentences that varied in length from three to five words. The acceptable patterns of sentences was as follows:

Core Sentence Patterns
A-C-E
A-D-C-E

A-D-C-G-E
A-C-G-E

Idiomatic Construction Patterns
E-C-*wug*-C
E-C-G-*wug*-C

Participants listened to the language for approximately 7 minutes. Subsequently, they completed 50 test trials in which they listened to two utterances and were asked to select the grammatical one. Scoring consisted of accuracy averaged across the 50 test trials.

Procedure

Participants were tested individually during two sessions approximately a week apart (mean delay = 7.27 days, *SD* = 1.34 days). They were tested on each of the three SL tasks languages in a pseudo-randomized order that was identical for each participant in both sessions. The SL tasks always occurred in the first, middle, and last positions, and the order of these tasks was counterbalanced across individuals. The auditory segmentation tasks used 16 test trials rather than the 8 test trials used in Experiment 1. Between the SL tasks, participants completed the other cognitive measures mentioned previously, which served the additional function of minimizing potential interference between SL tasks. As with Experiment 1, signed consent was obtained for all participants, and testing was conducted in accordance with the ethical standards established by the university's Institutional Review Board

Results

Time 1

Mirroring the analyses in Experiment 1, performance on each language was compared to chance to determine how well individuals had learned the languages. The results of the 2AFC tests from Time 1 are presented in **Table 7**. Participants performed above chance on all of the languages, which indicated that they had successfully learned the structure of the training sets[7]. A set of correlations was also performed to assess whether correlations between individual languages emerged. The correlations between individual languages are presented in **Table 8**. As reported previously in Experiment 1, none of the languages were significantly correlated with each other, *p*s > .05.

		Language 1	Language 3	Artificial Grammar Learning
Time 1	*M*	0.60	0.69	0.60
	SD	0.20	0.20	.070
	One-sample *t*	4.28**	8.39*	12.92**
	df	79	77	78
Time 2	*M*	0.63	0.68	0.61
	SD	0.25	0.25	.072
	One-sample *t*	4.46***	6.0***	11.90***
	df	62	64	62
Cross-time Comparisons	*r*	0.66***	0.59**	0.31*
	Paired-samples *t*	1.20	1.03	1.18
	df	61	61	59

Table 7: Performance on 2-Alternative Forced-Choice Assessment at Time 1 and Time 2 in Experiment 2.
Note: *$p < .05$, **$p < .01$, ***$p < .001$; 0.5 represents chance performance.

		Language 1	Language 3	Artificial Grammar Learning
Language 1	*r*	1		
	n	80		
Language 3	*r*	0.17	1	
	n	77	78	
Artificial Grammar Learning	*r*	–.023	–0.11	1
	n	77	77	79

Table 8: Correlations Between Individual Languages Assessed Via 2-Alternative Forced-Choice Test at Time 1 in Experiment 2.
Note: *$p < .05$, **$p < .01$, ***$p < .001$.

Time 2

The same analyses were performed to assess how well individuals had learned the languages. The results of the 2AFC tests from Time 2 are presented in **Table 7**. Participants performed above chance on all of the languages, indicating that they had successfully learned the structure of the training set. Correlations were also performed to assess interrelations between performance on the different languages. The correlations between individual languages are presented in **Table 9**. As in Experiment 1, none of the languages were significantly correlated with each other, *p*s > .05.

Comparing Time 1 and Time 2

As in Experiment 1, the test–rest reliability of the three SL tasks was assessed by examining the correlations between performance on each measure at both time points. The Time 1–Time 2 correlations for each of the measures is presented in **Table 7**. All three of the measures exhibited significant positive correlations between performance on Time 1 and Time 2. Test–retest reliability for the word segmentation tasks was improved over what had been observed in Experiment 1 (Language 1 $r = 0.66$; Language 3 $r = 0.59$), and was comparable to the best test–retest reliabilities reported in Siegelman and Frost (2015). Test–retest reliability for the artificial grammar learning task was quite low in comparison ($r = 0.31$). One possible explanation for this is that performance on the artificial grammar learning task showed restricted variability relative to performance on the two statistical segmentation tasks (See **Figure 3** for scatterplots of performance).

Correlations were also computed for composite measures incorporating the two word segmentation measures, as well as a composite created from all three SL measures. The correlations between performance at the first and second time points were numerically larger than any of the individual correlations (Word Segmentation Composite $r = 0.74$, $p < .001$; SL Composite $r = 0.73$, $p < .001$). In contrast to the findings of Experiment 1, no learning effects were observed between Time 1 and Time 2 for any of the individual measures or for either of the composite measures (all *p*s > .05).

Discussion

There are several results of interest from Experiment 2. Test–retest reliability for the word segmentation tasks was improved by increasing the number of test items. The reliabilities observed for these tasks were comparable to the highest reliabilities that have been reported for other SL measures (e.g., [42]). Furthermore, as was the case in Experiment 1, we found that a composite measure of word segmentation tasks, as well as a composite measure consisting of the word segmentation tasks and the artificial grammar learning task, was observed to have higher test-retest reliability than any of the individual measures. These findings are consistent with our conjecture that adding test items would generally improve the reliability of our SL measures. Nonetheless, it is noteworthy that the artificial grammar learning task did not have strong reliability on its own (despite having more test items than the word segmentation tasks). This suggests that having strong test–retest reliability for SL measures may not simply be a function of having more test items.

It is also worth noting that although test item repetition did increase test–retest reliability, performance on the three SL measures was uncorrelated. Although the reliability of the measures still places a cap on the ability of the measures to correlate with each other, these results are somewhat troubling for theories that posit statistical learning is a unitary capacity, given that these measures were selected to be highly similar (auditory SL of adjacent regularities). This is an important issue, which we will revisit in the General Discussion.

Finally, unlike Experiment 1, no learning effects were detecting using any of the individual measures or the composites. It is unclear why the learning effects that were observed in Experiment 1 were not found in Experiment 2, especially as performance at Time 1 was not at ceiling. However, this result is in keeping with other results in the literature suggesting that SL measures may not consistently produce learning gains across time (e.g., [42]). One possibility is that the inconsistency of the learning effects can be explained by a combination of weak effects and noisy measurement.

		Language 1	Language 3	Artificial Grammar Learning
Language 1	*r*	1		
	n	63		
Language 3	*r*	0.17	1	
	n	63	65	
Artificial Grammar Learning	*r*	–.006	0.23	1
	n	61	62	63

Table 9: Correlations Between Individual Languages Assessed Via 2-Alternative Forced-Choice Test at Time 2 in Experiment 2.

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

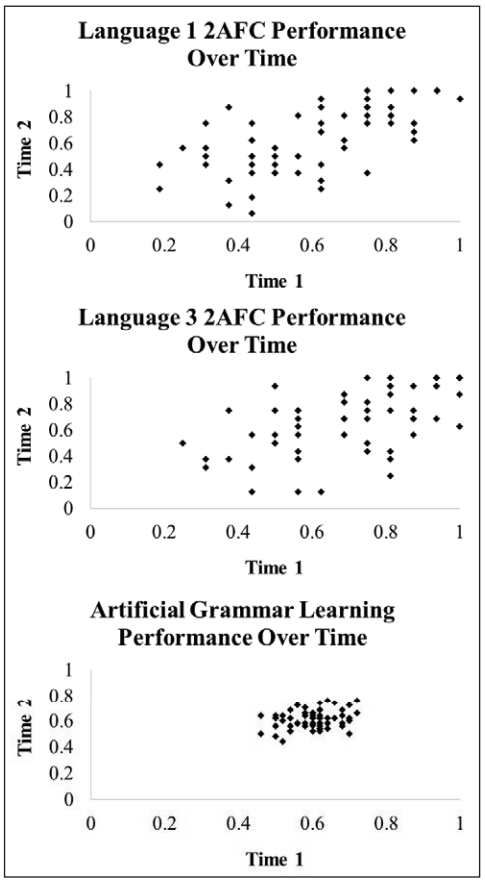

Figure 3: Relationship between 2-Alternative Forced Choice performance at Time 1 and Time 2 in Experiment 2 for the two word segmentation measures and the artificial grammar learning measure.

General Discussion

The work reported here addresses two issues about the measurement of individual differences in SL ability. First, are the generally low correlations between measures of SL due to task demands, the psychometric properties of the tasks being used, or the fact that SL is a highly fragmented construct? We addressed this question by studying performance on a set of SL tasks that are highly similar to each other (the auditory word segmentation tasks used in Experiment 1). That is, we constructed a measurement situation where all of our SL measures (with the possible exception of the AGL learning task in Experiment 2) had very similar learning demands, and very similar testing demands. Unlike previous work in this area (e.g., [42]), which used a number of distinct SL tasks, our Experiment 1 established a set of conditions that should optimize the chances of seeing significant correlations between SL tasks. The correlations we found were generally very low, suggesting that the low correlations between SL tasks were not necessarily being driven by differences in task demands, and were not necessarily being driven by fragmentation in the SL construct (since all of our measures in Experiment 1 assessed the same dimension of SL). In keeping with previous work (e.g., Siegelman & Frost, 2015 [42]), we found that the test–retest reliability for our measures was on the low side (reliabilities for our tasks were < 0.6), which is potentially consistent with the possibility that the low correlations between tasks was the result of poor psychometric properties of our measures.

The second issue we addressed in this work is the question of whether improving the reliability of our measures would increase the correlations between measures of SL. We endeavored to increase the reliability of our measures by adding to the length of the assessment of learning for our SL tasks (Experiment 2), and by generating composite measures that pooled performance across tasks (Experiments 1 and 2). The composite measures were generally more reliable than the individual SL measures, and the word segmentation tasks with longer tests in Experiment 2 were generally more reliable than their counterparts with shorter tests in Experiment 1. Despite the improved reliabilities, the correlations between measures of SL remained low.

The lack of a strong correlation between the word segmentation tasks in Experiment 2 (despite their reasonable level of reliability) is puzzling in light of the fact that the two tasks have extremely similar demands. One possible explanation is that although these measures were chosen to be highly similar (particularly relative to the differences between tasks seen in previous investigations), there were still some differences between these languages that may have contributed to the lack of correlation. For example, one language consisted of three-syllable words and the coherent test items appeared more frequently in the input than the foils. In contrast, the other language consisted of two-syllable words and the coherent test items and foils appeared with equal frequency in the exposure phase. It may be that these features play an important role in driving performance. For example, prior research has indicated that perceptual alterations—such as timing—can lead to differential learning of the same statistical structure (e.g., [44, 52]). More broadly, this possibility is consistent with accounts that suggest that "statistical learning" is actually composed of many different mechanisms, each operating over a different set of perceptually distinct input (e.g., [34]). If these features exert a strong influence on the mechanisms that are engaged during statistical learning, the subtle differences in the tasks might result in weak inter-language correlations, a possibility which is potentially consistent with prior research (e.g., [42, 64]). Such weak correlations could be difficult to detect (a sample size of close to 800 would be necessary to detect a correlation of 0.1 with power of 0.8; [65]).

Alternatively, it may be the case that the different statistical structures of the languages themselves invoke different kinds of learning mechanisms. For example, word length may place different demands on working memory, which may influence performance. Similarly, learning driven by frequency may be mediated by distinct mechanisms than learning driven by transitional probabilities (e.g., [45]). A final possibility worth considering is that participants are typically not given explicit instructions about what to do during the training or test segments of this type of SL task (e.g., what criteria does the participant

use to distinguish "good" from "bad" items?). As such, performance on these tasks may be affected to an unknown degree by the strategies that the participants employ at training or test. That is, even similar SL tasks (such as ours) may elicit different strategies, and will therefore be measuring different constructs. In this regard, it is interesting to note that our composite measures were typically more reliable than individual SL measures. Perhaps compositing across SL measures creates a situation where the strategic factors cancel out across tasks, and the resulting measure taps SL more directly.

Our results, in conjunction with the findings of Siegelman and Frost [42] pose a challenge to theoretical accounts of statistical learning. Most extant accounts of statistical learning either treat the construct as unitary, or partition the construct into a small number of sub-domains (e.g., [9]). For example, the account that we have previously proposed (The Extraction and Integration Framework; [9]) conceptualizes SL as two distinct but interrelated mechanisms. The lack of correlation between measures of SL is difficult to reconcile with either a unitary approach to SL or an account that posits a small number of sub-domains. Either SL is a highly fragmentary construct (perhaps to the extent that we might wonder whether it is a useful theoretical construct at all), or the lack of correlation between SL tasks is the result of measurement issues. Given that there has been so little effort to develop and validate measures of SL, it seems most profitable to pursue the latter hypothesis first. However, it is also important to consider the possibility that construct error rather than measurement error is that the heart of these null results, and that SL may not be a useful theoretical construct in the context of individual differences approaches (see [7], for a review exploring the possibility that SL is a mechanism involved in acquiring natural languages).

This effort is critical for both practical and theoretical progress. If SL measures are to be used as instruments to explore individual differences in domains such as language learning, it will be critical to leverage knowledge about measurement to ensure that tasks used possess the appropriate psychometric properties needed to yield meaningful correlations (See [66], for some initial suggestions about how the psychometric properties of SL measures may be improved in the context of an exploration of the psychometric properties of visual SL). Similarly, if we are to make progress in understanding the extent to which statistical learning is composed of one or several underlying mechanisms, it is necessary for us to understand how much variance in performance can be attributed to different task demands, psychometric characteristics, individual differences, and actual differences in the underlying learning mechanisms at work. Attempting to assess any one of these possible sources of variance without an understanding of the others is a difficult proposition at best. Regardless, the results reported here constitute an important step into specifying the psychometric properties of SL measures, which is a necessary step in developing a better understanding of the role that sensitivity to statistical structure plays in both language learning and use.

Supplementary Files

The supplementary files for this article can be found as follows:

- **Supplementary File 1: Appendix.** http://dx.doi.org/10.1525/collabra.41.s1 The appendix describes a set of exploratory analyses performed on data collected in Experiment 2. In addition to completing the statistical learning tasks, participants also completed a battery of language and cognitive measures. These analyses explore potential relationships between the statistical learning measures and measures of cognition and language.
- **Supplementary File 2: Experiment 1 Data.** http://dx.doi.org/10.1525/collabra.41.s2
- **Supplementary File 3: Experiment 2 Data.** http://dx.doi.org/10.1525/collabra.41.s3

Acknowledgements

The work reported here and preparation of this manuscript was supported by the National Science Foundation, through a grant awarded to E.D.T. (BCS0642415); by the National Science Foundation through a grant awarded to L.C.E. (0946825); and by the Institute of Education Sciences (R305F100027). We thank four anonymous reviewers for their insightful comments on an earlier version of this manuscript.

Competing Interests

The authors declare that they have no competing interests.

Notes

[1] We acknowledge that uncertainty exists over how SL is related to other forms of implicit learning as indexed by a variety of distinct tasks across a diverse set of literatures (e.g., [40]). In what follows, we presume that SL and implicit learning are closely related constructs, and that individual variation in SL is akin to the individual variation in implicit learning that others have studied (e.g., [27]).

[2] Although one of the languages used in the present research is tonal rather than syllabic, for simplicity we refer to all the languages as word segmentation tasks.

[3] Tonal refers to sequences of tones rather than speech containing lexical tones.

[4] Accuracy with Language 4 was higher than accuracy with the other languages (around 80% vs. around 60%). Consequently, a ceiling effect due to diminished variability may play a role in the lack of significant correlations.

[5] The slight discrepancy between the descriptive statistics reported here and reported within **Table 1** for Language 2 2AFC Time 1 ($M = 0.56$, $SD = 0.23$) are because the mean and standard deviation here are only for the subset of participants who contributed data for both time points whereas Table 1 reports means and standard deviations for all the participants who contributed data to Time 1.

[6] We collected data on a range of cognitive measures (Digit Span, Operation Span, Reading Span, & Nelson Denny Vocabulary Size). Because these measures were

collected for exploratory purposes, and are something of an aside from the main research questions that are addressed here, we report the outcome of these measures in the Appendix.

[7] During the first time point, the artificial grammar learning task showed an order effect such that performance was higher when it was presented in the first position relative to when it was presented in later positions. One possible explanation for this is the artificial grammar learning task is more complicated than the other SL measures and thus more affected by factors such as fatigue.

References

1. Misyak, J. B., and Christiansen, M. H. 2012. Statistical learning and language: an individual differences study. Language Learning, *62*(1), 302–331. DOI: http://dx.doi.org/10.1111/j.1467-9922.2010.00626.x
2. Reber, A. S. 1993. Implicit learning and tacit knowledge: An essay on the cognitive unconscious. Oxford University Press.
3. Saffran, J. R., Aslin, R. N., and Newport, E. L. 1996. Statistical learning by 8-month-old infants. Science, 274, 1926–1928. DOI: http://dx.doi.org/10.1126/science.274.5294.1926
4. Krueger, L. E. 1986. Why 2 × 2 = 5 looks so wrong: On the odd-even rule in product verification. Memory and Cognition, 14, 141–149. DOI: http://dx.doi.org/10.3758/BF03198374
5. Franklin, D. W., and Wolpert, D. M. 2011. Computational mechanisms of sensorimotor control. Neuron, 72, 425–442. DOI: http://dx.doi.org/10.1016/j.neuron.2011.10.006
6. Jones, J. L., and Kaschak, M. P. 2012. Global statistical learning in a visual search task. Journal of Experimental Psychology: Human Perception and Performance, 38, 152–160. DOI: http://dx.doi.org/10.1037/a0026233
7. Erickson, L. C., and Thiessen, E. D. 2015. Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. Developmental Review. DOI: http://dx.doi.org/10.1016/j.dr.2015.05.002
8. Romberg, A., and Saffran, J. R. 2010. Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 906–914. DOI: http://dx.doi.org/10.1002/wcs.78
9. Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. 2013. The extraction and integration framework: A two-process account of statistical learning. Psychological Bulletin, 139(4), 792. DOI: http://dx.doi.org/10.1037/a0030801
10. Chang, F., Dell, G. S., and Bock, K. 2006. Becoming syntactic. Psychological Review, 113(2), 234. DOI: http://dx.doi.org/10.1037/0033-295X.113.2.234
11. MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. 1994. The lexical nature of syntactic ambiguity resolution. Psychological Review, 101(4), 676. DOI: http://dx.doi.org/10.1037/0033-295X.101.4.676
12. MacDonald, M. C., and Christiansen, M. H. 2002. Reassessing Working Memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). Psychological Review, 109(1), 35–54. DOI: http://dx.doi.org/10.1037/0033-295X.109.1.35
13. Maye, J., Werker, J. F., and Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82, B101–B111. DOI: http://dx.doi.org/10.1016/S0010-0277(01)00157-3
14. Thiessen, E. D., and Erickson, L. C. 2013. Beyond word segmentation: A two-process account of statistical learning. Current Directions in Psychological Science, 22(3), 239–243. DOI: http://dx.doi.org/10.1177/0963721413476035
15. Vouloumanos, A. 2008. Fine-grained sensitivity to statistical information in adult word learning. Cognition, 107(2), 729–742. DOI: http://dx.doi.org/10.1016/j.cognition.2007.08.007
16. Peperkamp, S., Le Calvez, R., Nadal, J. P., and Dupoux, E. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. Cognition, 101(3), B31–B41. DOI: http://dx.doi.org/10.1016/j.cognition.2005.10.006
17. Saffran, J. R., and Thiessen, E. D. (2003). Pattern induction by infant language learners. Developmental Psychology, 39, 484–494. DOI: http://dx.doi.org/10.1037/0012-1649.39.3.484
18. Thiessen, E. D., and Saffran, J. R. 2007. Learning to learn: Acquisition of stress-based strategies for word segmentation. Language Learning and Development, 3, 75–102.
19. Yu, C., and Smith, L. B. 2007. Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414–420. DOI: http://dx.doi.org/10.1111/j.1467-9280.2007.01915.x
20. Jaeger, T. F., and Snider, N. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. Cognition, 127, 57–83. DOI: http://dx.doi.org/10.1016/j.cognition.2012.10.013
21. Kaschak, M. P. 2007. Long-term structural priming affects subsequent patterns of language production. Memory and Cognition, 35, 925–937. DOI: http://dx.doi.org/10.3758/BF03193466
22. Coyle, J. M., and Kaschak, M. P. 2008. Patterns of experience with verbs affect long-term cumulative structural priming. Psychonomic Bulletin & Review, 15(5), 967–970. DOI: http://dx.doi.org/10.3758/PBR.15.5.967
23. Kaschak, M. P., and Glenberg, A. M. 2004. This construction needs learned. Journal of Experimental Psychology: General, 133(3), 450. DOI: http://dx.doi.org/10.1037/0096-3445.133.3.450
24. Thompson, S. P., and Newport, E. L. 2007. Statistical learning of syntax: The role of transitional probability. Language Learning and Development, 3(1), 1–42. DOI: http://dx.doi.org/10.1080/15475440709336999

25. Misyak, J. B., Christiansen, M. H., and Tomblin, J. B. 2010a. On-line individual differences in statistical learning predict language processing. Frontiers in Psychology. DOI: http://dx.doi.org/10.3389/fpsyg.2010.00031
26. Misyak, J. B., Christiansen, M. H., and Bruce Tomblin, J. 2010b. Sequential Expectations: The Role of Prediction-Based Learning in Language. Topics in Cognitive Science, 2(1), 138–153. DOI: http://dx.doi.org/10.1111/j.1756-8765.2009.01072.x
27. Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., and Mackintosh, N. 2010. Implicit learning as an ability. Cognition, 116(3), 321–340. DOI: http://dx.doi.org/10.1016/j.cognition.2010.05.011
28. Conway, C. M., Bauernschmidt, A., Huang, S. S., and Pisoni, D. B. 2010. Implicit statistical learning in language processing: Word predictability is the key. Cognition, 114(3), 356–371. DOI: http://dx.doi.org/10.1016/j.cognition.2009.10.009
29. Kidd, E. 2012. Implicit statistical learning is directly associated with the acquisition of syntax. Developmental Psychology, 48(1), 171. DOI: http://dx.doi.org/10.1037/a0025405
30. Arciuli, J., and Simpson, I. C. 2012a. Statistical learning is related to reading ability in children and adults. Cognitive Science, 36(2), 286–304. DOI: http://dx.doi.org/10.1111/j.1551-6709.2011.01200.x
31. Kaschak, M. P., Kutta, T. J., and Jones, J. L. 2011. Structural priming as implicit learning: Cumulative priming effects and individual differences. Psychonomic Bulletin & Review, 18(6), 1133–1139. DOI: http://dx.doi.org/10.3758/s13423-011-0157-y
32. Baldwin, D., Andersson, A., Saffran, J., and Meyer, M. 2008. Segmenting dynamic human action via statistical structure. Cognition, 106(3), 1382–1407. DOI: http://dx.doi.org/10.1016/j.cognition.2007.07.005
33. Fiser, J., and Aslin, R. N. 2002. Statistical learning of higher-order temporal structure from visual shape-sequences. Journal of Experimental Psychology: Learning, Memory, & Cognition, 28, 458–467. DOI: http://dx.doi.org/10.1037/0278-7393.28.3.458
34. Conway, C. M., and Christiansen, M. H. 2006. Statistical learning with and between modalities: Pitting abstract against stimulus specific representations. Psychological Science, 17, 905–912. DOI: http://dx.doi.org/10.1111/j.1467-9280.2006.01801.x
35. Sell, A. J., & Kaschak, M. P. (2009). Does visual speech information affect word segmentation? Memory & Cognition, 37(6), 889–894. DOI: http://dx.doi.org/10.3758/MC.37.6.889
36. Thiessen, E. D. 2010. Effects of visual information on adults' and infants' auditory statistical learning. Cognitive Science, 34(6), 1093–1106. DOI: http://dx.doi.org/10.1111/j.1551-6709.2010.01118.x
37. Thiessen, E. D., and Saffran, J. R. 2004. Spectral tilt as a cue to word segmentation in infancy and adulthood. Perception and Psychophysics, 66, 779–791. DOI: http://dx.doi.org/10.3758/BF03194972
38. Lim, S. J., Fiez, J. A., and Holt, L. L. 2014. How may the basal ganglia contribute to auditory categorization and speech perception? Frontiers in Neuroscience, 8. DOI: http://dx.doi.org/10.3389/fnins.2014.00230
39. Frost, R., Armstrong, B. C., Siegelman, N., and Christiansen, M. H. 2015. Domain generality vs. modality specificity: The paradox of statistical learning. Trends in Cognitive Sciences, 19, 117–125. DOI: http://dx.doi.org/10.1016/j.tics.2014.12.010
40. Perruchet, P., and Pacton, S. 2006. Implicit learning and statistical learning: One phenomenon, two approaches. Trends in Cognitive Sciences, 10(5), 233–238. DOI: http://dx.doi.org/10.1016/j.tics.2006.03.006
41. Gebauer, G. F., and Mackintosh, N. J. 2007. Psychometric intelligence dissociates implicit and explicit learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 34. DOI: http://dx.doi.org/10.1037/0278-7393.33.1.34
42. Siegelman, N., and Frost, R. 2015. Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. Journal of Memory and Language, 81, 105–120. DOI: http://dx.doi.org/10.1016/j.jml.2015.02.001
43. Conway, C. M., and Christiansen, M. H. 2009. Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. European Journal of Cognitive Psychology, 21(4), 561–580. DOI: http://dx.doi.org/10.1080/09541440802097951
44. Emberson, L. L., Conway, C. M., and Christiansen, M. H. 2011. Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. The Quarterly Journal of Experimental Psychology, 64(5), 1021–1040. DOI: http://dx.doi.org/10.1080/17470218.2010.538972
45. Amso, D., Davidson, M. C., Johnson, S. P., Glover, G., and Casey, B. J. 2005. Contributions of the hippocampus and the striatum to simple association and frequency-based learning. NeuroImage, 27(2), 291–298. DOI: http://dx.doi.org/10.1016/j.neuroimage.2005.02.035
46. Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., and Bavelier, D. 2013. The neural correlates of statistical learning in a word segmentation task: An fMRI study. Brain and language, 127(1), 46–54. DOI: http://dx.doi.org/10.1016/j.bandl.2012.11.007
47. Turk-Browne, N. B., Scholl, B. J., Chun, M. M., and Johnson, M. K. 2009. Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. Journal of Cognitive Neuroscience, 21(10), 1934–1945. DOI: http://dx.doi.org/10.1162/jocn.2009.21131
48. Thiessen, E. D., and Saffran, J. R. 2003. When cues collide: Use of statistical and stress cues to word boundaries by 7- and 9-month-old infants. Developmental Psychology, 39, 706–716. DOI: http://dx.doi.org/10.1037/0012-1649.39.4.706
49. Arciuli, J., and Simpson, I. C. 2012b. Statistical learning is lasting and consistent over time. Neuroscience

Letters, 517(2), 133–135. DOI: http://dx.doi.org/10.1016/j.neulet.2012.04.045
50. Kim, R., Seitz, A., Feenstra, H., and Shams, L. 2009. Testing assumptions of statistical learning: is it long-term and implicit? Neuroscience Letters, 461(2), 145–149. DOI: http://dx.doi.org/10.1016/j.neulet.2009.06.030
51. Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. 1999. Statistical learning of tone sequences by human infants and adults. Cognition, 70(1), 27–52. DOI: http://dx.doi.org/10.1016/S0010-0277(98)00075-4
52. Thiessen, E. D., Hill, E. A., and Saffran, J. R. 2005. Infant-directed speech facilitates word segmentation. Infancy, 7, 49–67. DOI: http://dx.doi.org/10.1207/s15327078in0701_5
53. Weiss, D. J., Gerfen, C., and Mitchel, A. D. 2009. Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? Language Learning and Development, 5(1), 30–49. DOI: http://dx.doi.org/10.1080/15475440802340101
54. Gebhart, A. L., Aslin, R. N., and Newport, E. L. 2009. Changing structures in midstream: Learning along the statistical garden path. Cognitive Science, 33(6), 1087–1116. DOI: http://dx.doi.org/10.1111/j.1551-6709.2009.01041.x
55. Erickson, L. C., and Thiessen, E. D. 2013 (October). Adaptation to a novel lexical stress pattern: Evidence for consistency across the lifespan. Poster presented at the 8th Biennial Meeting of the Cognitive Development Society, Memphis, TN.
56. Fernald, A. 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child development, 1497–1510. DOI: http://dx.doi.org/10.2307/1130938
57. Aslin, R. N., Saffran, J. R., and Newport, E. L. 1998. Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9, 321–324. DOI: http://dx.doi.org/10.1111/1467-9280.00063
58. Schneider, W., Eschman, A., and Zuccolotto, A. 2002. E-Prime: User's guide. Psychology Software Incorporated.
59. Kaschak, M. P., Kutta, T. J., and Schatschneider, C. 2011. Long-term cumulative structural priming persists for (at least) one week. Memory and Cognition, 39, 381–388. DOI: http://dx.doi.org/10.3758/s13421-010-0042-3
60. Kaschak, M. P., Kutta, T. J., and Coyle, J. M. 2014. Long and short term cumulative structural priming effects. Language, Cognition, and Neuroscience, 29, 728–743. DOI: http://dx.doi.org/10.1080/01690965.2011.641387
61. Kaschak, M. P., and Saffran, J. R. 2006. Idiomatic syntactic constructions and language learning. Cognitive Science, 30(1), 43–63. DOI: http://dx.doi.org/10.1207/s15516709cog0000_44
62. Jones, J. L., and Kaschak, M. P. 2009. Do idiomatic constructions always aid language learning? Language Learning and Development, 5, 69–93. DOI: http://dx.doi.org/10.1080/15475440802637100
63. Johnson, E. K., and Tyler, M. D. 2010. Testing the limits of statistical learning for word segmentation. Developmental Science, 13(2), 339–345. DOI: http://dx.doi.org/10.1111/j.1467-7687.2009.00886.x
64. QFAB Bioinformatics. 2015. ANZMTG Statistical Decision Tree, Power Calculator (Version 1.0 (Web Application)). Retrieved from http://www.anzmtg.org/stats/PowerCalculator.
65. Siegelman, N., Bogaerts, L., and Frost, R. in press. Measuring individual differences in statistical learning: Current pitfalls and possible solutions. Behavior Research Methods. DOI: http://dx.doi.org/10.3758/s13428-016-0719-z

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://dx.doi.org/10.1525/collabra.41.opr

How to cite this article: Erickson, L, C, Kaschak, M, P, Thiessen, E, D and Berry, C A S, 2016 Individual Differences in Statistical Learning: Conceptual and Measurement Issues. *Collabra,* 2(1): 14, pp. 1–17, DOI: http://dx.doi.org/10.1525/collabra.41

Submitted: 30 March 2016 **Accepted:** 31 August 2016 **Published:** 27 October 2016

UNIVERSITY of CALIFORNIA PRESS | Collabra

Collabra is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS