Behavior Research Methods (2020) 52:68–81
https://doi.org/10.3758/s13428-019-01205-5

Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality

Inbal Arnon[1]

Published online: 12 February 2019
© The Psychonomic Society, Inc. 2019

Abstract

Do commonly used statistical-learning tasks capture stable individual differences in children? Infants, children, and adults are capable of using statistical learning (SL) to extract information about their environment. Although most studies have looked at group-level performance, a growing literature examines individual differences in SL and their relation to language-learning outcomes: Individuals who are better at SL are expected to show better linguistic abilities. Accordingly, studies have shown positive correlations between SL performance and language outcomes in both children and adults. However, these studies have often used tasks designed to explore group-level performance without modifying them, resulting in psychometric shortcomings that impact reliability in adults (Siegelman, Bogaerts, Christiansen, & Frost in *Transactions of the Royal Society B, 372*, 20160059, 2017a; Siegelman, Bogaerts, & Frost in *Behavior Research Methods, 49,* 418–432, 2017b). Even though similar measures are used to assess individual differences in children, no study to date has examined the reliability of these measures in development. This study examined the reliability of common SL measures in both children and adults. It assessed the reliability of three SL tasks (two auditory and one visual) twice (two months apart) in adults and children (mean age 8 years). Although the tasks showed moderate reliability in adults, they did not capture stable individual variation in children. None of the tasks were reliable across sessions, and all showed internal consistency measures well below psychometric standards. These findings raise significant concerns about the use of current SL measures to predict and explain individual differences in development. The article ends with a discussion of possible explanations for the difference in reliability between children and adults.

Keywords Statistical learning · Individual differences · Reliability · Domain generality · Children

Infants, children, and adults are constantly exposed to recurring patterns in their environment and manage to learn and generalize from them. This ability—often called *statistical learning* (SL)—is postulated to be one of the important mechanisms in language learning and in learning more generally (e.g., Erickson & Thiessen, 2015). Statistical learning in infants was demonstrated in a seminal study showing that 8-month-old infants can use distributional information about syllable co-occurrence to discover word boundaries (Saffran, Aslin, & Newport, 1996). This study led to a surge in research probing this ability in both infants and adults. Research over the past 20 years has shown that statistical learning is present from early infancy (Bulf, Johnson, & Valenza, 2011), is found across modalities (visual, auditory, and tactile; Conway & Christiansen, 2005; Kirkham, Slemmer, & Johnson, 2002), and can be used to learn a range of linguistic properties (e.g., phonetic categories, word order, and phrase structure; see Romberg & Saffran, 2010, for a review). These studies, which assessed learning at the group level, serve as a "proof of concept": They show that humans are capable of using distributional information to extract complex structure from their environment.

Recent years have seen growing interest in the predictive relation between SL and individual differences in language learning (Siegelman, Bogaerts, & Frost, 2017b; Siegelman

Highlights There is growing evidence for positive correlations between SL performance and language outcomes in development. However, these findings are based on a small set of tasks that were developed for assessing group-level performance, and may not be suited to measure individual differences. No study to date has examined the reliability of these tasks in children, a crucial prerequisite for using them to assess individual differences. The present study assessed the reliability of three SL tasks across modalities (visual, linguistic auditory, and nonlinguistic auditory) in children (mean age 8;2) and adults. The tasks showed reliability in adults, but not in children, raising serious concerns about their use to predict and explain individual differences in development.

✉ Inbal Arnon
Inbal.arnon@mail.huji.ac.il

[1] Department of Psychology, Hebrew University, Jerusalem, Israel

& Frost, 2015). Variation in SL is predicted to correlate with language outcomes: Individuals who are better at SL are expected to also show better linguistic abilities. Similarly, difficulty with SL may contribute to impaired language performance in children with atypical language development (Mainela-Arnold & Evans, 2014). Indeed, a growing number of studies find correlations between SL measures and language outcomes. In adults, performance on visual SL has been related to better literacy outcomes in L2 (Frost, Siegelman, Narkiss, & Afek, 2013), syntactic processing (Misyak & Christiansen, 2012; Misyak, Christiansen, & Tomblin, 2010), and speech perception (Conway, Bauernschmidt, Huang, & Pisoni, 2010). Similar findings have been reported for children, for whom visual SL is predictive of syntactic processing (Kidd, 2012; Kidd & Arciuli, 2016), early literacy skills (Arciuli & Simpson, 2012), and vocabulary size (Spencer, Kaschak, Jones, & Lonigan, 2015). Work with infants also suggests links between SL performance and later learning, with visual SL predicting later vocabulary size (Ellis, Robledo Gonzalez, & Deák, 2014; Shafto, Conway, Field, & Houston, 2012) and auditory SL predicting real-time language processing (Lany, Shoaib, Thompson, & Graf Estes, 2018).

Taken together, these findings suggest a strong link between SL performance and variations in language outcomes. However, this interpretation is dependent on the reliability of the SL measures used. If SL measures are not reliable, it is hard to interpret their correlations with other measures: We cannot know how much of the observed interindividual variation reflects measurement error and how much reflects stable, and meaningful variation. Despite extensive use over the past 20 years, relatively little work has examined or validated the psychometric properties of current SL measures. When used to assess group-level performance, this question is less pressing: We are looking precisely for effects that generalize beyond individual differences in performance. However, reliability is a crucial issue when such tasks are used to estimate individual differences. A series of recent studies has raised doubts about the suitability of commonly used SL tasks for assessing individual differences in adults (Siegelman, Bogaerts, Christiansen, & Frost, 2017a; Siegelman, Bogaerts, & Frost, 2017b). These studies highlight the fact that many studies of individual differences of SL adopt the same tasks used to explore group-level performance without modifying them, resulting in psychometric shortcomings. Many adult studies employ a word segmentation task modeled on the original Saffran et al. (1996) task (often called ASL), or a visual parallel that uses novel shapes instead of syllables (VSL, modeled on Turk-Browne, Jungé, & Scholl, 2005). In these tasks, participants have to detect recurring triplets (visual or auditory) in a continuous stream based on their statistical properties (see Table 1 in Siegelman, Bogaerts, & Frost, 2017b, for a summary of methods used in recent studies).

Although the implementation of these tasks differs across studies, they share several properties that undermine their suitability as a measure of individual differences (Siegelman, Bogaerts, Christiansen, & Frost, 2017a; Siegelman, Bogaerts, & Frost, 2017b). In particular, they have relatively few testing trials, all at the same level of difficulty; they repeat items during testing; they use two-alternative forced choice trials to assess learning after exposure; and show accuracy that is usually not much higher than chance, meaning that many participants perform at chance level. These characteristics reduce task reliability and may lead to the detection of spurious correlations on the one hand, and the underdetection of true correlations on the other (see Siegelman, Bogaerts, & Frost, 2017b, for simulations demonstrating this). This literature outlines several ways that the reliability of SL measures can be improved, for example by increasing the number and type of test trials (Siegelman, Bogaerts, Christiansen, & Frost, 2017), making the testing more implicit (Isbilen, McCauley, Kidd, & Christiansen, 2017), or using online measures of learning during exposure (Siegelman, Bogaerts, Kronenfeld, & Frost, 2018a). Despite these shortcomings, commonly used SL measures (the ASL and VSL discussed above) seem to capture stable individual variation in adults: They show test–retest reliability between sessions and internal consistency within sessions, though these measures are often below psychometric standards (test–retest of around .5: Siegelman & Frost, 2015; Potter, Wang, & Saffran, 2017; split-half reliability of around .8: Siegelman, Bogaerts, & Frost, 2017b).

Even though very similar measures are used to assess individual differences in children, and even though child data is often noisier and less stable, no study to date has examined their reliability in development. We do not know if commonly used SL tasks capture stable variation in children, a gap that limits our ability to interpret their reported correlations with language outcomes during childhood.[1] Moreover, there are several reasons to suspect that children will be more affected by the shortcomings of these tasks, leading to less stable performance and lower reliability. For starters, children show lower accuracy than adults (Raviv & Arnon, 2017), meaning that even more of them are at chance (making their scores less informative for predicting variation in other outcomes). The way that learning is assessed poses additional challenges: Children have a harder time with explicit judgments like the ones used in these tasks (in which they have to choose between two forms), and performance on them is often not a good indicator of their knowledge (see the current debate in the theory-of-mind literature; e.g., Southgate, Senju, &

[1] The tasks used with infants have similar properties but differ in that learning is assessed implicitly, a point that we will return to in the Discussion section. We focus here on tasks that assess learning in the same way in children and adults.

Table 1 Means and ranges for all tasks in both sessions for adults (including *SD*s in brackets and comparison to chance level)

Task	First session accuracy	Range	Second session accuracy	Range
Linguistic ASL	71% (13)***	40%–100%	74% (14)***	40%–100%
Nonlinguistic ASL	71% (15)***	44%–100%	69% (19)***	24%–100%
Visual SL	87% (15)***	44%–100%	88% (18)***	32%–100%
Working memory	10.2 (1.78)	6–14	10.4 (1.54)	8–13

*** Significantly above chance, $p < .001$

Csibra, 2007). This difficulty may be compounded by the repetition of test items and foils (increasing confusabiltiy), and by the relatively small number of test trials (so that fluctuations in attention can have a big effect on the overall score). Finally, the range of SL tasks in the child individual difference literature is even more limited than in the adult literature: Four of the five published studies with children (Arciuli & Simpson, 2012; Kidd, 2012; Kidd & Arciuli, 2016; Mainela-Arnold & Evans, 2014; Spencer et al., 2015) use either the ASL or the VSL with similar exposure and testing properties (four to six triplets, 36–64 testing trials, learning assessing learning using two-alternative forced-choice trials with repetition of test items and foils during testing. That is, our knowledge about the relation between SL and language outcomes in children is based on findings from a small set of tasks that are not ideally suited to measure individual differences.

In line with these concerns, the correlations between SL and language outcomes during development are small, and weaker than those found for adults. Although the correlations in adults tend to be moderate (r = .4–.6), those of children are lower, suggesting a weaker relationship between the variance captured by SL measures and language outcomes (ranging between r = .1 and .34, see Table 1 in Siegelman, Bogaerts, Christiansen, & Frost, 2017a). Moreover, the pattern of correlations is not consistent even when using the same language measures. For instance, Kidd (2012) and Kidd and Arciuli (2016) found no correlation between SL and vocabulary measures in children whereas Spencer et al. (2015) did. The low correlations and the fluctuations between studies may both stem from the tasks not having sufficient reliability.

The present study was designed to examine the reliability of commonly used SL measures in children and adults. This was done by assessing the reliability of three SL tasks (two auditory and one visual) that are closely modeled on ones used in the child individual-difference literature. The first study looked at adults, with the aim of providing a reliability baseline for the precise measures that would then be tested with children: It is theoretically possible that child-friendly tasks would also have lower reliability in adults, because of their shorter exposure and reduced number of items. The second study examined the reliability of the same three tasks in children aged seven to nine years (mean age 8 years 2 months [8;2]) in two different samples. In the first sample, the reliability of the three SL tasks was assessed in the same children. The second sample provided additional reliability estimates for a modified version of the linguistic auditory task (ASL), which is assumed to be most related to language outcomes. I focused on this age group because children of this age have shown learning at the group level (Raviv & Arnon, 2017), and their performance has been reported to be correlated with language outcomes (e.g., Kidd & Arciuli, 2016). Reliability was evaluated in two ways: (1) by looking at the internal reliability and consistency of tasks within each session (using split-half reliability and Cronbach's alpha coefficients) and (2) by examining their test–retest reliability two months apart. The study also assessed verbal working memory using an auditory digit span task in both sessions. Collecting this measure served as a sanity check, because it is known to be stable within individual children and is therefore expected to show high test–retest reliability (Gathercole, Willis, Baddeley, & Emslie, 1994). If SL tasks tap onto stable individual differences in children, they should be correlated across the two test sessions. If they do not, this would raise significant concerns about their use to predict and explain individual differences in development.

The SL tasks were closely modeled on those previously used to examine individual differences in children (e.g., Kidd, 2012; Kidd & Arciuli, 2016; Spencer et al., 2015). The tasks used were (1) a linguistic auditory task, in which participants were exposed to recurring triplets of syllables; (2) a nonlinguistic auditory task, in which the syllables were replaced with familiar sounds (dog barking, bell, drum), using stimuli from Siegelman, Bogaerts, Arciuli, Elazar, and Frost (2018b); and (3) a visual task, in which participants saw recurring triplets of familiar object drawings (e.g., car, door, plate). This task resembled the child-friendly visual SL task developed by Arciuli and Simpson (2011). All three tasks required learners to detect recurring triplets in a continuous temporal input. Although the tasks differed in the stimuli used (syllables vs. nonlinguistic sounds vs. drawings), they were comparable in terms of the distributional information learners were exposed to and in the number and nature of the test trials.

General method

Materials

Since the same SL tasks will be used with children and adults, they are described here, before the results of the two studies are reported. In all three SL tasks, participants were exposed to a continuous stream made up of five recurring triplets. The transitional probabilities (TPs) between elements within a triplet were always 1, whereas the TPs between triplets were .25 (because elements were not repeated across triplets and because each triplet could be followed by any of the other four). Following exposure, participants' knowledge of the triplets was assessed using 25 two-alternative forced choice trials. The stimuli are described for each task separately, and then in the next section the identical testing phase is described (see the Appendix for the full stimulus list). Verbal working memory was assessed using the forward digit span task, in which participants have to recall lists of numbers growing in size, from three to nine (Kaufman, 1994). The test is discontinued if the participant fails on two consecutive trials. Participants' scores represented the number of sequences correctly recalled (with the maximal score being 14). The procedure was identical to that used in Kaufman (1994), with one change: we only used the digits 1-5 with children, to prevent arithmetic ability from affecting performance (Havron & Arnon, 2017).

Linguistic auditory task The auditory stimuli consisted of a synthesized "alien" language, containing five unique trisyllabic words (*gedino*, *dukame*, *kimuga*, *nalobi*, *tobelu*), made up of 15 different syllables (taken from Glicksohn & Cohen, 2013; see the Appendix for all items). Syllables were created using the PRAAT synthesizer (Boersma & van Heuven, 2001) and were matched on pitch (~ 76 Hz), volume (~ 60 dB), and duration (250–350 ms). The words were created by concatenating the syllables using MATLAB, to ensure that there were no co-articulation cues to word boundaries. The words were matched for length (average word length 860 ms, range = 845–888 ms). Words were concatenated together in a semirandomized order (with the constraint that no word would appear twice in a row) to create an auditory familiarization stream. The exposure phase lasted 2:20 min, with each word repeated 32 times with no breaks between words and no prosodic or co-articulation cues in the stream to indicate word boundaries. All participants were exposed to the same five triplets: Using the same familiarization stream is a common feature of ASL tasks (starting from Saffran et al., 1996, and repeated across many studies). Here the stream was used in both sessions because group performance on this task seems to be affected by the exact syllable combinations in ways that are not fully understood (Erickson, Kaschak, Thiessen, & Berry, 2016; Siegelman et al., 2018b). The concern was that changing the triplets would introduce variation that could not be predicted (I will return to the possible implications of this in the follow-up study).

Nonlinguistic auditory task This task was very similar to the linguistic task except that syllables were replaced with familiar nonlinguistic sounds (e.g., bell, dog barking). The auditory stimuli contained five unique triplets made up of 15 different sounds. Unlike the linguistic task, triplets were generated anew for each participant, so that each participant heard a different set of triplets. Triplets were changed between sessions so that participants did not hear the same triplets in both. Average sound length was 500 ms (range 450–550 ms). Triplets were concatenated in a semi-randomized order, with the constraint that no triplet would appear twice in a row. The exposure phase lasted 3 min, with each triplet repeated 24 times.

Visual task This task had similar properties to the previous two, but in the visual domain. The visual stimuli consisted of a continuous temporal stream of black and white drawing of familiar objects (e.g., plane, door), containing five unique triplets of drawings (a total of 15 different drawings). Line drawings were selected from the Snodgrass and Vanderwart (1980) database that had high naming agreement based on Alario and Ferrand (1999). All names had high frequency in Hebrew and early age of acquisition (Maital, Dromi, Sagi, & Bornstein, 2000). All images were equally sized and were presented in the center of the screen. Each drawing appeared on the screen for 500 ms, with a 100-ms break between figures—resulting in a 1,800-ms presentation time for each triplet. The triplets were generated anew for each participant and changed between sessions. For each participant, the five triplets were concatenated together in a semirandomized order (with the constraint that no triplet would appear twice in a row). The exposure phase lasted 3:30 min, with each triplet repeated 24 times.

The test phase

The test phase was the same for the three tasks and included 25 two-alternative forced choice trials in which participants had to choose between two triplets (separated by 500 ms). On each trial, participants heard a real triplet (that had appeared in the exposure stream) either followed or preceded by a foil triplet (the order was counterbalanced so that on half of the trials the real triplets appeared first). Foil triplets were constructed by taking the first syllable/sound/drawing from one triplet, followed by the second syllable/sound/drawing from another triplet, and the third syllable/sound/drawing from a third triplet. Thus, each element in the foil triplets appeared in a similar position in real triplets, but with different surrounding syllables (e.g., "kilome" or "dubega"). This created a difference in the statistical properties of the real triplets and the foils: Whereas the TPs between every two adjacent elements within a real triplet were 1, the TPs between every two syllables in a foil test item were 0, because

participants had never heard these elements one after the other during familiarization. If participants learned the statistical properties of the stream, they should be able to distinguish between real triplets and foils. Scores on each task could range from 0% accuracy (0/25 trials correct) to 100% (25/25 trials correct). Trials were presented in a random order, with the constraint that the same triplet/foil did not appear in two consecutive trials

Study 1: The reliability of child-friendly SL tasks in adults

Method

Participants A total of 52 adults participated in both testing sessions (mean age 23 years; 33 females and 19 males). All were university students and received payment for participation. All were native Hebrew speakers and none had any learning, hearing, or language impairments.

Procedure Participants were tested in a quiet room in the lab while seated in front of a computer. They completed all four tasks (three SL tasks and working memory tasks) in both sessions. Task order was semirandomized. There were three possible orders for the SL tasks (linguistic auditory, visual, nonlinguistic auditory; visual, nonlinguistic auditory, linguistic auditory; and nonlinguistic auditory, linguistic auditory, visual). Verbal working memory was tested in between the SL tasks. Participants were tested in different orders in each session. Each session took 30 min to complete. In the two auditory tasks, participants were told they will be learning a novel alien language (linguistic auditory) or song (nonlinguistic auditory). Following exposure, participants were asked to say which of two word/sound sequences was more like the language/song they had just heard. In the visual task, they were told that they were about to see objects that aliens are taking back to their country. Following exposure, participants were asked to help the spaceship commander remember which objects were taken into the spaceship together. After hearing/seeing both possibilities, participants were asked to press either "1" or "2," depending on whether they thought the correct triplet was the first or the second they had heard. In cases in which participants felt they didn't know the answer, they were encouraged to guess.

Results

Performance on SL tasks in the two sessions Task order did not have a significant effect on performance in any of the SL tasks in both sessions, so results from the three orders were collapsed (p's > .4 for all tasks, obtained using mixed-effect regression models predicting trial accuracy from order for each task separately). Gender also did not affect performance (ps > .3 for all tasks). Adults showed learning in all tasks (see Table 1). Performance did not improve significantly for any of the tasks (ps > .2), as has been previously found (Siegelman & Frost, 2015). Performance was better on the visual task than on the two auditory ones [visual vs. linguistic auditory: Session 1 $t(51) = 5.21$, $p < .001$, Session 2 $t(51) = 4.19$, $p < .001$; visual vs. nonlinguistic auditory: Session 1 $t(51) = 6.77$, $p < .001$, Session 2 $t(51) = 6.4$, $p < .001$], in line with previous findings with (Siegelman & Frost, 2015). The ranges and SDs for all tasks were similar in both sessions, indicating that there wasn't a change in the distribution of performance.

Assessing the reliability of the SL tasks The reliability of SL was evaluated by looking at the internal reliability and consistency of tasks within each session and by examining their test–retest reliability across the two sessions. To examine the internal reliability and consistency, the split-half and Cronbach's alpha coefficients were calculated for each task in each session (using the psychometric::alpha function in R): Both measures give an indication of how well each item predicts overall performance and of whether performance on different items is correlated, as would be expected if it reflects learning of the statistical structure of the input. To evaluate the test–retest reliability of the SL tasks, I looked at the correlation of performance on the same task in the two sessions. Table 2 shows the three reliability measures for all tasks in both sessions for adults.

Importantly, all three tasks showed moderate reliability, though it was generally lower than advised psychometric standards. The measures were also not entirely consistent: The linguistic auditory task had the lowest internal reliability (Cronbach's alpha of .57–.63, below the advised values of .8–.95; see Streiner, 2003), but the visual one had the lowest between session reliability (.45, below the advised .7 value). Although all three test–retest correlations were significant, their values were lower than what is expected from tools assessing individual traits (Ellis et al., 2014; Nunnally & Bernstein, 1994). Unsurprisingly, the estimates are somewhat lower than found for similar SL tasks that were modified to better capture individual differences by increasing the test trials and varying their difficulty and kind (Siegelman, Bogaerts, & Frost, 2017b). These results show that our tasks have moderate reliability in adults while highlighting their limitation as a tool for assessing individual differences (even in adults).

Correlations between the tasks Table 3 shows the correlation between the three SL tasks in the two sessions. The pattern of correlations was stable across sessions, another indication of task stability: The two auditory tasks were not correlated, whereas visual task and the nonlinguistic auditory task were. This pattern of correlation may seem counterintuitive because the tasks do not group together on the basis of modality; however, it is compatible with a set of recent findings suggesting that tasks using linguistic stimuli (like syllables) behave differently from tasks using nonlinguistic stimuli, regardless of

Table 2 Internal consistency and reliability measures in adults for all tasks in both sessions (with 95% CIs in brackets)

	Linguistic auditory		Nonlinguistic auditory		Visual	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Internal consistency	.57 [.42, .71]	.63 [.50, .74]	.68 [.57, .78]	.79 [.72, .86]	.83 [.76, .88]	.91 [.88, .94]
Split-half reliability	.43 [.18, .63]	.54 [.31, .71]	.69 [.52, .81]	.62 [.41, .76]	.72 [.55, .83]	.82 [.70, .89]
Test–retest	$r = .61^{***}$ [.41, .76]		$r = .70^{***}$ [.53, .82]		$r = .45^{***}$ [.20, .64]	

$^{***}p < .001$

modality (Siegelman et al., 2018a; Shufaniya & Arnon, 2018). An additional stable pattern was the positive correlation between the SL tasks and working memory, a pattern that has been reported in some adult studies (Misyak & Christiansen, 2012) but not others (Siegelman & Frost, 2015). We will return to both findings in the Discussion section.

Study 2a: Reliability of SL measures in children

Method

Participants In all, 44 children participated in the first testing session. Three children were absent from school during the second testing session, so the final sample comprised of 41 children (22 girls, 19 boys). Children were in second or third grade (mean age 8;2, range 7;2–9;0); all were native Hebrew speakers and none had known learning, hearing or language impairments. Parental consent was obtained for all participating children. Children received a small educational reward.

Procedure Children were tested in a quiet room in their school, while wearing noise-cancelling headphones. The procedure was identical to that of adults with the only difference being that an experimenter sat next to the child and read out the instructions to them. Children completed all four tasks (three SL tasks and the auditory digit span) in both sessions, each child was randomly assigned to one of the three possible task orders. Verbal working memory was assessed in between the SL tasks. Children were tested in different orders in each session. Each session took 30 min to complete. At the start of each session, children were told they will be playing a few games with the experimenter and can stop at any time. During the linguistic auditory task, children were told that they were about to hear an alien language. Following exposure, children were told that they were about to hear an alien who is not a good speaker of the alien language, and that they must help him by telling him which of the two words he says sounds more like the alien language they just heard. In the nonlinguistic auditory task, children were told to learn an alien song and then say which sound sequence is more like the alien song they just heard. In the visual task, children were told that they were about to see objects that aliens are taking back to their country. Following exposure, children were asked to help the spaceship commander remember which objects were taken into the spaceship together. After hearing/seeing both possibilities, children were asked to press either "1" or "2" according to whether they thought the correct triplet was the first or the second they heard. In cases in which children felt they didn't know the answer, the experimenter encouraged them to try and guess.

Results

Performance on SL tasks in the two sessions Table 4 shows group-level performance on all tasks. Task order did not have a significant effect on performance in any of the SL tasks in both sessions, so results from the three orders were collapsed ($ps > .2$). Gender also did not affect performance ($ps > .6$). Children showed learning across modalities and stimulus types, though accuracy was lower than for adults. Children were significantly above chance for all tasks in the first session and for the linguistic–auditory and visual task in the second session [see Table 4; first session: visual, $t(40) = 7.48$, $p <$

Table 3 Simple bivariate Pearson correlations in adults between the different SL tasks in both sessions (significant correlations in bold)

	First session			Second session		
	Nonlinguistic	Visual	WM	Nonlinguistic	Visual	WM
Linguistic	.07	– .09	**.28***	.20	– .09	.23
Nonlinguistic		**.41****	**.32***		**.39****	**.26***
Visual			**.38****			**.31***

$^{*}p < .05$, $^{**}p < .01$

Table 4 Means and ranges for all tasks in children (including comparison to chance performance) in both sessions

Task	First session accuracy	Range	Second session accuracy	Range
Linguistic ASL	57% (10)**	40%–84%	63% (10)***	40%–84%
Nonlinguistic ASL	59% (13)**	36%–92%	52% (11)	32%–80%
Visual SL	69% (15)***	36%–92%	69% (16)***	32%–96%
Working memory	6.5 (1.39)	4–11	7.1 (1.13)	5–10

** Significantly above chance, $p < .01$, *** Significantly above chance, $p < .001$

.001; nonlinguistic auditory, $t(40) = 4.43$, $p < .001$; linguistic auditory, $t(40) = 4.71$, $p < .001$; second session: visual, $t(40) = 7.24$, $p < .001$; nonlinguistic auditory, $t(40) = 1.44$, $p = .15$; linguistic auditory, $t(40) = 7.33$, $p < .001$]. Only the linguistic auditory task improved between sessions [$t(40) = 2.74$, $p < .01$]. The ranges and *SD*s for all tasks were similar in both sessions, indicating that there wasn't a change in the distribution of performance.

In line with prior findings (Raviv & Arnon, 2017), children's accuracy was higher on the visual task than on the two auditory tasks in both sessions [Session 1: visual vs. linguistic, $t(40) = 3.53$, $p < .001$, visual vs. nonlinguistic, $t(40) = 3.20$, $p < .01$; Session 2: visual vs. linguistic auditory, $t(40) = 2.26$, $p < .05$; visual vs. nonlinguistic auditory, $t(40) = 6.35$, $p < .001$]. The relation between the linguistic and nonlinguistic auditory tasks changed between sessions: In the first session, performance did not differ on the two tasks [$t(40) = -0.71$, $p > .4$] whereas in the second session performance was better on the linguistic task than on the nonlinguistic one [$t(40) = 3.85$, $p < .001$]. In sum, children showed learning on all tasks (though lower than adults); they were better in the visual task than the auditory one; and did not perform consistently better on one of the auditory tasks.

Assessing the reliability of the SL tasks Turning to our main research question, the reliability of the SL tasks was evaluated using the same reliability measures used with adults (Cronbach's alpha coefficients and split-half reliability within session and test–retest reliability across sessions). Table 5 shows the three reliability measures for all tasks in both sessions.

There are several important patterns to notice. First, in contrast with the adult data, these tasks are not reliable in children. All reliability measures were well below psychometric standards: Both the Cronbach's alpha coefficients and the split half-reliability varied greatly between tasks (from – .04 for the linguistic auditory task to .68 for the visual one) and were below recommended values for standard tests (Streiner, 2003). Test–retest reliability was also lower than had been found for adults, and well below psychometric norms. In fact, only the linguistic auditory task showed any correlation between sessions, with a low correlation of $r = .33$ (meaning that only 32% of the variance was shared between test and retest). Importantly, this task differed from the other two in that triplets were repeated between sessions. Even though participants were learning the same "words," the correlation in performance between sessions was low (we will return to this point below). Crucially, verbal working memory did show the expected test–retest reliability between the two sessions ($r = .67$, $p < .001$), which was similar in magnitude to that reported in the literature for this age group (Gathercole et al., 1994). The second striking pattern is the lack of consistency on the three measures. The auditory linguistic task, which showed the highest test–retest reliability, had the lowest values on the two other measures. The visual task, which showed the highest internal consistency, had no test–retest reliability.

To further explore these findings, an aggregated SL score was calculated for each participant based on the average performance on the three tasks. Taking an aggregated measure may increase the stability of the measure. When using this aggregated measure, the correlation between sessions was still low ($r = .31$, $p = .046$), and in fact was lower than the correlation between the two linguistic auditory tasks, indicating that the other two tasks were not contributing much to the correlation. Moreover, the empirical and theoretical motivation for using an aggregate score is weak, given that performance on the different tasks was not systematically correlated within an individual (see Table 6). I calculated how many children were above chance on each of

Table 5 Internal consistency and reliability measures in children for all tasks in both sessions (with 95% CIs in brackets)

	Linguistic ASL		Nonlinguistic ASL		Visual	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Internal consistency	.05 [– .33, .37]	.27 [.01, .52]	.43 [.20, .62]	.2 [– .12, .46]	.68 [.56, .79]	.72 [.62, .82]
Split-half reliability	– .04 [– .34, .26]	.22 [– .08, 0.5]	.37 [.07, .60]	.08 [– .22, .38]	.59 [.36, .76]	.46 [.18, .67]
Test–rest reliability	$r = .33$, $p = .035$ [.02, .57]		$r = .24$, $p = .1$ [.06, .51]		$r = .01$, $p > .9$ [– .29, .31]	

Table 6 Simple bivariate Pearson correlations in children between the different SL tasks in both sessions, with 95% CIs in brackets

	First session			Second session		
	Nonlinguistic	Visual	WM	Nonlinguistic	Visual	WM
Linguistic	**.38***	– .14	.04	– .20	.28	.07
Nonlinguistic		.12	– .11		**.33***	.03
Visual			– .014			– .007

* Significantly above chance, $p < .05$

the tasks. Even when children as a group were above chance, many individual learners may be performing at chance level, making it difficult to interpret any correlation between their scores. Following Siegelman, Bogaerts, and Frost (2017b), the binomial distribution was used to determine chance level for an individual learner. Since there were 25 test trials in each task, the threshold was set to 17 or more correct responses ($p < .05$). Very few individuals were above chance in the same task in both sessions (another indication of low reliability): only five children for the linguistic auditory task, one for the nonlinguistic auditory task, and 14 for the visual one. The examination of individual patterns highlights another problematic aspect of these tasks (which has been previously noted by Siegelman, Bogaerts, Christiansen, & Frost, 2017a): The relatively low performance rates mean that many participants are at chance, making their scores uninformative for predicting variation in other measures. We will return to this point in the Discussion.

Correlations between the tasks Table 6 shows the correlations between the three SL tasks in the two sessions. In contrast with the adult findings, and in line with their low reliability, the pattern of correlations between SL tasks was not stable across the two sessions (cf. Table 3). In the first session, the two auditory tasks were correlated, but neither correlated with the visual task. However, an opposite pattern was observed in the second session: Now the two auditory tasks were correlated with the visual task, but not with each other. The lack of stable correlation pattern makes it hard to interpret the relation between the three tasks.

The relation between the SL tasks and working memory was stable across sessions, and different from what was found with adults: None of the SL tasks were correlated with verbal working memory in both sessions, replicating previous findings with children (Kidd, 2012; Kidd & Arciuli, 2016). Importantly, the lack of reliability makes it hard to interpret any of these correlations in a meaningful way.

Study 2b: Reliability of a linguistic–auditory task without repeated triplets in children

Our results indicate that all three tasks have low to nonexistent reliability in children. They also suggest that the linguistic auditory task has slightly better reliability: It was the only one to show any correlation between the two sessions. Although this could be related to the linguistic nature of the stimuli (and hence have interesting theoretical implications), it is more likely that it reflects a methodological difference between the tasks. The linguistic task was the only one in which triplets were repeated between sessions, so participants were tested on the same language in both sessions. It was also the only task in which accuracy was higher in the second session. The improved accuracy and reliability could have been driven by children's memory of specific triplets rather than by an increased stability of the measure. To investigate this, an additional study was run (with a new sample of children) looking only at the linguistic auditory task, but this time using different triplets in each session. If triplet repetition was responsible for the correlation between sessions, it should not be found with this modified task. Children also completed the digit span in both sessions since this is expected to show reliability between sessions.

Method

Participants In all, 38 children participated in the first testing session. Two children were absent from school during the second testing session, so the final sample comprised 36 children (21 girls, 15 boys). Children were in second or third grade (mean age 7;9, range 7;2–8;9); all were native Hebrew speakers, and none had known learning, hearing or language impairments. Parental consent was obtained for all participating children. Children received a small educational reward.

Materials: Linguistic auditory task The auditory stimuli were created in the exact same way as in the previous two study, but now there were two different "languages," each made up of five unique trisyllabic words. Language 1 was identical to the one in the previous study (*gedino*, *dukame*, *kimuga*, *nalobi*, *tobelu*; language2). Language 2 was created from the same inventory of consonants and vowels, but without reusing any of the syllables used in Language 1 (*bakomi*, *detula*, *golike*, *mudano*, *tinebu*). All syllables were created using the PRAAT synthesizer (Boersma & van Heuven, 2001) and were matched on pitch (~ 76 Hz), volume (~ 60 dB), and duration (250–350 ms). The words were created by concatenating the

syllables using MATLAB and were matched for length. For both languages, words were concatenated together in a semi-randomized order (with the constraint that no word would appear twice in a row) to create an auditory familiarization stream. The exposure phase lasted 2:20 min, with each word repeated 32 times. To control for possible differences between the languages, which language participants heard first was counterbalanced. Half of them heard Language 1 in the first session and Language 2 in the second session, and the second half heard Language 2 first and Language 1 second. Importantly, each participant was exposed to a different language in the two sessions.

Procedure Identical to the previous study. Children were tested in a quiet room in their school. They were tested twice, two months apart. All children completed the linguistic auditory task followed by the digit span. The study took around 10 min to complete.

Results

Performance in the two sessions Children showed learning in both sessions [first session: 59%, range 40%–76%, $t(35) = 6.28$, $p < .001$; second session: 57%, range 32%–72%, $t(35) = 4.40$, $p < .001$], with accuracy rates similar to the ones found in the previous study. Performance was similar for the two languages, suggesting both were equally easy to learn [60% vs. 56%, $t(35) = -1.8$, $p > .07$]. Importantly, accuracy did not improve between sessions [59% vs. 57%, $t(35) = 1.02$, $p > .3$], indicating that triplet repetition had driven the improvement in the previous study.

Assessing reliability Reliability was evaluated using the same measures used before. As predicted, when triplets were not repeated, the correlation between sessions was no longer significant ($r = -.15$ [−.46, .17], $p > .3$), indicating that the low correlation previously found did not reflect increased stability. The task was not reliable on any of the other measures, as well, and in fact showed reliability that was even lower than in the previous study (split-half: first session = .08 [−.23, .39], second session = .04 [−.29, .36]; Cronbach's alpha: first session = −.3 [−.8, .15], second session = −.08 [−.54, .29]).[2] Crucially, working memory did show the expected test–retest reliability between the two sessions ($r = .41$, $p < .01$). These results confirm that the task is not reliable in children and further undermine its use as a measure of individual differences.

[2] Note that the Cronbach's alpha for this modified task is negative. This can happen when the mean correlation between items is negative, which can happen if success on one item predicts failure on another. This is another worrying indicator for the lack of internal consistency in this task. It also raises a more general concern about what the test measures: such a negative correlation is highly unexpected if performance reflects knowledge of the transitional probabilities (which is the same for all triplets).

Discussion

This study set out to assess the reliability of three SL tasks modeled on ones previously used in the child individual difference literature. Although the use of such tasks to predict individual differences in language outcomes during development is growing, no study to date has shown that they measure a stable property within a child. This is crucial, in light of previous reports that such tasks have psychometric shortcomings as measures of individual differences, even in adults (Siegelman, Bogaerts, Christiansen, & Frost, 2017a; Siegelman, Bogaerts, & Frost, 2017b), and the possibility that these shortcomings are more pronounced in children. To examine this, both children and adults were tested on two auditory SL tasks and a visual one two months apart. I also assessed verbal working memory in both sessions to make sure that it shows the expected reliability. Two types of reliability were of concern: within session (assessed using split-half and Cronbach's alpha coefficients) and between sessions (using test–retest reliability).

The first study showed that our three SL tasks had moderate reliability in adults. These findings replicate previous investigations of task reliability in adult learners in two ways: by indicating that SL measures developed to assess group-level performance can capture stable variation in adults (Potter et al., 2017; Siegelman & Frost, 2015), and by illustrating their shortcomings: Reliability was lower than psychometric norms, and lower than had been found when using SL tasks modified to increase their psychometric validity (Siegelmen, Bogaerts, & Frost, 2017). The second study looked at the reliability of the same tasks in children and found a strikingly different pattern: Whereas children showed learning as a group on all tasks, all reliability measures were well below accepted norms, raising concern about their use as measures of individual differences in development. The lack of stability was also reflected in the relation between the different tasks. Unlike in the adult data, no stable correlations emerged between the different SL tasks: In the first session, the two auditory tasks were correlated with each other (but not with the visual task), suggesting a modality-based division. In the second session, however, the only correlation was across modality (between the nonlinguistic and the visual task). These findings indicate that commonly used SL tasks are not reliable in children, across modality, and stimulus type.

The only task that showed any test–retest correlation was the linguistic auditory task, which is based on Saffran et al. (1996) and has been used in numerous articles since. This was also the only task in which learners were exposed to the same triplets in both sessions: The slightly higher reliability could have reflected this methodological difference rather than an improved reliability for linguistic stimuli (which would have practical and theoretical implications). To further investigate this point, additional data were collected from a new sample of children (at the same ages), on a linguistic auditory task without repeating triplets. As suspected, when triplets were not repeated, the linguistic

auditory task was also not reliable (no correlation between sessions). The previously found correlation was driven by the repetition of triplets and is not a reflection of the better reliability of this task. Moreover, the internal consistency for this task was very low, which is worrying given its widespread use as a measure of SL. In contrast, verbal working memory showed the expected high reliability across sessions, in both child samples, suggesting the lack of reliability for the SL tasks is not related to our sample, but reflects something more meaningful (and problematic) about the SL tasks themselves. This low reliability is consistent with several patterns in the existing developmental data. The correlations tend to be low (ranging between 0.1 and 0.32, see (Siegelman, Bogaerts, Christiansen, & Frost, 2017a) and vary quite a bit even when using the same language measures. For instance, Kidd (2012) and Kidd and Arciuli (2016) found no correlation between SL and vocabulary measures in children, whereas Spencer et al. (2015) did (though small ones). Importantly, none of the previous child individual difference studies assessed the reliability of the tasks used.

The results here indicate that the SL tasks examined cannot be used as a reliable measure of individual differences in children. Since these tasks share important psychometric properties with ones previously used in the developmental literature, they raise a more general concern about existing findings on the relation between SL and language outcomes. Low reliability could lead both to the detection of correlations that are not really there (spurious correlations) and to underestimating true correlations (due to measurement error), making it hard to draw strong conclusions from existing reports of correlations between SL measures and linguistic outcomes. The same problem holds for two additional theoretical questions: the domain generality of SL in development and its relation to other cognitive skills. The lack of reliability argues caution in interpreting correlations with other measures or other SL tasks. To give an example, based on the results of the first session alone, one could argue for a modality-sensitive characterization of SL: Individual performance was correlated within a modality but not between modalities. The second session, however, gave rise to a different pattern, not compatible with this conclusion. Our findings do not undermine the basic claim that humans are capable of extracting distributional information from their input, and that this mechanism plays a role in language acquisition. The lack of reliability does not invalidate the findings at the group level: The present tasks are informative for studying which relations (and when) children can learn (though their psychometric weakness means that they are prone to noise and measurement error). However, they cast doubt on their appropriateness as a measure for individual differences in development.

Before discussing the source of the low reliability in children, let me briefly address the pattern of correlations between the three SL tasks in our adult data. Examining performance on several SL tasks across modalities can shed light on the modality-specific nature of SL. Although SL is found across modalities (e.g., Conway & Christiansen, 2005), there is evidence for modality-based differences in performance in both children (Raviv & Arnon, 2017) and adults (Emberson, Conway, & Christiansen, 2011; Frost, Armstrong, Siegelman, & Christiansen, 2015). From the perspective of individual differences, adults' performance on visual and auditory tasks is not correlated (Erickson et al., 2016; Siegelman & Frost, 2015), a finding used to argue that SL is not a unitary capacity, but one that is sensitive to modality. Our findings give rise to more complex pattern: In both sessions, performance was not correlated on the two auditory tasks, but was correlated between the visual and the nonlinguistic auditory tasks. That is, performance patterned together on the basis of stimulus type (linguistic vs. nonlinguistic) rather than modality (auditory vs. visual). This pattern is in line with a set of recent findings suggesting that the nature of stimuli, and in particular whether it is linguistic or nonlinguistic, is an important factor in explaining SL performance. SL tasks using linguistic stimuli (syllables) are more affected than visual tasks by the exact stimuli used and show evidence of L1 influences (Siegelman et al., 2018b). Developmentally, the effect of age on performance seems different for tasks using linguistic stimuli: Whereas performance on visual and nonlinguistic auditory SL improves during childhood (ages 5–12 years), performance on linguistic auditory SL does not (Shufaniya & Arnon, 2018). Since studies finding modality-based differences often use linguistic auditory tasks (Emberson et al., 2011; Siegelman & Frost, 2015), previously reported differences (and similarities) may be driven not only by modality by also by the specific stimuli use. Our findings offer further evidence for similarities across modalities and point to the linguistic nature of the stimuli as a possible cause for modality-based differences. Note, though, that the linguistic task differed from the two other tasks in having slightly shorter exposure (24 vs. 32 repetitions of each triplets) and shorter stimuli duration (250 ms per syllables vs. 500 ms per nonlinguistic sound/image), making it less similar to the other two.[3] Although the shorter stimulus duration on the linguistic auditory task did not prevent learning (children were above chance across sessions and samples), it could have impacted accuracy (Arciuli & Simpson, 2011; Emberson et al., 2011). Future work will be needed to assess the impact of those methodological differences on the pattern of correlations in this study.

Returning to question of reliability, where does the task instability come from, and why is it more pronounced in children? One possibility is that SL itself is not a stable individual

[3] Although the optimal design is one in which stimulus duration is identical across the three tasks, this would have introduced two other (and potentially more influential) problems. Making the linguistic auditory syllables longer would have (a) made our task unlike ASL tasks used in the literature, in which syllable duration is usually around 200 ms, and (b) made our ASL stimuli less natural: syllables in natural language are shorter than 500 ms (the presentation duration of the visual and nonlinguistic stimuli), and shorter than the duration of familiar sounds. For those reasons, we opted to keep the linguistic auditory stimuli at the 250-ms presentation duration.

property, but rather one that is affected by an unknown combination of other cognitive abilities, like memory and attention. Children may be more vulnerable to such influences, explaining why reliability was higher in adult learners in this study and in studies using similar tasks (Siegelman, Bogaerts, & Frost, 2017b; Siegelman & Frost, 2015). However, a more likely interpretation for the lack of reliability is that SL is a stable property also in children, but one that is not well measured by the tasks we currently use. Indeed, low reliability in children has also been found in other learning tasks that are reliable in adults: A recent article assessed the reliability of commonly used procedural memory measures (SRT, Hebbian learning, and contextual cuing), and found that they have very low reliability in children and do not show stable correlations with language or literacy measures (West, Vadillo, Shanks, & Hulme, 2017).

The present study does not allow us to discern which aspects of SL tasks are responsible for their low reliability in children, but points to several plausible directions. The low reliability may be driven by test properties (number of test trials, level of difficulty), its explicit nature, or a combination of both. The fact that many children did not show learning at an individual level in both sessions (around 80% in the two auditory tasks and 50% in the visual one), and that they showed worse performance than adults, indicates that the task was difficult. This is not a unique feature of our tasks: Group accuracy rates were similar or higher than those found with children of similar ages in other studies (Kidd & Arciuli, 2016; Raviv & Arnon, 2017; Spencer et al., 2015). These accuracy levels are problematic, because they do not generate the wide distribution of performance needed for individual difference studies. Although it is clear that the tasks need to be made easier, there is not a straightforward way to do so. Reducing the number of triplets may alleviate memory demands, but it also impacts the statistical structure in a way that makes triplet boundaries *less* salient (having only three triplets instead of five means that the TPs between triplets are now .5 instead of .25). Increasing exposure times may also not solve the problem. Accuracy rates in our study were similar to those found when exposing children of the same age to a word segmentation task with four times the exposure length of the present study (59% in Saffran, Newport, Aslin, Tunick, & Barrueco, 1997, vs. our 57%). In fact, using online measures to track the trajectory of visual statistical learning in adults found that learning is already evident after seven repetitions of each triplet and does not increase in magnitude with increased exposure (Siegelman et al., 2018a). Similar patterns were found in the auditory domain, in which learning plateaued after only three repetitions (Batterink, 2017; though see Batterink & Paller, 2017, for enhanced neural activation following increased exposure). That is, it is unclear whether increasing exposure length will improve learning significantly. Introducing a delay between exposure and testing may be another way to raise accuracy rates (thereby increasing the proportion of scores that are above chance): Implicit learning seems to improve when tested after a 12-h delay, and even after a one-week delay (Nemeth & Janacsek, 2011).

Although task difficulty needs to be addressed, several findings suggest that it is not the major factor driving the low reliability in children. First, test–retest reliability was not higher for the visual task, which had higher accuracy overall and in which performance was similar to that of adults (Raviv & Arnon, 2017). Second, the proportion of individual children showing above chance performance was similar to that found in adults (Siegelman, Bogaerts, & Frost, 2017b), for whom task reliability was higher. One intriguing possibility is that children show greater fatigue effects during testing (and more individual variation in the magnitude of such effects), and that this contributes to the low task reliability. A recent article has pointed to the possible impact of fatigue on task reliability. Török, Janacsek, Nagy, Orbán, and Nemeth (2017) documented reactive inhibition (accumulative performance deterioration) in adults performing an SRT task and showed that such effects vary among individuals, and that modeling them (by including individual reactive inhibition terms) revealed larger learning effects (Török et al., 2017). Children may show greater fatigue effects than adults and more individual variation in the magnitude of effects, both of which could contribute to the low reliability. Another possibility is that the children are more affected than adults by the explicit nature of the task and that reliability will be improved by the development of more implicit measures of learning. Such measures have been used with adults (self-paced visual presentation: Karuza, Farmer, Fine, Smith, & Jaeger, 2014; rapid serial auditory presentation: Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015) and have shown both reliability and stable correlations with offline performance (Siegelman et al., 2018a). Given similar debates on the difference between online and offline measures of behavior in children (e.g., in the theory-of-mind literature), this seems like a promising avenue for addressing the low reliability. My lab is currently developing online measures that can be used to assess both auditory and visual SL over a range of ages, from infancy to adulthood.

Interestingly, even the move to more implicit methods may not be enough. A similar debate about the ability of SL measures to predict individual variation has taken place within the literature on infant speech (see Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014). The logic is similar to that used in the SL literature: Being better at detecting linguistic regularities (phonemes in this case) should lead to better (or faster) language learning outcomes. This claim is supported by findings showing positive correlations between speech perception measures and vocabulary size (e.g., Tsao, Liu, & Kuhl, 2004). However, as in the case of the SL literature, there are few data on the stability and reliability of speech perception measures in infancy. In their recent article, Cristia et al. offered a theoretical and methodological critique of these findings highlighting the psychometric weakness of these speech perception measures, which like SL tasks, were developed to assess group-level learning and not individual differences: "Making the jump from correlation to causation requires multidisciplinary

approaches and an improvement of the measurements used" (Cristia et al., 2014, p. 1331). Similar critiques are offered in a recent study questioning the reliability of sequence learning in adults and its ability to serve as a stable measure of individual differences (Bogaerts, Siegelman, Ben-Porat, & Frost, 2018).

These debates highlight the difficulty of using measures developed to assess group-level performance in the study of individual differences, as well as the impact of these methodological challenges on our theoretical understanding. There is a pressing need to further investigate the psychometric properties of existing SL tasks and, more importantly, to develop novel tasks that are better suited to assess individual differences. Until that is done, little can be concluded about the relation between variation in SL and individual differences in language-learning outcomes.

Author note Thanks to Noam Siegelman for comments and help with the statistical analyses, and Louisa Bogaerts and Ram Frost for comments and helpful discussions. Additional thanks to Zohar Aizenbud and Amir Efrati for assistance in programming the experiments and coordinating the testing, as well as to the research assistants who collected the data: Yuval Braeman, Noa Bar, Shira Zicherman, Hilla Merhav, Amir Efrati, Amir Shufaniya, and Hana Gerchikov. Special thanks to Maytal Wiener, who collected the data for the second child study. I also thank the children, parents, and teachers at the Givat Mesu'aa and David primary school. The research was funded by an Israeli Science Foundation grant to the first author (Grant No. 584/16).

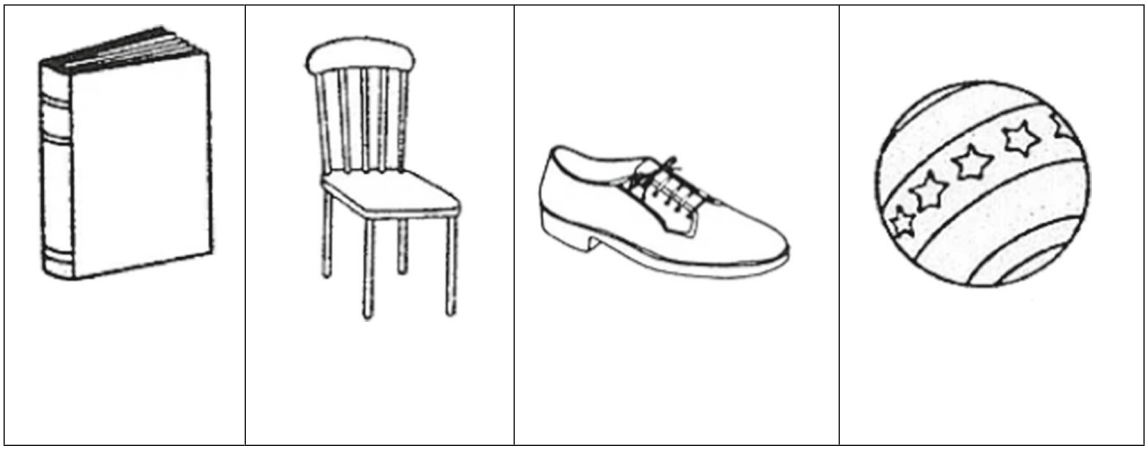
Appendix: Stimuli used in the three SL tasks

Linguistic auditory task (syllables): du, ka, me, ge, di, no, ki, mu, ga, na, lo, bi, to, be, lo

Nonlinguistic auditory task (familiar sounds): "bird tweet," "running water," "goat bleat," "opening door," "dog bark," "bouncing ball," "trumpet," "cat meow," "duck quack," "frog quack," "cuckoo clock," "bell," "chord," "whistle," and "cow moo."

Visual task: Black-and-white drawings (taken from Alario & Ferrand, 1999) of a house, book, cake, dog, cat, airplane, shoe, fish, ball, banana, fork, flower, bottle, chair, and butterfly.

Example images:



Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*, 531–552. https://doi.org/10.3758/BF03200732

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science*, *14*, 464–473. https://doi.org/10.1111/j.1467-7687.2009.00937.x

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

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science*, *28*, 921–928. https://doi.org/10.1177/0956797617698226

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*, 341–347.

Bogaerts, L., Siegelman, N., Ben-Porat, T., & Frost, R. (2018). Is the Hebb repetition task a reliable measure of individual differences in sequence learning? *Quarterly Journal of Experimental Psychology*, *71*, 892–905. https://doi.org/10.1080/17470218.2017.1307432

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

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

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*, 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*, 1330–1345. https://doi.org/10.1111/cdev.12193

Ellis, E. M., Robledo Gonzalez, M., & Deák, G. O. (2014). Visual prediction in infancy: What is the association with later vocabulary? *Language Learning and Development*, *10*, 36–50. https://doi.org/10.1080/15475441.2013.799988

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, *64*, 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. S. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra, 2*(1), 14. https://doi.org/10.1525/collabra.41.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation: A new measure of statistical learning in speech segmentation. *Experimental Psychology*, *62*, 346–351. https://doi.org/10.1027/1618-3169/a000295

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*, 1243–1252. https://doi.org/10.1177/0956797612472207

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103–127. https://doi.org/10.1080/09658219408258940

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review*, *20*, 1161–1169. https://doi.org/10.3758/s13423-013-0458-4

Havron, N., & Arnon, I. (2017). Minding the gaps: Literacy enhances lexical segmentation in children learning to read. *Journal of Child Language*, *44*, 1516–1538. https://doi.org/10.1017/S0305000916000623

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

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

Kaufman, A. (1994). Intelligent testing with the WISC-III. New York: Wiley

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

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*, 184–193. https://doi.org/10.1111/cdev.12461

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.

Lany, J., Shoaib, A., Thompson, A., & Graf Estes, K. (2018). Infant statistical-learning ability is related to real-time language processing. *Journal of Child Language*, *45*, 368–391. https://doi.org/10.1017/S0305000917000253

Mainela-Arnold, E., & Evans, J. L. (2014). Do statistical segmentation abilities predict lexical–phonological and lexical–semantic abilities in children with and without SLI? *Journal of Child Language*, *41*, 327–351. https://doi.org/10.1017/S0305000912000736

Maital, S. L., Dromi, E., Sagi, A., & Bornstein, M. H. (2000). The Hebrew Communicative Development Inventory: Language specific properties and cross-linguistic generalizations. *Journal of Child Language*, *27*, 43–67. https://doi.org/10.1017/S0305000999004006

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

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

Nemeth, D., & Janacsek, K. (2011). The dynamics of implicit skill consolidation in young and elderly adults. *Journal of Gerontology*, *66B*, 15–22. https://doi.org/10.1093/geronb/gbq063

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill

Potter, C. E., Wang, T., & Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, *41*, 913–927. https://doi.org/10.1111/cogs.12473

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*, e12593:1–13. https://doi.org/10.1111/desc.12593

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906–914. https://doi.org/10.1002/wcs.78

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Shafto, C. L., Conway, C. M., Field, S. L., & Houston, D. M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy*, *17*, 247–271. https://doi.org/10.1111/j.1532-7078.2011.00085.x

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, *42*, 3100–3115. https://doi.org/10.1111/cogs.12692

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017a). Towards a theory of individual differences in statistical learning. *Transactions of the Royal Society B*, *372*, 20160059. https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., & Frost, R. (2017b). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*, 418–432. https://doi.org/10.3758/s13428-016-0719-z

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018a). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, *42*(Suppl. 3), 692–727. https://doi.org/10.1111/cogs.12556

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018b). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*, 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing*, *28*, 467–490. https://doi.org/10.1007/s11145-014-9533-0

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*, 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Török, B., Janacsek, K., Nagy, D. G., Orbán, G., & Nemeth, D. (2017). Measuring and filtering reactive inhibition is essential for assessing serial decision making and learning. *Journal of Experimental Psychology: General*, *146*, 529–542.

Tsao, F., Liu, H., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, *75*, 1067–1084.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*, 552–564. https://doi.org/10.1037/0096-3445.134.4.552

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21*, e12552:1–13. https://doi.org/10.1111/desc.12552