Journal of Experimental Child Psychology 200 (2020) 104964

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Measuring children's auditory statistical learning via serial recall

Evan Kidd [a,b,c,*], Joanne Arciuli [c,d], Morten H. Christiansen [c,e], Erin S. Isbilen [e], Katherine Revius [b], Michael Smithson [b]

[a] Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, the Netherlands
[b] Research School of Psychology, The Australian National University, Canberra, ACT 2601, Australia
[c] ARC Centre of Excellence for the Dynamics of Language, The Australian National University, Canberra, ACT 2601, Australia
[d] Caring Futures Institute, Flinders University, Bedford Park, SA 5042, Australia
[e] Department of Psychology, Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 8 December 2019
Revised 10 July 2020

Keywords:
Statistical learning
Auditory statistical learning
Verbal memory
Serial recall
Individual differences
Language

ABSTRACT

Statistical learning (SL) has been a prominent focus of research in developmental and adult populations, guided by the assumption that it is a fundamental component of learning underlying higher-order cognition. In developmental populations, however, there have been recent concerns regarding the degree to which many current tasks reliably measure SL, particularly in younger children. In the current article, we present the results of two studies that measured auditory statistical learning (ASL) of linguistic stimuli in children aged 5–8 years. Children listened to 6 min of continuous syllables comprising four trisyllabic pseudowords. Following the familiarization phase, children completed (a) a two-alternative forced-choice task and (b) a serial recall task in which they repeated either target sequences embedded during familiarization or foils, manipulated for sequence length. Results showed that, although both measures consistently revealed learning at the group level, the recall task better captured learning across the full range of abilities and was more reliable at the individual level. We conclude that, as has also been demonstrated in adults, the method holds promise for future studies of individual differences in ASL of linguistic stimuli.

* Corresponding author at: Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, the Netherlands.
E-mail address: evan.kidd@mpi.nl (E. Kidd).

https://doi.org/10.1016/j.jecp.2020.104964
0022-0965/Crown Copyright © 2020 Published by Elsevier Inc. All rights reserved.

Introduction

Humans have a remarkable ability to learn, which is supported by a range of mechanisms that are suited to specific learning tasks and domains. One notable form of learning is *statistical learning* (SL)—the process of using probabilistic co-occurrence to group elements present in the environment. SL is operational in a rudimentary form in neonates (Bulf, Johnson, & Valenza, 2011; Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009), is robustly observed across different modalities during the first year of life (Emberson, Misyak, Schwade, Christiansen, & Goldstein, 2019; Kirkham, Slemmer, & Johnson, 2002; Saffran, Aslin, & Newport, 1996), and at least in the visual domain continues to develop throughout childhood (Arciuli & Simpson, 2011; Raviv & Arnon, 2018). Studies that report associations between SL and other cognitive skills (e.g., spoken language, reading) suggest that SL could be an important component of human cognition that varies across individuals (e.g., Arciuli & Simpson, 2011; Conway, Bauernschmidt, Huang, & Pisoni, 2010; Frost, Armstrong, & Christiansen, 2019).

A common experimental paradigm for measuring SL is the embedded triplet task, in which stimuli are presented sequentially during a familiarization phase, forming recurring triplets of items within a continuous stream (e.g., Saffran et al., 1996). This familiarization phase is followed by a surprise test phase. In infants, learning is measured using preference methods (e.g., head-turn preference, habituation–dishabituation). In adults, learning has typically been measured via methods that require explicit decision making, such as two-alternative forced-choice (2AFC) trials, where participants make a decision between a target (i.e., an embedded triplet that was presented during familiarization) and a foil (comprising stimuli that did not appear as a triplet during familiarization).

The inappropriateness of using infant methods with children has meant that many studies of SL in children have used modified versions of the embedded triplet task typically used with adults. This strategy has met with some success (e.g., Arciuli & Simpson, 2011, 2012; Evans, Saffran, & Robe-Torres, 2009; Qi, Sanchez, Georgan, Gabrieli, & Arciuli, 2019; Raviv & Arnon, 2018; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). However, as the field has progressed, task reliability has become a central concern, most prominently when measuring SL as an individual capacity. Whereas some SL tasks that use nonlinguistic stimuli and 2AFC trials demonstrate good psychometric properties in children (e.g., internal consistency; Qi et al., 2019; Torkildsen, Arciuli & Wie, 2019), recent work investigating auditory statistical learning (ASL) for linguistic stimuli (henceforth *linguistic ASL*) during childhood using the 2AFC task has revealed comparatively poor results (Arnon, 2020). Methodological differences across studies no doubt contribute to the inconsistent results (see Arciuli & Conway, 2018), and there is a need for sensitive and reliable methods that capture linguistic ASL at younger ages. In the current research, we compared the relative sensitivity and reliability of 2AFC and serial recall to measure linguistic ASL, based on an approach that has been successfully used with child and adult populations (Isbilen, McCauley, Kidd, & Christiansen, 2017, 2020; Majerus, van der Linden, Mulder, Meulemans, & Peters, 2004).

SL: Reflection- and processing-based measurements

SL is a form of implicit learning and thus is assumed to occur below the threshold of conscious awareness (Reber, 1993). Most tasks modeled on the embedded triplet paradigm contain two phases: (a) a familiarization phase during which implicit learning occurs, and (b) a subsequent test phase that aims to measure this learning. When measured via 2AFC, participants are asked to explicitly compare a target sequence and a foil, and then indicate which sequence is more familiar. Because such *reflection-based* measurements require participants to make a judgment about what they have implicitly learned, they may only indirectly tap into participants' knowledge of the learned material (Christiansen, 2019). There is ongoing discussion about how measurement interacts with modality and other methodological details because there has been some success in measuring SL at the individual level using reflection-based measurement, at least in older children (e.g., Torkildsen et al., 2019) and adults (e.g., Siegelman et al., 2017). However, requiring explicit decision making in young children has likely contributed to some of the task reliability problems concerning linguistic ASL, given that there appears to be a lower bound at which it is possible to observe consistent learning effects in children (~7 years of age; Raviv & Arnon, 2018).

Other *processing-based* measures of ASL exist, including those that use reaction times (e.g., Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; Lammertink, van Witteloostuijn, Boersma, Wijnen, & Rispens, 2019; Qi et al., 2019) or electroencephalography (e.g., Batterink, Reber, Neville, & Paller, 2015; Batterink & Paller, 2017; Mandikal Vasuki, Sharma, Ibrahim, & Arciuli, 2017a, 2017b), which allow learning to be measured over the course of familiarization. Some research has shown that processing- and reflection-based measurements are not always associated, most prominently for linguistic ASL (Batterink et al., 2015; Isbilen et al., 2017, 2020). By contrast, investigations of nonlinguistic ASL in adults (Mandikal Vasuki et al., 2017b) and in children (Mandikal Vasuki et al., 2017a) show a closer alignment of results from processing- and reflection-based measures. In the case of language-like stimuli, the results suggest that ASL leads to knowledge that can be explicitly reflected on but that its relationship to accrued implicit knowledge may be weak.

In this article, we report on a child-friendly processing-based task for measuring linguistic ASL that is inspired by work on verbal memory and which has been successfully used with adults. In Isbilen et al. (2017, 2020), adults participated in a Saffran-style embedded triplet task (Saffran et al., 1996) in which they listened to 11.5 min of a continuous speech stream that contained six embedded trisyllabic sequences. After this familiarization phase, participants completed subsequent test phases comprising a reflection-based 2AFC task and a processing-based serial recall task in which they repeated either two trained target sequences concatenated together or a random sequence of 6 syllables. Isbilen et al. (2017) called this latter task *statistically induced chunking recall* (SICR). They observed successful learning at the group level for both the 2AFC and SICR tasks, the latter denoted by significantly greater recall of target sequences compared with that of random foil sequences. Notably, whereas the 2AFC task did not show acceptable test–retest reliability in the auditory domain (r = .19), the SICR task did (.63 < rs < .81), which appeared to reflect the acquisition of sequence-specific statistics and not simply verbatim memory for syllables.

In addition to showing promising reliability in adult participants, serial recall has two features that suggest it would make an appropriate processing-based measure of linguistic ASL. First, there is good evidence for a direct link between the speech perception and production systems. Neuroimaging studies with adults show that motor areas involved in speech production are activated during perception tasks (e.g., Glanz et al., 2018; Wilson, Saygin, Sereno, & Iacoboni, 2004), and there is strong evidence identifying this link as an important basis for early language development (e.g., Vihman, 2017; Vilain, Dole, Loevenbruck, Pascalis, & Schwartz, 2019). Thus, we can assume a relatively direct link between speech perception during familiarization and speech production at test. Second, there is strong evidence to suggest that serial recall of verbal information taps into linguistic information stored in long-term memory and used for language production (e.g., Acheson, Hamidi, Binder, & Postle, 2011; Jones, Gobet, Freudenthal, Watson, & Pine, 2014; Jones, Gobet, & Pine, 2007; Majerus & van der Linden, 2003; Potter & Lombardi, 1990; Szewczyk, Marecka, Chiat, & Wodniecka, 2018). Thus, the use of serial recall has the potential to contribute to a component of SL that is not readily observable using reflection- or processing-based measures such as reaction times and electroencephalography—the output of SL.

Evidence for this assertion comes from three sources. The first is the literature on Hebbian learning. First introduced by Hebb (1961), the Hebbian repetition effect (HRE) describes an empirical effect whereby participants repeat sequences of syllables that either follow a fixed order and are repeated—the Hebb sequence—or follow random sequences that are not repeated. Across the course of the experiment, repetition of the Hebb sequence improves relative to the random sequences, suggesting long-term incidental learning. Page and Norris (2009) argued that performance in the Hebb paradigm and performance in vocabulary acquisition both are subserved by the same underlying processes—specifically, that the acquisition of phonological codes is best described as Hebbian learning (Szmalec, Page, & Duyck, 2012). Thus, the HRE describes the process of grouping sublexical information into higher-order chunks based on co-occurrence information (Smalle et al., 2016), a feature of SL. Furthermore, like SL, the HRE is implicit (McKelvie, 1987). Several studies have shown that the HRE is sensitive to developmental differences in language development (e.g., Bogaerts, Szmalec, De Maeyer, Page, & Duyck, 2016; Mosse & Jarrold, 2008; Smalle et al., 2016; Smalle, Muylle, Szmalec, & Duyck, 2017; Smalle, Page, Duyck, Edwards, & Szmalec, 2018). However, an important limitation of the HRE as an individual differences measure is that at least one study has reported low test–retest reliability (Bogaerts, Siegelman, Ben-Porat, & Frost, 2018).

The second source comes from a long history of work on repetition as a general methodology for child language research, which shows that it is sensitive to linguistic representations across the entire system. Accordingly, nonword repetition taps long-term knowledge for phonemic sequences at multiple grain sizes (e.g., Baddeley, Gathercole, & Papagno, 1998; Gathercole & Baddeley, 1989; Jones et al., 2007, 2014; Majerus & van der Linden, 2003; Szewczyk et al., 2018), and sentence repetition is also sensitive to fine distinctions in children's grammatical knowledge (e.g., Kidd, Brandt, Lieven, & Tomasello, 2007), with the assumption being that repetition implicates parsing routines underlying sentence comprehension and production (e.g., Acheson & MacDonald, 2009; Potter & Lombardi, 1990).

Finally, there is evidence from Majerus et al. (2004) that implicit phonological learning can take place in children and adults after a relatively short amount of passive exposure and that this effect is observable through serial recall. In that study, adults and 8-year-old children listened to a 30-min continuous sequence of syllables that manipulated sublexical phonological rules—specifically, how phonemes could combine within syllables and how syllables could combine into pseudowords. Participants then completed a repetition task, demonstrating greater recall of grammatically legal terms than grammatically illegal items.

The current study

SL has been implicated as foundational to development across many domains, most prominently language and literacy (e.g., Arciuli, 2018; Saffran et al., 1996). However, efforts to reliably measure SL in children have not always been as successful as in other cognitive domains (e.g., working memory, IQ), most notably at the level of the individual. In younger children, and for linguistic ASL in particular, this appears to be at least partially attributable to the difficulty in measuring implicit processes via reflection-based methods. Thus, there is a need for more sensitive and reliable measures for developmental populations. With these points in mind, we investigated the use of serial recall as an alternative assessment of linguistic ASL in children aged 5 to 8 years and compared performance with learning as measured by the standard 2AFC task. Following Isbilen et al.'s (2017, 2020) work with adults, we used a developmental version of the SICR task. In Study 1, children listened to three blocks of continuous speech containing embedded triplets of co-occurring syllables that lasted just over 6 min in total, after which their learning was tested using 2AFC test trials. Children then completed the SICR repetition test in which they repeated either target sequences or foils. We predicted that we would observe significant learning effects for both measures but that, because it is a processing-based measure, repetition would produce a more robust learning effect. We also investigated the relationship between the two measures of learning and their relationship to age. In Study 2, we followed up one particularly notable result in a separate sample of children, in addition to testing the adequacy of both the 2AFC and SICR measures in capturing individual differences in linguistic ASL.

Study 1

Method

Participants

A total of 143 children (69 female) aged 5;6 (years;months) to 7;7 (*M* = 6;8, *SD* = 4.45 months) were recruited from the first two grades of primary schools in Canberra, Australian Capital Territory (ACT), for participation in a larger longitudinal study tracking the development of SL across childhood. We did not conduct an a priori power analysis to determine our final sample size because the size of the learning effect for the different dependent measures was hypothesized to vary and in the case of SICR was unknown. Therefore, we aimed to recruit as many children as possible in order to provide a reliable estimate of these effects and report post hoc power analyses and confidence intervals (CIs) to determine the precision of the effect size given the sample, with narrow CIs indicating greater precision (Steiger & Fouladi, 1997). Recruitment criteria included (a) monolingual English speakers with no more than 1 day per week exposure to another language and (b) no existing or past diagnosis of a language or cognitive impairment, learning difficulty, or hearing loss. Consistent with the demographics of the region, children's socioeconomic status was high relative to the rest of Australia, with all children drawn from areas in the 89th percentile or higher (*M* = 93.86, range = 89–100) as measured by the 2011 Socio-Economic Indexes for Areas (SEIFA) of the Australian Bureau of Statistics. However, relative to areas within the ACT, the SEIFA scores were more variable (*M* = 67.86, range = 45–99).

Materials

The ASL task was completed on a laptop computer (Dell Latitude XT3), with the auditory stimuli presented via E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA, USA). The auditory stimuli consisted of 12 syllables, which were synthesized using the MBROLA speech synthesizer. Children listened to the auditory stimuli through circumaural headphones (Sennheiser HD 280 Pro). Children's recall of the test items was recorded using a Zoom H4n Pro handheld recorder.

Procedure

All children individually completed the ASL task in a quiet area of their school. The task was presented as a game in which children were introduced to an alien character called "Bok", and were told that they would be listening to alien signals coming through Bok's spaceship radio. The task employed the embedded triplet paradigm (Saffran et al., 1996) and consisted of three phases: (a) a familiarization phase, (b) a 2AFC test phase, and (c) a repetition test phase that, following Isbilen et al. (2017), we refer to as SICR. We describe each phase below (see also Fig. 1).

Familiarization phase. During the familiarization phase, children listened to a continuous sequence of syllables that, unbeknown to them, consisted of four trisyllabic triplets: *kibudu, modipa, takapo,* and *lomari* (for a complete list of targets and foils, see Appendix A). The average length of each syllable was 305 ms, with an average intersyllable gap of 120 ms. Following Saffran et al. (1996), triplets were defined by high transitional probabilities (TPs) within triplets (TPs = 1.0) and lower TPs between triplets (TPs = .33). The full familiarization exposure consisted of 72 tokens of each triplet; however, piloting of the task revealed that children found it difficult to listen to the full sequence. Therefore, we broke up the familiarization phase into three blocks where children heard, on average, 24 tokens of each triplet.[1] Each familiarization block lasted approximately 2 min. Each break in between segments lasted approximately 20 s.

2AFC test phase. At the end of the familiarization phase, children completed a standard 2AFC task that aimed at measuring their learning of triplets presented during the familiarization phase. Across 32 trials, children heard one target triplet and one foil, which were presented 1000 ms apart. Children's task was to indicate which group of syllables was from Bok's radio signal. Foils were constructed from the same syllables as the targets (e.g., the foil *kaburi* consisted of syllables taken from the three target triplets *takapo, kibudu,* and *lomari*) but crucially differed from the targets in their statistical structure, such that the within-foil TP between syllables was 0.

SICR test phase. The SICR component of the task required children to repeat syllable sequences across two conditions. In the first condition, children were asked to repeat 8 3-syllable sequences that were either (a) an embedded target triplet, or (b) a foil triplet. The foils differed from the foils used during the 2AFC test phase but were constructed using the same principle (i.e., each contained syllables from target triplets, such that TP = 0 in the training set). In the second condition, children were asked to repeat 16 6-syllable sequences that consisted of either (a) 2 3-syllable target triplets concatenated

[1] The full exposure phase does not divide equally by 3; thus, Blocks 1 and 2 contained 24 tokens of each trisyllablic target word plus an additional 2 triplets, and Block 3 had 23 tokens of each target word. This is why the blocks differed slightly in length.

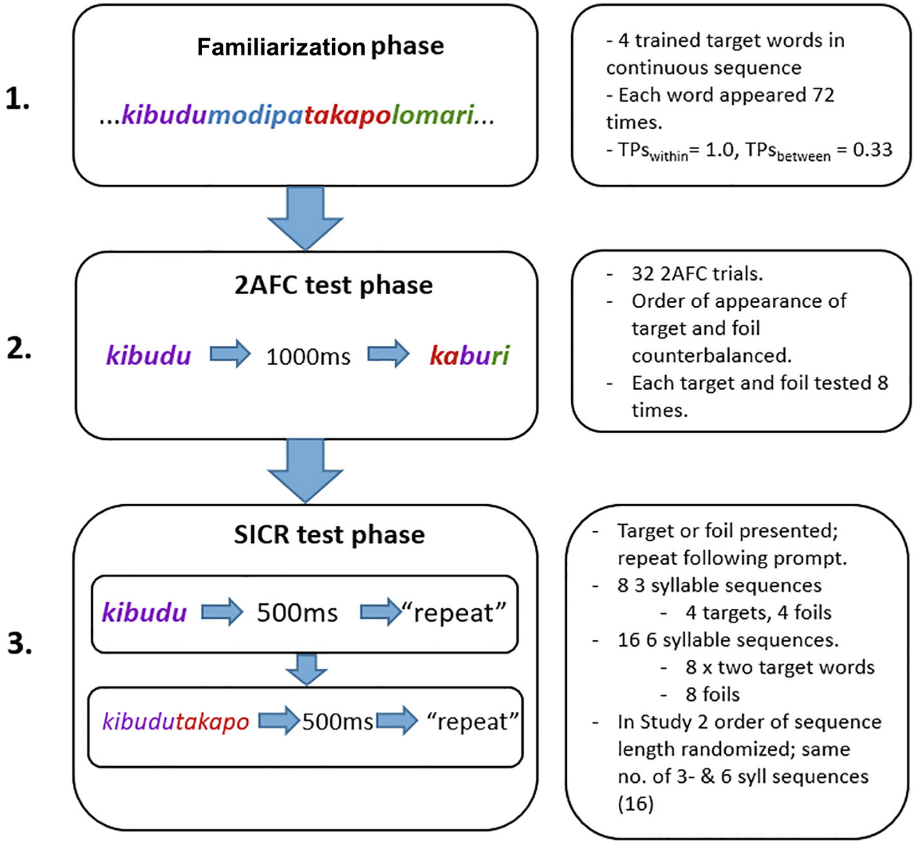
1.
Familiarization phase
...kibudumodipatakapolomari...
- 4 trained target words in continuous sequence
- Each word appeared 72 times.
- TPs within = 1.0, TPs between = 0.33
2.
2AFC test phase
kibudu
1000ms
kaburi
- 32 2AFC trials.
- Order of appearance of target and foil counterbalanced.
- Each target and foil tested 8 times.
3.
SICR test phase
kibudu
500ms
"repeat"
kibudutakapo
500ms
"repeat"
- Target or foil presented; repeat following prompt.
- 8 3 syllable sequences
- 4 targets, 4 foils
- 16 6 syllable sequences.
- 8 x two target words
- 8 foils
- In Study 2 order of sequence length randomized; same no. of 3- & 6 syll sequences (16)

Fig. 1. Schematic overview of the ASL (auditory statistical learning) task. 2AFC, two-alternative forced-choice; SICR, statistically induced chunking recall; TPs, transitional probabilities.

together or (b) a 6-syllable foil that was once again generated from syllables in the training set and had TPs = 0.

The choice to include both 3-syllable and 6-syllable sequences was motivated by piloting. Following Isbilen et al. (2017, 2020), we originally included only 6-syllable sequences; however, some children were reluctant to attempt to repeat the items, indicating that they perceived the longer sequencs as too difficult. This is consistent with work on nonword repetition and verbal working memory, which has shown a marked increase in difficulty in repeating nonwords of more than 4 syllables (Gathercole, Willis, Baddeley, & Emslie, 1994) or sequences of more than three or four items in length (Cowan, 2001). Therefore, we added in the 3-syllable sequences, which served to introduce children to the concept of repetition using sequences of a more manageable length. As such, the presentation of 3-syllable and 6-syllable conditions was blocked, with the 3-syllable condition being first. For each item, children were instructed to listen first and then repeat the item when they saw a cross on the computer screen, which appeared 500 ms after the offset of the item.

Scoring

Responses on the 2AFC component of the task were automatically scored by E-Prime (correct = 1, incorrect = 0). Children's productions in the SICR component of the task were transcribed and then coded, following Isbilen et al. (2017, 2020), with some accommodations made for differences in adult versus child speech. The SICR task is reminiscent of Hebbian repetition learning (Hebb, 1961) in that participants repeat sequences that either follow or do not follow a specified distribution. Following work in the Hebbian learning literature (e.g., McKelvie, 1987; Norris, Page, & Hall, 2018; Smalle et al., 2016, 2018) and Isbilen et al. (2017, 2020), children were awarded 1 point for every syllable that they correctly recalled in the correct serial order. For instance, if the target was *lomaritakapo* and a child produced *romabitakaku,* the child was awarded a score of 3/6 syllables correct and a score of 0.5. This is because the syllables *ma, ta,* and *ka* occurred in the correct serial order. As in Hebbian studies with both children and adults, participants sometimes do not produce the exact number of target syllables. In this case, participants were still awarded a score of 1 for each syllable that was correctly recalled relative to those syllables that were recalled. For instance, for the 6-syllable foil item *kipabumodudi,* one child produced *kapamodi* (transcribed as *ka-pa-xx-mo-xx-di,* where *xx* represents a missing syllable). In this case, the child was awarded a score of 3/6 syllables correct (0.5) because the syllables *pa, mo,* and *di* occurred in the correct serial order.

Because we used synthesized speech, there were often perception effects on children's productions (as there are in adults; Isbilen et al., 2017, 2020). For instance, some participants heard /b/ as /v/ and consistently produced /v/ for /b/ in their productions (e.g., producing *bupadi* as *vupadi*). In these cases, consistent phoneme substitution (production of /v/ in 75% or more of instances) was counted as correct. Sometimes, children also produced blended consonants (e.g., producing *lomari* as *blomari*). These cases were coded in favor of participants if the original phoneme was preserved (i.e., /l/), under the assumption that the blend could be a perception issue derived from using synthesized speech. All vowels in the materials were long (monopthongs), but Australian English has a significant inventory of diphthongs and it is not unlikely that a speaker with a broad Australian English accent would produce *lomari* as *lomarai.* In these cases, coders used their judgment as to whether the repetition was an accurate reproduction of the intended vowel. Finally, repeated phonemes or syllables (e.g., *l, lo, lomari*) were not penalized if it was clear that they were hesitations or false starts.

The data of 10 participants (7% of total sample) were retranscribed and coded by a second blind coder for intercoder reliability, computed using Cohen's kappa. There was excellent agreement among the coders. Cohen's kappa for individual participants ranged from substantial to nearly perfect (.75–.98), with the average kappa for target items being .83 (95% CI [.79, .88]) and the average kappa for foils being .91 (95% CI [.87, .94]). Kappa did not differ for 3-syllable repetitions (.89, 95% CI [.83, .95]) and 6-syllable repetitions (.86, 95% CI [.82, .89]).

Results

Data and analyses can be access on the Open Science Framework (https://osf.io/nz6qj/). We first report whether we observed learning in both dependent variables: performance on the 2AFC trials and the SICR component. Four children were removed from the analysis of the 2AFC component of the task because research notes indicated that they were not attending or that they were answering according to a clear strategy (e.g., by indicating that the correct answer was always the first or second sequence played in the 2AFC trials). Overall mean performance on the 2AFC trials was 52.4% (SD = 0.10), which was significantly above chance, $t(138) = 2.81$, $p = .006$, $d = 0.24$, 95% CI (d) = [0.07, 0.41], power$_{(1-\beta)}$ = .81, 95% CI $(1-\beta)$ = [.13, 1.00] (all comparisons report two-tailed p values).[2] Fig. 2 is a pirate plot depicting children's performance on the SICR trials by length (3 vs. 6 syllables) and type (target vs. foil). Pirate plots show a combination of the raw data, the central tendency (i.e., the mean as indicated by the solid black bar), the inference band around the mean, and the (smoothed) data density (indicated by the shaded gray area) (see Phillips, 2018).

Fig. 2 shows that children performed better on the targets than on the foils for both the 3-syllable and 6-syllable items. Two children were removed from the analysis because they were not attending to the task, and an additional two children did not repeat any foils. Two paired-samples t tests, comparing performance on the targets versus the foils for each item length type, revealed that children recalled the targets significantly better than the foils {3 syllables: M_{diff} = .072, $t(138) = 4.51$,

[2] Note that the result was still significant when the entire sample was included {M = .525, SD = .10, $t(142) = 3.02$, $p = .003$, d = 0.25, 95% CI (d) = [0.09, 0.42], power$_{(1-\beta)}$ = .84, 95% CI $(1-\beta)$ = [.18, 1.00]}.

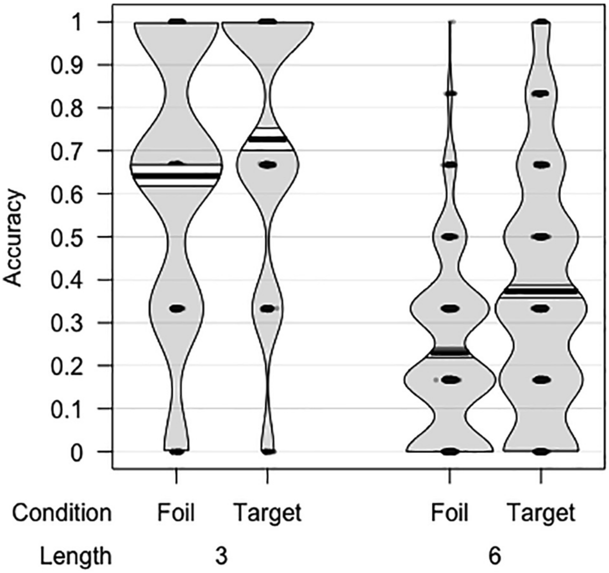

Fig. 2. Pirate plot depicting children's accuracy during the SICR (statistically induced chunking recall) test phase by item type (target vs. foil) and sequence length (3- vs. 6-syllables).

p < .001, d = 0.38, 95% CI (d) = [0.21, 0.55], power$_{(1-\beta)}$ = .99, 95% CI (1 − β) = [.69–1.00]; 6 syllables: M_{diff} = .141, t(138) = 12.06, p < .001, d = 1.02, 95% CI (d) = [0.82, 1.23], power$_{(1-\beta)}$ = 1.00, 95% CI (1 − β) = [1.00, 1.00]}.

Overall, we see significant learning in every measure but distinct differences in the size of the effects. The 2AFC result yielded a small effect, whereas the SICR trials yielded larger effect sizes, particularly in the 6-syllable condition. There are two possible explanations for the differences. The first is that the SICR task is a more sensitive measure of linguistic ASL, and the second is that all measures are equally sensitive but the children continued to learn throughout the experiment and therefore their knowledge of the embedded triplets improved relative to foils. There are several lines of evidence against the latter interpretation. First, across several studies with adults, Isbilen et al. (2017, 2020) found no reliable evidence that performance on one test type influences learning in the other, which is consistent with other studies that reported a dissociation between 2AFC and other processing-based measures of linguistic ASL (Batterink et al., 2015). In the data we report here, we found small but significant correlations between 2AFC performance and SICR performance, but in each case the shared variance was small (<3.8%; see Tables 1 and 2 in the next section), suggesting that the two measures were only weakly related. Finally, additional analyses reported in Appendix B showed no evidence of additional learning throughout either the 2AFC or SICR phase.

An alternative way to determine whether the two tasks are differentially sensitive is to determine whether there are asymmetries in the presence of the learning effects across the two measures. We considered this possibility by splitting the data into two groups: (a) those children who performed at or below chance (≤50%) on the 2AFC trials, and (b) those children who performed above chance (>50%). If the SICR measure is more sensitive, we should see evidence of learning even in children who did not show evidence of learning on the 2AFC trials. This is exactly what we found. The non-learning group comprising 65 children (29 female; M_{age} = 6;7, SD = 4.5 months, range = 5;8–7;4) had an average performance on the 2AFC trials of .44 (SD = .058), which was significantly below chance, t(64) = − 8.30, p < .001, d = − 1.03, 95% CI (d) = [−1.33, −0.73]). However, these children still showed evidence of learning via the SICR measure, performing numerically better on the repetition of targets than on the repetition of foils in the 3-syllable condition {M_{diff} = .045, t(64) = 1.83, p = .07, d = 0.23, 95% CI (d) = [−0.02, 0.47], power$_{(1-\beta)}$ = .47, 95% CI (1 − β) = [.05, .96]} and significantly better

Table 1
Bivariate correlations among 2AFC performance, SICR performance by item length (3 vs. 6 syllables) and item type (target vs. foil), and age (in months).

	3-Syllable target	3-Syllable foil	6-Syllable target	6-Syllable foil	Age
2AFC	.194*	.140	.177*	−.013	.289**
3-Syllable target		.497**	.353**	.316**	.055
3-Syllable foil			.380**	.418**	.043
6-Syllable target				.617**	.258**
6-Syllable foil					.119

Note. 2AFC, performance on two-alternative forced-choice component of ASL (auditory statistical learning) task; SICR, statistically induced chunking recall.
* $p < .05$.
** $p < .01$.

Table 2
Final model estimates for predicting ASL 2AFC performance from age and SICR scores.

Parameter	β	*SE*	95% CI (Wald)	Wald χ^2	*df*	*p*
Intercept	−8.25	2.48	[−13.12, −3.39]	11.08	1	.001
Log–age	1.90	0.57	[0.78, 3.01]	11.18	1	.001
Log–3 syllable target	0.05	0.02	[0.01, 0.09]	5.13	1	.024
Scale	1.14					

Note. ASL, auditory statistical learning; 2AFC, two-alternative forced-choice; SICR, statistically induced chunking recall; CI, confidence interval.

in the 6-syllable condition {M_{diff} = .11, $t(63)$ = 7.13, p < .001, d = 0.89, 95% CI (d) = [0.60, 1.18], power$_{(1-\beta)}$ = 1.00, 95% CI (1 − β) = [1.00, 1.00]}.

Likewise, children who performed above chance (n = 74, 38 female; M_{age} = 6;8, SD = 4.4 months, range = 5;6–7;7) on the 2AFC trials {M = .60, SD = .062, $t(73)$ = 13.49, p < .001, d = 1.57, 95% CI (d) = [1.23, 1.91]} also showed evidence of significant learning on the SICR measure {3 syllables: M_{diff} = .097, $t(72)$ = 4.65, p < .001, d = 0.54, 95% CI (d) = [0.30, 0.79], power$_{(1-\beta)}$ = .995, 95% CI (1 − β) = [.72, 1.00]; 6 syllables: M_{diff} = .17, $t(73)$ = 9.93, p < .001, d = 1.15, 95% CI (d) = [0.86, 1.45], power$_{(1-\beta)}$ = 1.0, 95% CI (1 − β) = [1.00, 1.00]}. Therefore, the data suggest that the SICR measure is more sensitive than 2AFC performance when measuring linguistic ASL. Thus, although many children were making incorrect decisions on the 2AFC trials, their learning of the sequences acquired during the exposure phase was still present and observable through sequence repetition in the SICR task.

Regression analyses

We next report the results from two regression analyses that investigated the relationships between our two measures of ASL—2AFC and SICR performance—and age. Table 1 presents the bivariate correlations between all variables.

Table 1 shows that 2AFC performance was significantly but weakly correlated with SICR performance on 3-syllable and 6-syllable targets but not on foils, and was also significantly correlated with age. All SICR measures correlated with each other. Finally, only performance on the SICR 6-syllable target items significantly correlated with age.

Predicting 2AFC performance. We first analyzed whether age and SICR performance explained overlapping or independent variance in 2AFC performance using regression. In these analyses, we used the log odds of SICR scores and the log of age, which allowed us to sensibly interpret the model output. We first estimated logistic regression models with 2AFC predicted by log age and the log odds of each SICR component separately. Log age was significant in each of the models; however, only 3-syllable target item performance significantly predicted ASL performance (χ^2 = 5.131, p = .024), whereas repetition of 6-syllable target items and 3-syllable and 6-syllable foils did not. Moreover, including *both* 3-syllable and 6-syllable target log odds into the regression model did not improve model fit. Table 2 presents the summary statistics for the best-fitting model.

Thus, we see that both age and 3-syllable SICR performance positively and independently predict 2AFC performance, although the association between 3-syllable SICR and 2AFC is weak. This is best demonstrated by the log odds ratio (.046), which implies that doubling the performance on 3-syllable SICR target items results in a modest increase in 2AFC performance (i.e., $2^{.046}$ = 1.032).

SICR performance. We next analyzed the relationship among sequence length (3 vs. 6 syllables), item type (target vs. foil), and age. We fit the data using a generalized estimating equations repeated-measures logistic regression. The parameters of the best model, based on goodness of fit, are shown in Table 3.

Table 3 shows significant main effects for item type and sequence length, which were subsumed by a significant item by length interaction, confirming that the difference between target and foil repetition was greater in the 6-syllable sequence condition than in the 3-syllable sequence condition. A significant age by item interaction effect suggested that the learning effect (i.e., difference between target items and foils) increased with age.

Discussion

In Study 1, children aged 5;6 to 7;7 completed an embedded triplet task where learning occurred during familiarization and was measured subsequently by 2AFC test trials and serial recall, based on studies that have used repetition to index sequence learning (e.g., Isbilen et al., 2017, 2020; Majerus et al., 2004) in addition to studies of Hebbian learning (e.g., Hebb, 1961; Page & Norris, 2009). Some previous studies of linguistic ASL have suggested that 2AFC is not sensitive to children's performance in this age range (e.g., Arnon, 2020; Raviv & Arnon, 2018), and thus there is a need for measures that more adequately measure linguistic ASL in this age group. The promise of the repetition paradigm was borne out; we observed stronger effects in our repetition task compared with 2AFC trials, and we found that the SICR task revealed significant learning even in a subgroup of children who were performing at or below chance on the 2AFC task. Thus, we can conclude that the SICR paradigm is a significantly more sensitive measure of children's linguistic ASL than 2AFC.

We observed several results of note. A key finding was that 2AFC performance and SICR performance were only weakly related. The overlap between the two measures likely reflects ASL-related changes to phonological memory representations for the linguistic stimuli (Majerus et al., 2004), but the weak association may reflect differences in the precision with which reflection and processing measures tapped into the learned material. This is consistent with our post hoc power analysis, where we found that learning as measured via 2AFC had a large confidence interval around the computed power, suggesting that the estimate of the effect cannot be determined with high precision despite the large sample size. By contrast, SICR fared much better, especially the 6-syllable component. We interpret this to indicate that SICR better tapped the implicitly acquired material.

In a regression analysis of SICR performance, we also found that the learning effect (a) increased with age and (b) interacted with sequence length. Whereas age-related changes have been shown in visual statistical learning and ASL of nonlinguistic stimuli (Arciuli & Simpson, 2011; Raviv & Arnon, 2018; Shufaniya & Arnon, 2018), evidence for age-related changes in linguistic ASL is mixed and potentially measure-dependent (Raviv & Arnon, 2018; Smalle et al., 2017, 2018). The multicomponent nature of SL means that the developmental effect we observed could be due to development in several subcomponents of learning as well as the quality of children's linguistic representations (Arciuli, 2017). Thus, although the capacity to identify statistical relationships between linguistic units such as syllables is evident early in life, variables that influence the ability to find those relationships and influence SL in general, such as processing speed/fluency (Arciuli & Simpson, 2011) and the entrenchment of linguistic representations (Jost & Christiansen, 2016; Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018), are still subject to developmental change and could have contributed to the developmental effect we observed.

The item type by length interaction is consistent with models of linguistic ASL that rely on chunking processes to extract word-like units from speech (e.g., Jones, 2012; Perruchet & Vinter, 1998;

Table 3
Model parameters for analysis of SICR performance as predicted by age (months), item type (target vs. foil), and item length (3- vs. 6 syllables).

	β	*SE*	95% CI (β)	*z*	*p*
Intercept	−2.26	0.69	[−3.59, −0.93]	−3.33	<.001***
Age	0.01	0.009	[−0.003, 0.03]	1.56	.12
Item	−1.99	0.82	[−3.60, −0.38]	−2.42	.016*
Length	1.80	0.06	[1.67, 1.92]	27.73	<.001***
Item * Length	−0.28	0.09	[−0.45, −0.11]	−3.25	.001**
Age * Item	0.03	0.01	[0.014, 0.054]	3.29	.001**

Note. SICR, statistically induced chunking recall; CI, confidence interval.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Robinet, Lemaire, & Gordon, 2011), which compress reoccurring sequences of information and store them in long-term memory as higher-order units (e.g., storing sequences of syllables as words), thereby making their processing more efficient given the computational demands of serial recall (Cowan, 2001; Miller, 1956). Therefore, one potential explanation of the effect is that children's successful extraction of regularities from the speech stream allowed children to process larger sequences of information, such that the 6-syllable items, which are typically very difficult for children this age (Gathercole, 2006), become more tractable due to the presence of acquired subsequences in long-term memory (Christiansen, 2019; Majerus et al., 2004).

Study 2

In Study 2, we aimed to replicate the effects found in Study 1 in addition to determining the test–retest reliability of the 2AFC and SICR tasks. There is an increasingly growing literature on individual differences in SL and how they relate to other cognitive domains such as spoken language and literacy. Some studies have reported significant associations between measures of SL and children's natural language and literacy development (e.g., Arciuli & Simpson, 2012; Frost et al., 2020; Kidd & Arciuli, 2016; Qi et al., 2019; Torkildsen et al., 2019), whereas others have not (e.g., Lammertink, Boersma, Rispens, & Wijnen, 2020; Mimeau, Colman, & Donlan, 2016). Part of this inconsistency is likely due to the psychometric properties of the tasks used. Notably, Arnon (2020) recently reported that measuring linguistic ASL via 2AFC resulted in rather poor reliability. SICR has been shown to have good test–retest reliability in adults (Isbilen et al., 2017, 2020); here we investigated its psychometric properties (internal consistency, test–retest reliability) in children and compared it with 2AFC.

Method

Participants

Determining a sample size for Study 2 was complicated by the variability in effect sizes across the dependent measures in Study 1 (0.24 < *d*s < 1.02). Our primary goal was to replicate the effects found for the SICR measure and determine test–retest reliability of the task; therefore, our target sample size was determined by a combination of (a) SICR effect sizes from Study 1 and (b) a few comparable past studies. Our lower estimate for a sample size was N = 10 (based on 6-syllable SICR, d = 1.02, dependent-samples t test, alpha = .05, power = .80, calculated using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007), and our upper estimate was N = 57 (based on 3-syllable SICR, d = 0.38). We aimed for an approximate midpoint here of N = 40, which was still more than adequate if we used the upper limit of the estimate of d = 0.55 for the 3-syllable SICR (N = 28). This number is comparable to Arnon's (2020) test–retest study of linguistic ASL as measured via 2AFC, which tested 36 to 41 children. It is also comparable to Isbilen et al.'s (2020) adult test–retest study, which tested 42 participants using 2AFC and SICR.

Our final sample comprised 37 children (22 female) who were recruited from one primary school in Canberra. The area in which the school was located was in the 84th percentile for the SEIFA score nationally but in the 29th percentile for the ACT. The same inclusion criteria as per Study 1 were used. All children were in the second year of formal schooling (Grade 1) and had a mean age of 6;9 (*SD* = 3. 65 months, range = 6;2–7;5). None had participated in Study 1.

Materials and procedure

Children were tested on the same ASL task used in Study 1, with two modifications made to the SICR component. First, the number of 3-syllable target and foil trials was doubled to 16 (i.e., each target and foil was tested twice), such that there were an equal number of 3-syllable and 6-syllable sequences tested. Second, the 3-syllable and 6-syllable items were intermixed, although each list began with four 3-syllable trials (two targets and two foils). Children were tested on the ASL task twice, with an average of 8.89 days (range = 6–12) in between testing sessions, but were tested on different counterbalanced versions of the test across the two sessions.[3] The SICR data of 4 children (11% of the sample) were double-coded for reliability, which showed substantial agreement for both target items (Cohen's kappa = .75) and foils (Cohen's kappa = .82).

Results

One child was absent at Time 2 and thus contributed data to analyses at Time 1 only. The general results from Study 1 replicated across both time points. The 2AFC data showed that children performed above chance on both occasions {Time 1: $M = .533$, $SD = .097$, $t(36) = 2.05$, $p = .048$, $d = 0.34$, 95% CI (d) = [0.003, 0.67]; Time 2: $M = .57$, $SD = .072$, $t(35) = 5.89$, $p < .001$, $d = 0.98$, 95% CI (d) = [0.58, 1.38]}. The difference in children's performance across Time 1 and Time 2 was not statistically significant, $t(35) = 1.987$, $p = .055$, $d = 0.33$, 95% CI (d) = [−0.007, 0.67].

Fig. 3 presents pirate plots for the SICR data by item type and sequence length for Time 1 and Time 2 testing.

Fig. 3 shows a very similar pattern across the two time points. In both cases, children repeated a greater proportion of target sequences in comparison with foils, and this difference appeared to be larger for the 6-syllable sequences. Once again, the basic learning effects replicated across both time points {Time 1: 3 syllables, $M_{\text{diff}} = .063$, $SD = .17$, $t(34) = 2.25$, $p = .031$, $d = 0.38$, 95% CI (d) = [0.03, 0.72]; 6 syllables, $M_{\text{diff}} = .15$, $SD = .15$, $t(34) = 6.14$, $p < .001$, $d = 1.04$, 95% CI (d) = [0.62, 1.45]; Time 2: 3 syllables, $M_{\text{diff}} = .05$, $SD = .13$, $t(34) = 2.18$, $p = .036$, $d = 0.37$, 95% CI (d) = [0.02, 0.71]; 6 syllables, $M_{\text{diff}} = .22$, $SD = .16$, $t(34) = 8.01$, $p < .001$, $d = 1.35$, 95% CI (d) = [0.89, 1.81]}.

We repeated our sensitivity analysis from Study 1, comparing those children who performed at or below chance on the 2AFC task ($n = 14$; $M_{\text{age}} = 6;9$, $SD = 3.8$ months; 9 female) at Time 1 with those who performed above chance ($n = 23$; $M_{\text{age}} = 6;9$, $SD = 3.7$ months; 13 female). We ran these analyses only on the Time 1 data because there were too few children at or below chance at Time 2 to run any inferential statistics ($n = 6$). The findings from Study 1 replicated; children who performed poorly on the 2AFC task were significantly below chance {$M = .44$, $SD = .04$, $t(13) = -6.51$, $p < .001$, $d = -1.50$, 95% CI (d) = [−2.60, −0.90]} but showed evidence of learning on the SICR task, particularly the 6-syllable condition {3 syllables: $M_{\text{diff}} = .04$, $SD = .09$, $t(12) = 1.65$, $p = .13$, $d = 0.46$, 95% CI (d) = [−0.13, 1.02]; 6 syllables: $M_{\text{diff}} = .16$, $SD = .14$, $t(12) = 4.13$, $p = .001$, $d = 1.14$, 95% CI (d) = [0.42, 1.84]}. The SICR results of the above-chance group {2AFC: $M = .59$, $SD = .07$, $t(22) = 6.22$, $p < .001$, $d = 1.30$, 95% CI (d) = [0.73, 1.85]} were qualitatively indistinguishable {3 syllables: $M_{\text{diff}} = .08$, $SD = .20$, $t(21) = 1.80$, $p = .09$, $d = 0.38$, 95% CI (d) = [−0.05, 0.81]; 6 syllables: $M_{\text{diff}} = .15$, $SD = .15$, $t(21) = 4.50$, $p < .001$, $d = 0.96$, 95% CI (d) = [0.44, 1.46]}. Thus, we once again observed that SICR can provide evidence of learning during familiarization in cases where 2AFC does not.

We next tested whether the item type by sequence length interaction found in Study 1 replicated in Study 2. We analyzed the data using generalized estimating equations, including the fixed effects of

[3] One child was tested on the same version of the test across the two time points, but the results did not change when this child was removed, and so the data were kept in the final analyses.

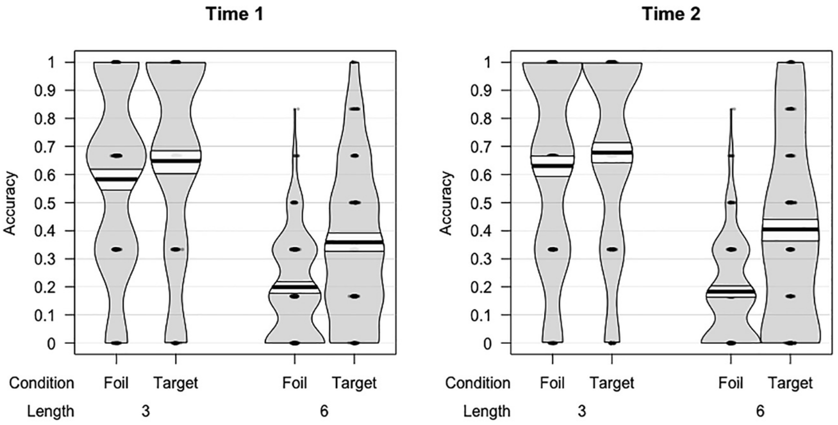

Fig. 3. Pirate plots depicting children's accuracy during the SICR (statistically induced chunking recall) test phase at Time 1 (left panel) and Time 2 (right panel) of testing by item type (target vs. foil) and sequence length (3- vs. 6-syllables).

(a) item type (target vs. foil), (b) sequence length (3 vs. 6 syllables), age (in months), and test session (Time 1 vs. Time 2). Age and test session and their interactions with other variables did not contribute to the model and so were removed. The parameter estimates for the final model are shown in Table 4.

Table 4 shows that the item type by length interaction effect was replicated in the same direction as in Study 1. Notably, the effect did not interact with test session, suggesting that the learning effect is stable across the two testing sessions.

We next analyzed some of the psychometric properties of the two dependent measures, which we estimate by computing (a) the internal consistency at each time point, as measured by Cronbach's alpha (α), and (b) the test–retest reliability of the of the 2AFC and SICR measures. Although the data from both the 2AFC and SICR measures at the item level are proportions, in reality they still form a discrete variable with a restricted number of categories. Therefore, we report two values for each measure at each time point: α based on nonparametric Kendall's tau (τ) and α based on Pearson's *r*, with the former being a more conservative estimate. The generally accepted interpretation of α is as follows: questionable, $.60 \leq \alpha < .70$; acceptable, $.7\ 0 \leq \alpha < .80$; good, $.80 \leq \alpha < .90$; excellent, $.90 \leq \alpha < 1.00$. The results are presented in Table 5.

Table 5 shows poor internal consistency for the 2AFC task; in fact, at Time 2 the alphas are negative, suggesting especially poor interitem correlations. By contrast, the SICR target measures show a more promising pattern of internal consistency across Time 1 and Time 2. Notably, for the 6-syllable targets we see acceptable to good reliability in three of four estimates. We find the same results when the two measures (3-syllable and 6-syllable SICR targets) are combined (SICR overall).

We next determined the test–retest reliabilities of the 2AFC and SICR measures. We report both aggregate reliabilities (i.e., participant performance averaged across items —the most commonly reported test–retest statistic in this space; see Arnon, 2020; Siegelman & Frost, 2015) and item-level test–retest reliability (i.e., correlations between items across time). The Spearman–Brown "prophecy" formula (Brown, 1910; Spearman, 1910) predicts that item-level reliability will be lower than aggregate-level reliability,[4] and thus we can expect high aggregate test–retest reliability when test–retest reliability at the item level is moderate.

For the 2AFC, test–retest reliability was low at the aggregate and item levels. The aggregate correlation across the time points was almost zero ($r = .053$; see Fig. 4). The average item-level correlation was $r = .073$. By contrast, aggregate-level test–retest reliability for SICR was high (3-syllable target items: $r = .79$; 6-syllable target items: $r = .79$), as shown in Fig. 5.

[4] $r_t = \frac{nr_i}{1+(n-1)r_i}$, where r_t is aggregate test–retest reliability, n is the number of items being aggregated, and r_i is the test–retest reliability of an item.

Table 4
Parameter estimates for final model predicting SICR performance in Study 2.

	B	*SE*	95% CI (β)	*z*	*p*
Intercept	−1.60	0.12	[−1.85, −1.36]		
Item	1.13	0.14	[0.85, 1.41]	7.96	<.001
Length	1.76	0.13	[1.50, 2.02]	13.30	<.001
Item * Length	−0.74	0.094	[−0.92, −0.56]	−7.87	<.001

Note. SICR, statistically induced chunking recall; CI, confidence interval.

Table 5
Internal consistency statistics for ASL 2AFC and SICR measures across time.

	Time 1		Time 2	
	Kendall's τ	Pearson's *r*	Kendall's τ	Pearson's *r*
ASL 2AFC	−.05	−.04	−.49	−.88
SICR 3 syllables	.47	.58	.60	.72
SICR 6 syllables	.66	.71	.74	.82
SICR overall	.75	.80	.83	.88

Note. ASL, auditory statistical learning; 2AFC, two-alternative forced-choice; SICR, statistically induced chunking recall.

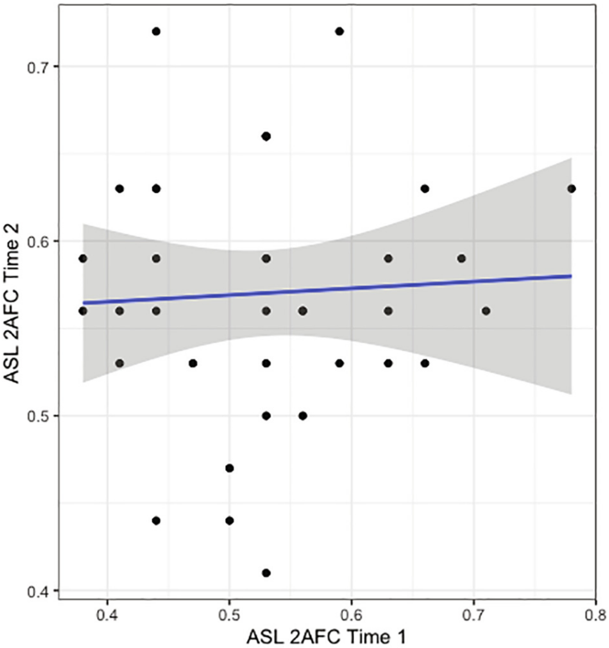

Fig. 4. Scatterplot depicting the association between performance on the 2AFC (two-alternative forced-choice) component of the ASL (auditory statistical learning) task at Time 1 and that at Time 2 of testing. The blue line denotes the least-squares line of best fit.

However, the SICR task heavily implicates short-term phonological working memory, which we can estimate by looking at foil repetition, which was also relatively stable across time (3 syllables: $r = .72$; 6 syllables: $r = .57$; for all bivariate correlations, see Appendix C). We tested whether the correlations across the target items remained significant after controlling for foil repetition in the same

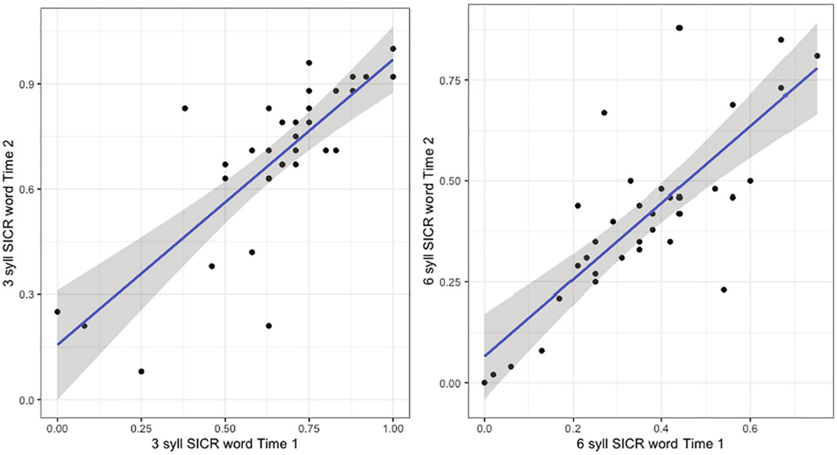

Fig. 5. Scatterplots depicting the association between performance on the SICR (statistically induced chunking recall) component (left panel: 3 syllables; right panel: 6 syllables) of the ASL (auditory statistical learning) task at Time 1 and that at Time 2 of testing. The blue lines denote the least-squares lines of best fit.

condition at Time 1 (a proxy for short-term memory for random syllables) and age (in months). In both cases, the partial correlations were lower but still statistically significant (3 syllables: r = .57, df = 31, p < .001; 6 syllables: r = .65, df = 31, p < .001).

Average item-level reliabilities for SICR were lower, with a slight difference according to syllable length (3 syllables: r = .22; 6 syllables; r = .33). Foil repetition also showed similar item-level reliability (3 syllables: r = .20; 6 syllables: r = .21).

Discussion

In Study 2, we aimed to replicate the results found in Study 1 and determine the reliability of the 2AFC and SICR in our task. We observed significant learning effects across the 2AFC and SICR measures, with similar effect sizes to those observed in Study 1. The one exception was the 2AFC measure at Time 2, which yielded a large effect size in comparison with the smaller effects observed in Study 1 and at Time 1 of Study 2. By contrast, the effect sizes for the SICR measures were stable across the two time points. Although we can only speculate at this point, 2AFC may be more susceptible to practice effects, which is also what Arnon (2020) found in one test–retest study that used linguistic stimuli. In this instance, the second exposure to the stimuli may improve participants' ability to make familiarity judgments about the learned material, resulting in more participants performing above chance (although somewhat unevenly given that the poor test–retest reliability suggests a nonuniform improvement across participants). In line with this suggestion, we found that 2AFC performance at Time 2, but not at Time 1, was significantly correlated with children's repetition of target items on the SICR task. Specifically, correlations between SICR 3-syllable and 6-syllable target repetition at Time 1 and Time 2 and 2AFC at Time 1 were low and nonsignificant (–.05 < r < .06), whereas for 2AFC at Time 2 the same correlations all were positive and larger in magnitude (.24 < r < .51) and with one exception were significantly different from 0 (see Appendix C).

The item type by sequence length interaction effect in the SICR task was replicated across both testing sessions. As in Study 1, we found that the learning effects from the familiarization phase were stronger for sequences of greater length. The interaction suggests that storage of learned items eases the processing of larger amounts of linguistic input because it can be processed as larger encoded chunks rather than a long string of smaller units (i.e., a random 6-syllable sequence), as is characteristic of verbal working memory (e.g., Acheson & MacDonald, 2009; Baddeley et al., 1998; Majerus & D'Argembeau, 2011; Majerus & van der Linden, 2003; Majerus et al., 2004; see Christiansen, 2019, for discussion). Because SICR indexes the output of linguistic ASL, the results provide evidence in favor of chunking models of SL (Perruchet & Vinter, 1998; Robinet et al., 2011), which predict that the output of SL is chunk-based information stored in long-term memory that can be rapidly accessed and used to process larger sequences of subsequent information. We consider the broader implications of the effect in relation to theories of language acquisition in the General Discussion.

Our second important result concerned the reliability observed in the SICR task, which for several measures (i.e., 6-syllable target and overall recall) reached or approached psychometric standards for internal consistency and also had high test–retest reliability. Importantly, the test–retest reliability was still of moderate magnitude when we controlled for children's age and their foil repetition, suggesting that the task indexes children's learning of the specific statistical structure of the sequence above and beyond age and memory for random syllable sequences (i.e., their short-term memory). Item-level reliability was lower than aggregate-level reliability. Although expected, the difference is likely to indicate individual variability (i.e., individuals perform differently on items across time, possibly due to a combination of random noise and differences in learning; see Siegelman, Bogaerts, Armstrong, & Frost, 2019) and item-level variability (e.g., individuals learning different items to varying degrees across the sessions). Aggregate-level performance is commonly used to measure individual differences in ASL, and thus its high reliability in children is notable, consistent with the adult version of the task (Isbilen et al., 2017, 2020). Thus, the task holds promise as a measure of individual differences in linguistic ASL.

General discussion

There has been considerable recent interest in the robust measurement of SL in both developmental and adult populations, notably stemming from a focus on individual differences. Although this approach has resulted in some success, progress has been hampered in the linguistic domain by shortcomings in task reliability (e.g., Arnon, 2020; Siegelman & Frost, 2015). In the current article, we have presented the results of a task that uses serial recall to measure SL of co-occurring linguistic elements. In contrast to reflection-based measures of linguistic ASL, SICR produced consistent and reliable results across and within individuals.

Measuring linguistic ASL through serial recall has several advantages. First, it takes advantage of established empirical links between speech perception and production across the lifespan (e.g., Glanz et al., 2018; Vihman, 2017; Vilain et al., 2019; Wilson et al., 2004). Second, the method has a long tradition in research on verbal memory and learning. Importantly, this means that the dependent measure is the output of SL, with the potential to reveal novel insights into the computations underlying the mechanism and how these computations might in part support the acquisition of language.

One example along this line of reasoning is the item type by sequence length interaction we found across both studies, which showed that the learning effect was greater for longer sequences of syllables. The result is reminiscent of a classic effect in psycholinguistics, whereby memory for semantically meaningful grammatical sentences is superior to memory for semantically anomalous but grammatical sentences, which in turn is superior to memory for random strings of words (Miller & Isard, 1963). This original result was interpreted to reflect the psychological existence of grammatical (and "semantic") rules; the suggestion was that serial memory for grammatical sequences was greater than that for random word lists because we possess implicit rules denoting how grammatical and semantic units combine.

We assume that the children in our studies did not begin the experiment with any prior knowledge of the specific distributional regularities in the input stream; they needed to rapidly acquire them via SL. In this respect, the item type by sequence length interaction is consistent with Christiansen and Chater's (2016) now-or-never (NoN) bottleneck theory of language acquisition and language processing. The central tenet of the NoN bottleneck is that the learner is constrained by limited memory resources but must rapidly process large amounts of linguistic input "in the moment." The learner does so by "chunk-and-pass" processing in which chunks of linguistic information at lower levels of description (e.g., phonemes, syllables) are compressed and recoded into higher-level chunks (e.g., as words). The chunking process occurs over frequently co-occurring units, as in the co-occurring syllables in our target items, allowing faster processing of subsequent input. Thus, one interpretation of our data is that a driver of SL is the identification and storage of co-occurring syllables as compressed chunks, as might be described by bigrams (e.g., *loma, mari*) or trigrams (e.g., *lomari*). These chunks provide a small advantage to children when they repeat 3-syllable targets versus foils but are particularly useful when the processing system is overburdened with information, as in the 6-syllable condition (for an account of SL that identifies an important role for capacity limitations, see Frank, Goldwater, Griffiths, & Tenenbaum, 2010).

Chunking is not the only possible explanation of the data. The exact mechanism underlying the statistical computations is a matter of ongoing debate in the literature. A broad division can be drawn between those approaches that make use of TPs to segment and those approaches that use chunking (for an overview, see Frank et al., 2010). Although we have called our repetition task "chunking recall," this denotation describes the general tendency for humans to group co-occurring elements in memory (as in serial recall). Thus, our data do not unambiguously provide evidence for one process over the other given that it is in principle possible that the learning effect derives from knowledge of transitional probabilities, chunked sequences, or both. Interestingly, recent work on visual SL in adults has shown that individuals vary in their reliance on TPs and chunking (Siegelman et al., 2019), and it is possible that this flexibility is present in the auditory domain. We note, however, that several studies suggest that chunking models provide a better fit to adult human data (e.g., Frank et al., 2010; French, Addyman, & Mareschal, 2011; Perruchet, Poulin-Charronnat, Tillman, & Peereman, 2014). It may be possible that both sensitivity to TPs and storage of chunks might take place in parallel, as implemented in the computational model of early language acquisition by McCauley and Christiansen (2019). Their chunk-based learner uses backward TPs to discover multiword chunks, which are used to capture aspects of language comprehension and production cross-linguistically. Future research is needed to determine the degree to which different individuals may rely differentially on TPs and/or chunks and whether this might change across development.

As with any task, there are also some limitations. For example, although the SICR task holds promise as a measure of individual differences, it would not be useful for individuals who have significant speech production difficulties. In addition, the version of our SICR task that we report here uses syllables and phonemes from English; different versions of the SICR task would need to be developed for speakers of languages other than English. Moreover, the data we report here pertain only to reliability and not to other psychometric properties such as predictive validity, which is of great interest for investigations of how individual differences in SL are linked with development in cognitive domains such as spoken language and literacy.

Conclusion

SL features prominently in many developmental theories of cognitive processes. One promise of SL is that it provides a powerful mechanism by which children may both analyze and abstract over their input. In the current study, we built on previous work (e.g., Isbilen et al., 2017, 2020; Majerus et al., 2004), measuring linguistic ASL via serial recall, and showed that, in young children, it is a more sensitive and reliable measure than 2AFC. Notably, because the SICR measure showed good reliability at the individual level, the measure holds promise for future studies of individual differences in linguistic ASL. We hope that this SICR measure will be useful in developmental studies of ASL.

Acknowledgments

We thank Shanthi Kumarage and Tanya Price for help with testing and thank Steven Hecht, Gary Jones, and two anonymous reviewers for helpful comments. This research was supported by an Australian Research Council grant (DP160101917) awarded to E.K., J.A., and M.S. The order of authorship following the first author is alphabetical.

Appendix A. Target triplets and foils

Target triplets: kibudu, modipa, takapo, lomari
2AFC foils: kaburi, dilota, makidu, popamo

SICR 3-syllable foils: ribuki, molopa, bumaka, tadipo

SICR 6-syllable targets: takapolomari, kibudutakapo, modipakibudu, lomaritakapo, lomarimodipa, takapokibudu, modipalomari, kibudumodipa

SICR 6-syllable foils: kipabumodudi, pobutakikadu, dupotabukaki, ripomataloka, bupadidukimo, lomomaporidi, dimamoparilo, tarimalopoka

Appendix B. No evidence of learning across 2AFC and SICR tasks

An anonymous reviewer suggested that because the foils changed across the 2AFC and SICR phases, children may continue to learn the target items throughout the task. Here we present two analyses of performance on the 2AFC and SICR tasks, demonstrating no evidence of learning across either measure.

2AFC

We compared children's performance on the first half of the 2AFC task (Trials 1–16: M = .530, SD = .137) with that on the second half of the task (Trials 17–32: M = .516, SD = .126). A paired-samples t test revealed no difference between the two test halves, $t(138)$ = 0.983, p = .327, d = 0.083, 95% CI (d) = [−0.08, 0.25]. This result suggests no evidence for learning across the 2AFC test trials. Because frequentist statistics do not provide evidence for the null hypothesis, we also conducted a Bayesian paired-samples t test, which revealed a Bayes factor of BF_{01} = 6.61, suggesting that the null hypothesis is 6.61 times more likely than the alternative hypothesis.

SICR

We next analyzed the 6-syllable SICR trials. This is the most stringent and best test of the hypothesis that children continue to learn across the test phases because (a) it was the final phase (i.e., it followed the 3-syllable phase) and (b) it was the only SICR phase in which the targets occurred more than once (i.e., the 3-syllable phase simply tested each target once, whereas during the 6-syllable phase each target occurred four times). Fig. B1 shows the average proportions correct for the 6-syllable target and foil trials.

Fig. B1 shows that whereas foil repetition was fairly constant across trial number, the repetition of the targets *decreased* as the 6-syllable phase progressed. If children continued to learn through the experiment, they should be improving in their repetitions of the targets relative to the foils, which predicts an item type by sequence interaction, where the difference between the targets and foil accuracy is larger for later trials. We tested this hypothesis by comparing accuracy performance on targets and foils in the first and second halves of the phase. We conducted a linear mixed-effects model in R Version 3.6.1 (R Core Team, 2014) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015), with item type (contrast coded: foil = − 0.5 vs. target = +0.5), experiment half (contrast coded: first = − 0.5 vs. second = +0.5), and their interaction entered as fixed effects, along with random intercepts for participants and items, and a random slope for word type in participants. Although the main effect of item type was significant, with children performing significantly better on targets than on foils (β = 0.15, SE = 0.034, t = 4.57, p < .05), the effect of experimental half was not significant (β = − 0.013, SE = 0.034, t = − 1.11, *ns*), and nor was the interaction (β = − 0.02, SE = 0.02, t = − 1.26, *ns*). Note that these last two results, although not significant, are in the opposite direction of what would be predicted if children were continuing to learn throughout the experiment.

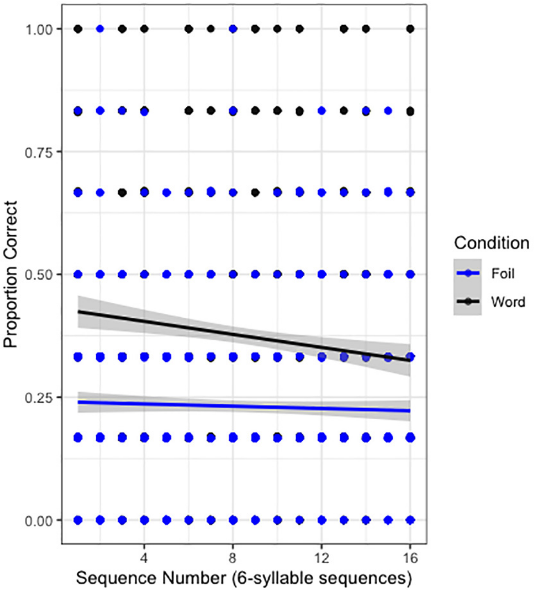

Fig. B1. Proportion SICR items correct (targets vs. foils) by sequence number for 6-syllable trials.

Appendix C. Bivariate correlation between 2AFC scores and SICR measures in Study 2.

	2AFC T1	2AFC T2	3-Syllable target T1	3-Syllable foil T1	6-Syllable target T1	6-Syllable foil T1	3-Syllable target T2	3-Syllable foil T2	6-Syllable target T2	6-Syllable foil T2
2AFC T1	–									
2AFC T2	.053	–								
3-Syllable target T1	.014	.329	–							
3-Syllable foil T1	−.015	.256	.691***	–						
6-Syllable target T1	.043	.491**	.719***	.626***	–					
6-Syllable foil T1	−.089	.309	.606***	.651***	.591***	–				
3-Syllable target T2	−.049	.419*	.798***	.735***	.710***	.659***	–			
3-Syllable foil T2	−.008	.158	.752***	.722***	.519**	.547***	.833***	–		
6-Syllable target T2	.019	.514**	.698***	.624***	.793***	.739***	.801***	.613***	–	
6-Syllable foil T2	.123	.322	.573***	.525**	.694***	.568***	.666***	.601***	.714***	–

Note. 2AFC, two-alternative forced-choice; SICR, statistically induced chunking recall; T1, Time 1; T2, Time 2.

$^{*}p$ < .05.

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

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

References

Acheson, D. J., Hamidi, M., Binder, J. R., & Postle, B. R. (2011). A common neural substrate for language production and verbal working memory. *Journal of Cognitive Neuroscience, 23*, 1358–1367.

Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin, 135*, 50–68.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*, 20160058.

Arciuli, J. (2018). Reading as statistical learning. *Language, Speech, and Hearing Services in Schools, 49*, 634–643.

Arciuli, J., & Conway, C. (2018). The promise and challenge of statistical learning for elucidating atypical language development. *Current Directions in Psychological Science, 27*, 492–500.

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.

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

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods, 52*, 68–81.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex, 90*, 31–45.

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

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.

Bogaerts, L., Szmalec, A., De Maeyer, M., Page, M. P. A., & Duyck, W. (2016). The involvement of long-term serial-order memory in reading development: A longitudinal study. *Journal of Experimental Child Psychology, 145*, 139–156.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition, 121*, 127–132.

Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11*, 468–481.

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

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

Cowan, N. (2001). The magical number four in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 21*, 87–185.

Emberson, L. L., Misyak, J. B., Schwade, J., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science, 22*, e12847.

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

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

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

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in human speech segmentation. *Cognition, 117*, 107–125.

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

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

Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., ... Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology, 120*, 101291.

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

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

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The Children's Test of Non-Word Repetition: A test of phonological working memory. *Memory, 2*, 103–127.

Glanz, O., Derix, J., Kaur, R., Schulze-Bonhage, A., Auer, P., Aertsen, A., & Ball, T. (2018). Real-life speech production and perception have a shared premotor-cortical substrate. *Scientific Reports, 8*, 8898.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). Oxford, UK: Blackwell.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). In *Testing statistical learning implicitly: A novel chunk-based measure of statistical learning* (pp. 564–569). Austin, TX: Cognitive Science Society.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically-induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science, 44*, e12848.

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

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

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

Jost, E., & Christiansen, M. H. (2016). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment, memory and automaticity: The psychology of linguistic knowledge and language learning* (pp. 227–244). Boston: Walter de Gruyter.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87*, 184–193.

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes, 22*, 860–897.

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.

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing*. https://doi.org/10.1007/s11145-020-10018-4. Advance online publication.

Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics, 40*, 279–302.

Majerus, S., & D'Argembeau, A. (2011). Verbal short-term memory reflects the organization of long-term memory: Further evidence from short-term memory for emotional words. *Journal of Memory and Language, 64*, 181–197.

Majerus, S., & van der Linden, M. (2003). Long-term memory effects on verbal short-term memory: A replication study. *British Journal of Developmental Psychology, 21*, 303–310.

Majerus, S., van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language, 51*, 297–306.

Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R., & Arciuli, J. (2017a). Musicians' online performance during visual and auditory statistical learning tasks. *Frontiers in Neuroscience, 11*. https://doi.org/10.3389/fnhum.2017.00114.

Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R., & Arciuli, J. (2017b). Statistical learning and auditory processing in children with music training: An ERP study. *Clinical Neurophysiology, 128*, 1270–1281.

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

McKelvie, S. J. (1987). Learning and awareness in the Hebbian digits task. *Journal of General Psychology, 114*, 75–88.

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

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2*, 217–228.

Mimeau, C., Colman, M., & Donlan, C. (2016). The role of procedural memory in grammar and numeracy skills. *Journal of Cognitive Psychology, 28*, 899–908.

Mosse, E. K., & Jarrold, C. (2008). Hebb learning, verbal short-term memory, and the acquisition of phonological forms in children. *Quarterly Journal of Experimental Psychology (Hove, England), 61*, 505–514.

Norris, D., Page, M. P. A., & Hall, J. (2018). Learning nonwords: The Hebb repetition effect as a model of word learning. *Memory (Hove, England), 26*, 852–857.

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 3737–3753.

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

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

Phillips, N. D. (2018). YaRrr! The pirate's guide to R. Available at: https://bookdown.org/ndphillips/YaRrr/.

Potter, M. C., & Lombardi, L. (1990). Regeneration in short-term recall of sentences. *Journal of Memory and Language, 29*, 633–654.

Qi, Z., Sanchez, Y., Georgan, W., Gabrieli, J., & Arciuli, J. (2019). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*, 101–115.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21* e12593.

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

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDL chunker: A MDL-based cognitive model of inductive learning. *Cognitive Science, 35*, 1352–1389.

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

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

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

Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? *Insights from Bayesian modeling. Cognition, 192*, 104002.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring indiviudal 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., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213.

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

Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M. G., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01925.

Smalle, E. H., Muylle, M., Szmalec, A., & Duyck, W. (2017). The different time course of phonotactic constraint learning in children and adults: Evidence from speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1821–1827.

Smalle, E. H., Page, M. P. A., Duyck, W., Edwards, M., & Szmalec, A. (2018). Children retain implicitly learned phonological sequences better than adults: A longitudinal study. *Developmental Science, 21*, e12634.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp 221–257). Mahwah, NJ: Lawrence Erlbaum.

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

Szmalec, A., Page, M. P. A., & Duyck, W. (2012). The development of long-term lexical representations through Hebbian repetition learning. *Journal of Memory and Language, 67*, 342–354.

Teinonen, T., Felham, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical learning in neonates revealed by event-related potentials. *BMC Neuroscience, 10*, 21.

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

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology, 108*, 1–27.

Vilain, A., Dole, M., Loevenbruck, H., Pascalis, O., & Schwartz, J. (2019). The role of production abilities in the perception of consonant categories in infants. *Developmental Science, 22*, e12830.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701–702.