Cognitive Development 66 (2023) 101335

Contents lists available at ScienceDirect

Cognitive Development

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

The sources and consequences of individual differences in statistical learning for language development

Evan Kidd [a,b,c,*], Joanne Arciuli [c,d], Morten H. Christiansen [c,e], Michael Smithson [f]

[a] Max Planck Institute for Psycholinguistics, the Netherlands
[b] School of Literature, Languages, and Linguistics, The Australian National University, Australia
[c] The ARC Centre of Excellence for the Dynamics of Language, Australia
[d] Caring Futures Institute, Flinders University, Australia
[e] Cornell University, United States
[f] College of Health and Medicine, The Australian National University, Australia

ARTICLE INFO

Keywords:
Language development
Statistical learning
Longitudinal

ABSTRACT

Statistical learning (SL)—sensitivity to statistical regularities in the environment—has been postulated to support language development. While even young infants are capable of using distributional statistics to learn in linguistic and non-linguistic domains, efforts to measure SL at the level of the individual and link it to language proficiency in individual differences designs have been mixed, which has at least in part been attributed to problems with task reliability. In the current study we present the first prospective longitudinal study of the relationship between both non-linguistic SL (measured with visual stimuli) and linguistic SL (measured with auditory stimuli) and language in a group of English-speaking children. One-hundred and twenty-one (N = 121) children in their first two years of formal schooling (M_{age} = 6;1 years, Range: 5;2 – 7;2) completed tests of visual SL (VSL) and auditory SL (ASL) and several control variables at time 1. Both forms of SL were then measured every 6 months for the next 18 months, and at the final testing session (time 4) their language proficiency was measured using a standardised test. The results showed that the reliability of the SL tasks increased across the course of the study. A series of path analyses showed that both VSL and ASL independently predicted individual differences in language proficiency at time 4. The evidence is consistent with the suggestion that, when measured reliably, an observable relationship between SL and language proficiency exists. Theoretical and methodological issues are discussed.

Humans have a powerful ability to rapidly learn the distributional regularities present in their environment. This skill – *statistical learning* (SL) – has been argued to be part of the multifaceted toolkit children bring to the problem of language learning (Lidz & Gagliardi, 2015; Saffran et al., 1996; Tomasello, 2003). While early research established the existence and parameters under which infants and children successfully exhibit SL (for reviews see Romberg & Saffran, 2010; Saffran & Kirkham, 2018), a separate but more recent line of research has investigated individual differences in SL and how they might relate to natural language proficiency (see Arciuli, 2017; Arciuli & Conway, 2018; Kidd et al., 2018; Siegelman, 2020; Siegelman et al., 2017). Guided by the assumption that SL as a skill systematically varies amongst individuals *and* that this variation explains significant variance in language proficiency,

* Correspondence to: School of Literature, Languages and Linguistics, College of Arts and Social Sciences, The Australian National University, Acton 2601, Australian Capital Territory, Australia.
E-mail address: evan.kidd@anu.edu.au (E. Kidd).

https://doi.org/10.1016/j.cogdev.2023.101335
Received 4 August 2022; Received in revised form 20 April 2023; Accepted 1 May 2023
Available online 14 May 2023

research in this space has yielded mixed results, which is attributable to several related problems associated with measuring SL as an individual capacity (Arciuli & Conway, 2018; Bogaerts et al., 2022). In the current paper we bring much needed longitudinal data to bear on this problem (Arciuli & Von Koss Torkildsen, 2012), which allowed us to track development in both ASL and VSL across the early school years and determine if and how learning in these two domains relates to language proficiency in middle childhood.

1. Individual differences in SL and their relationship to language

Although it was once assumed that implicit learning processes that characterize SL were not subject to individual variation (Reber, 1993), the weight of current evidence suggests the existence of measurable individual differences in SL. For instance, Kaufman et al. (2010) showed that individual differences in probabilistic sequence learning, as measured by the alternating serial reaction time task, were selectively related to processing speed and verbal analogical reasoning and not to other components of IQ and cognitive skill (e.g., Working Memory, WM). While it is true that participants vary in their performance on SL tasks, precisely how and why is unclear. It was originally (perhaps implicitly) assumed that SL constituted a single domain-general skill that computed distributional statistics across perceptual domains (e.g., visual, auditory); however, it seems likely that SL is a multi-componential process. Such issues were explored by Arciuli and Simpson (2011), who suggested "… a learning mechanism that plays a key role in language processing ought to show a degree of participant- and task-related variability, especially during childhood (p. 464–465)", going on to argue that "SL may be a multi-componential facility with some components maturing more gradually than others" (p. 471). Indeed, the lack of cross-domain associations relating to a variety of SL tasks points to a multi-componential process that operates differently depending on the type of perceptual input and the distributional properties to be learned (Arciuli, 2017; Arciuli & Conway, 2018; Bogaerts et al., 2022; Christiansen, 2019; Frost et al., 2015; Misyak & Christiansen, 2012; Siegelman & Frost, 2015; Siegelman et al., 2018).

More recent work on individual differences in SL has focused on its measurement in behavioural tasks. Adult research has shown that, while performance on different SL tasks is typically uncorrelated, performance within tasks is relatively reliable within individuals (Isbilen et al., 2020; Kalra et al., 2019; Perfors & Kidd, 2022; Siegelman et al., 2017; Siegelman & Frost, 2015), raising the possibility that an individual's aptitude for SL is in important ways linked to the distributional and modality-specific properties of each task (Arciuli & Conway, 2018; Bogaerts et al., 2022).

Task reliability in developmental populations appears more problematic. Notably, several tasks have been shown to have poor reliability on several psychometric indices (e.g., internal consistency, test-retest reliability, see Arnon, 2020; West et al., 2018). This poorer reliability is likely attributable both to idiosyncratic properties of individual tests and to the fact that reliability of the tasks may increase with age. For instance, while Qi et al. (2019) and Von Koss Torkildsen et al. (2019) found good internal consistency (i.e., Conbach's $\alpha > .8$) in VSL in their samples of children, Arnon (2020) found poor internal consistency in a sample of similarly aged children. However, there were differences in the way the task was executed across the studies that may have led to differences in the results (see Arciuli & Conway, 2018; Von Koss Torkildsen et al., 2019, for broader discussions of these points). For example, the stimuli differed and the number of test trials was lower in Arnon (2020) than in Qi et al. (2019) and Von Koss Torkildsen et al. (2019). These features, along with the nature of task instructions, may matter in capturing stable individual differences (e.g., Arciuli et al., 2014). West et al. (2018, see also West, Shanks et al., 2021) have reported poor test-retest reliability for the serial reaction time task (Nissen & Bullemer, 1987) in children aged 8 years, whereas the test shows some consistency in adults (Kalra et al., 2019; Siegelman & Frost, 2015; for discussion see Conway et al., 2019). Kidd et al. (2020) developed an auditory SL task where participants repeated sequences of syllables following familiarisation, which yielded acceptable internal consistency and excellent short-term test-retest reliability in children aged, on average, 6;9 years.

Task reliability is likely to have hampered consistent observation of a link between individual differences in SL and language in developmental populations, if that link exists. On the one hand, some behavioural studies have observed concurrent and longitudinal relationships between SL and spoken language. Early work reported significant (though inconsistent) correlations between the acquisition of probabilistic regularities in the visual domain and early language acquisition in infants (Ellis et al., 2014; Shafto et al., 2012). Work by Lany and colleagues (Lany, 2014; Lany et al., 2018) has shown that infants' existing linguistic knowledge is related to their performance on SL tasks (see also Hoareau et al., 2019). More recently, Gerbrand et al. (2022) reported that associative learning and YSL in the first year of life predicted vocabulary knowledge at 18 months. Similarly, in a large prospective longitudinal study, Frost et al. (2020) reported that 17-month-old infants' successful segmentation of words containing non-adjacent dependencies from continuous speech predicted their vocabulary size up to 12 months later. In older children, significant relationships have been observed between SL and vocabulary as well as grammatical knowledge (e.g., Clark & Lum, 2017; Conti-Ramsden et al., 2015; Evans et al., 2009; Kidd, 2012; Kidd & Arciuli, 2016; Stojanovik et al., 2018). Such findings support the suggestion that SL abilities support language acquisition (Arciuli & Conway, 2018; Christiansen & Chater, 2016; McCauley & Christiansen, 2019; Saffran et al., 1996; Tomasello, 2003; Yang & Piantadosi, 2022).

However, some behavioural studies have not observed the hypothesised SL-language link. A meta-analysis by West, Melby-Lervåg, et al. (2021) suggests that this may be due to task-specific reasons; while some measures showed significant effect sizes others did not, underlying the point that accurately measuring SL is crucial for any study investigating the SL-language link. Indeed, many of the tasks used in these studies were originally developed to investigate group-level effects and are thus designed to minimise between-participant variability (Hedge et al., 2018). Consequently, they may not be accurately estimating the size of the effect because the tasks are not adequately capturing the range of variability in SL. Relatedly, it is important to point out that task selection of SL tasks is rarely if ever considered in significant detail (Arciuli & Conway, 2018; Bogaerts et al., 2022). Rather, researchers have taken tasks that test for learning of distributional regularities in either the linguistic or non-linguistic domains as general measures of SL, and in this sense treat the concept as a kind of black box which performs statistical computations. Much of the work to date is characterized by a very small

number of closely related tasks that only capture a fraction of the statistical patterns that are likely to influence the development of language and cognition (Frost et al., 2019). An important step forward is thus to select multiple reliable tasks to better determine the nature of the SL-language relationship (Siegelman, 2020; Von Koss Torkildsen et al., 2019).

2. The current study

In the current study we present, to our knowledge, the first longitudinal study of the relationship between SL and language in school-age children. Specifically, we followed a cohort of over 100 children who were in the first two years of primary school for 18 months, testing their SL every 6 months to track how change in SL relates to language proficiency. There were several key aspects to our design, which we outline here. At time 1, we measured children's SL across two modalities using tests of ASL and VSL. As outlined above, it seems that measures of these two constructs are unrelated, even when they test the same types of distributional statistics in the same manner. The fact that both have been found to correlate with language suggests that they may be related to language proficiency in different ways and may thus explain separate variance. We tested this possibility in our study.

At time 1 we also included several covariates to rule out the possibility that any observed SL-language relationship could be explained by other uncontrolled variables. These included: children's verbal working memory, non-verbal IQ, and language proficiency. At time 4 we measured language proficiency using a standardised test. This is an important addition: just as reliability of measurement if important for SL, so too is it important to reliably measure language proficiency.

Thus, our aim in the study was to measure ASL and VSL, language, and several covariates at time 1 and determine how SL is both related to these variables and is longitudinally related to language proficiency 18 months later. A second aim was to track developmental change in both ASL and VSL across the course of the study. Following past work (e.g., Kidd & Arciuli, 2016, among others), we expected to observe positive associations between SL and language, such that SL would explain independent variance in language proficiency both concurrently and longitudinally. We also predicted that SL in both modalities performance would increase with age, although whether growth would be the same across modalities is unclear (Arciuli & Simpson, 2011; Raviv & Arnon, 2018).

3. Method

3.1. Participants

One-hundred-and-twenty-one children (51 females, 70 males) were recruited from the first two years of six primary schools in a medium-sized city in Australia. The recruitment criteria were: (i) children were typically developing and had no identified cognitive or language impairment, and (ii) children were acquiring Australian English as a first language with no more than one day per week exposure to a second language. All but two children were monolingual; the two bilingual children were retained in the data set because their vocabulary knowledge was within the normal range for their age. Table 1 shows the number of children and their age at each testing session, showing that 14 children dropped out of the study across time. Note that sometimes children missed tasks within testing sessions, and so our analyses typically have a slightly lower N (though never less than 100).

The children's language proficiency at time 1, as measured by the Peabody Picture Vocabulary Scale 4th edition (Dunn & Dunn, 1997), was within the normal range (Mean standardised score = 116.8, SD = 11.39, Range: 91 – 150).[1] This was also the case for their non-verbal IQ, as measured by the Ravens Coloured Progressive Matrices (Raven et al., 1998, Mean standardised score = 105.62, SD = 12.94, Range: 80 – 135). The one child who scored less than 1.5 standard deviations from the mean (i.e., < 85) on non-verbal IQ did not complete time 4 testing, when the children were tested on their language outcomes, and thus did not contribute data to the longitudinal analyses.

Consistent with the demographics of the city, the children were drawn from areas that are high in socio-economic status (SES) relative to the rest of Australia. Based on the Australian Bureau of Statistic's Socio-Economic Indexes for Areas (SEIFA) (ABS, 2016) scores of relative advantage and disadvantage, participants lived in areas ranked in the 91st percentile or higher (M = 95.5, Range: 91 – 100, where higher scores indicate relatively greater advantage and less disadvantage) compared to all other areas within Australia. However, relative to areas within the immediate administrative territory, the SEIFA scores were more variable (M = 68.83, Range: 40–100). Ethnicity was not recorded.

3.2. Materials

3.2.1. SL tasks

We measured SL in the auditory and visual domain, aiming to equate task difficulty as much as possible. We point out that our VSL task used visual objects and our ASL task used auditorily presented synthesised syllables. In this sense, our tasks confound the auditory/visual dichotomy on the one hand, and the linguistic/non-linguistic dichotomy on the other. This was somewhat unavoidable, given our goals to study the relationship between SL and language and the age of the children, who could not be tested on written

[1] Note that the Peabody Picture Vocabulary Scale 4th edition does not have Australian norms, and so we used the USA norms to roughly estimate the typical range of development but use raw scores in our main analyses (see Results). Similarly, there are no Australian norms for the Raven's Coloured Matrices that span our age range, and so we use the UK norms for the purpose of estimating the typical range, again using raw scores in our main analyses.

Table 1
Age, sample size, and gender distribution of participants across the study.

	Time 1	Time 2	Time 3	Time 4
Mean age (years; months)	6;2	6;8	7;4	7;10
SD (months)	4.49	4.29	4.53	4.33
Age Range (months)	[62,86]	[68,91]	[74,98]	[80,103]
N[a]	121	117	117	107
Gender	51 females	49 females	49 females	45 females

[a] Only 101–102 participants were included in our longitudinal models because children missed sessions across the study.

language. Our selection of tasks is, however, largely consistent with past studies, who have overwhelmingly measured ASL and VSL in this manner (e.g., Arciuli & Simpson, 2011; Arnon, 2020; Evans et al., 2009, among many others). Children were tested on the SL every six months, consistent with our aim of measuring developmental change in SL. We chose six-month intervals for both theoretical and practical reasons. Theoretically, past studies have shown, at least for VSL, small increments of age-related developmental change within the age range we tested (Arciuli & Simpson, 2011; Raviv & Arnon, 2018), such that developmental change is likely observable at longer rather than shorter time intervals (e.g., weeks). Practically, the length of the study was restricted by the length of funding (3 years). Six-month intervals allowed us to include four testing time points given our original aim of recruiting at least 100 children.

3.2.1.1. Visual statistical learning task (Arciuli & Simpson, 2011). We used Arciuli and Simpson's (2011; 2012) embedded triplet 'Alien Attack' task to measure VSL, presented via E-Prime 2 (Schneider et al., 2002). The task consists of two phases: (i) familiarisation, and (ii) a surprise test phase. During familiarisation, children see a continuous stream of 12 cartoon aliens appearing in the centre of a computer screen, one at a time (each appearing for 800 ms, with an interstimulus interval of 200 ms). They were told that they will see a set of aliens lining up to enter their spaceship, and they should press the space bar every time they see the same kind of alien appear twice in a row. This acted as a cover task, but also served to maintain children's attention. Unbeknownst to the children, the aliens occur in 4 triplets semi-randomly throughout the familiarisation, referred to here as *ABC, DEF, GHI,* and *JKL*, with the two constraints: following Turk-Browne et al. (2005), a given triplet did not occur twice in a row (i.e., there was no *ABCABC*) and pairs of triplets were not repeated (i.e., no *DEFJKLDEFJKL*). Each triplet occurred 24 times during familiarisation. For 6 of the 24 instances, one of the aliens was presented twice in a row, as the basis of the cover task. Thus, there were 24 cover task trials in total. Triplets were defined by high transitional probabilities (TPs) within triplets, but low TPs between triplets. Because of the cover task, the within-triplet TPs were 0.92. The between-triplet TPs were 0.33. The familiarisation stream lasted in 5 mins and 12 s in total. At times 1 – 3 the children were given a short break (typically lasting less than 30 s) halfway through familiarisation, to ensure that they stayed on task. At time 4 this was not deemed necessary, and the children sat through the familiarisation stream in its entirety. At no point before or during the familiarisation phase were children told to learn or remember anything.

Immediately following the completion of the familiarisation stream the children completed a two alternative forced choice (2AFC) test. This test phase measures the incidental learning of the embedded triplets presented during familiarisation, aligning with the idea that language acquisition and processing involves sensitivity to transitional probabilities that occur in natural language (e.g., Aslin et al., 1998). Children were presented with a triplet as it was seen in the familiarisation phase (i.e., three aliens appeared one at a time for 800 ms each, with a 200 ms interstimulus interval) and one of four impossible triplet foils, which were created by taking one alien from each triplet (these were: *AEI, DHL, GKC,* and *JBF*). The internal TPs of the foils was 0, but note that they were not part-words, which was also the same in our ASL task (see below). This is consistent with many similar studies of SL, but admittedly sets only minimal demands on participants for distinguishing between trained words and foils in the 2AFC task. The presentation of the triplet and foil was separated by 1000 ms, and their order of appearance was counterbalanced. Following the presentation of all six aliens a new screen appeared that prompted participants to identify which of the two triplets had appeared previously in the familiarization phase, with no time constraints imposed. Each base triplet was paired with one of the impossible triplets on two separate occasions; each attested triplet was seen 8 times for a total of 32 forced choice trials. The presentation of the triplet pairs was randomized for each participant. Participants' responses during the test phase were tallied to provide a proportion correct score.

3.2.1.2. Auditory statistical learning task (Kidd et al., 2020). We used Kidd et al.'s (2020) ASL task, presented using E-Prime 2 (Schneider et al., 2002). Like the VSL task, the task uses the embedded triplet paradigm presented in the context of an alien theme but yields two possible measures of ASL. Children are introduced to the task as a game in which they will listen to alien signals coming through an alien's spaceship radio. The task has three phases: (i) a familiarisation phase, (ii) a 2AFC test, and (iii) a sequence repetition phase referred to as *Statistically Induced Chunking Recall* (SICR, see also Isbilen et al., 2020). During the familiarisation phase, children listened to four trisyllabic words (*kibudu, lomari, modipa, takapo*)[2] that occurred within a continuous speech stream of syllables, over circumaural noise-cancelling headphones (Sennheiser HD 280 Pro). The syllables were created using the MBROLA speech synthesiser; there was no intonational patterning in the speech. The average length of each syllable was 305 ms, with an average intersyllable gap of 120 ms. The triplets were defined by high within-word TPs (1.0) and low between-word TPs (0.33). The full familiarisation phase consisted of 72 tokens of each word, divided into three blocks of approximately 24 tokens each. Blocks were created to maintain the

[2] The words differed at time 1, see Appendix for details.

children's attention, following extensive piloting. Each block lasted approximately 2 min, with an average break between blocks of 20 s

At the end of the familiarisation phase, the children completed 32 2AFC trials, in which they were asked to distinguish between an attested word and a foil. On each trial, a trained word and a foil were played 1000 ms apart, and children indicated which of the words was part of the alien's radio signal. Foils were constructed from the same pool of syllables as the test words, but crucially has TPs = 0 given the familiarisation stream. This component of the task was scored automatically using E-Prime (1 = correct, 0 = incorrect).

Following the 2AFC test, children completed the SICR task. SICR was designed to measure recall of recurring syllables, based on the idea that language acquisition and processing involves the chunking of recurring units of language, which are then more easily recalled as higher order units (i.e., syllables into words, words into phrases, etc…) (Ambridge & Lieven, 2015; Christiansen & Chater, 2016). In the task, children hear sequences of attested syllables and foils, and are simply required to repeat what they heard. If, during familiarisation, the children were chunking the attested sequence, their recall of the attested sequences should be significantly better than foils (Christiansen, 2019).

The procedure for the SICR component followed Kidd et al. (2020) (Study 1). In the first 8 trials, children received either a three-syllable attested sequence (i.e., one familiarised word) or a three-syllable foil. In a trial, children heard the to-be-repeated sequence, and following a 500 ms gap, a high-pitched tone played, which cued children to repeat what they had heard. These initial trials served to introduce the children to the task, since asking children to repeat long sequences can be difficult (Gathercole, 2006). These trials were not analysed: although performance on them does typically show a learning effect (see Kidd et al., 2020), scores on three-syllable sequences are less reliable within individuals than on longer sequences. Additionally, performance may implicate chunking less than on longer sequences, since three syllables is within most 6-year-olds' verbatim memory span. Thus, after the initial 8 shorter trials, the children completed a further 16 six-syllable sequences. These were either: (i) two trained trisyllabic words concatenated together, or (ii) a six-syllable foil sequence containing TPs = 0. Once again, children heard the sequence, and after 500 ms heard a high-pitched tone, which was their signal to repeat the sequence as best they could. Children's repetitions were recorded using a Zoom H4n Pro handheld recorder.

The SICR task was scored offline from the audio recordings. Children were scored 1 for every correctly repeated syllable that occurred in the correct serial order. Thus, if the test sequence was *modipatakapo*, and the child produced *mo**da**pa**to**kapo*, they would receive a score of 4/6 = 0.67. As is typical in studies of verbal learning, participants did not always produce six syllables in their repetition. In this case, children were still awarded 1 point for every correctly recalled syllable that occurred in a correct position relative to the other syllables that were recalled. For example, for the item *modipatakapo*, if a child produced *moditapo*, they would receive 4/6, since all syllables recalled were in their correct relative serial position (i.e., *mo-di-xx-ta-xx-po*, where *xx* represents missing syllables). Following Kidd et al. (2020), consistent substitution of phonemes (e.g., /v/ for /b/) was coded as correct. At each time point 10 children's repetitions were re-coded by a second blind coder for reliability. At each time point, the agreement was high (overall Cohen's Kappa: T1:.90, T2:.87, T3:.93, T4:.90, with no difference across trained and foil items or syllable number, all Kappas >.83).

3.2.1.3. Exogenous variables. We measured several features of children's language and cognitive skills at time 1, which served as exogenous variables in our longitudinal models, in addition to children's age (in months). These were:

3.2.1.4. Vocabulary (Peabody Picture Vocabulary Test 4th edition, Dunn & Dunn, 1997). The Peabody Picture Vocabulary Test (4th edition) (PPVT) is a measure of vocabulary comprehension and is commonly used as a measure of verbal ability. In the task, children are shown an array of four pictures, and are asked to point to the picture matching a verbal label provided by the experimenter. The PPVT has excellent psychometric properties (split-half reliability, alpha (α), and test–retest reliability all > .9). Children receive 1 point for every correct trial, with a ceiling of 226. We report raw scores here, which we used in our models because we are interested in estimating total vocabulary knowledge independent of age (i.e., we are ranking children on total vocabulary knowledge at the time of testing).

3.2.1.5. Grammar (Kidd & Arciuli, 2016). Grammatical comprehension was measured using Kidd and Arciuli's (2016) test of multiple syntactic structures (see also Boyle et al., 2013). In the task, children are shown two pictures that depict reversable transitive events (e.g., a chicken kissing a mouse, a mouse kissing a chicken), which are described by one of four syntactic structures that vary in their syntactic complexity: (i) an active sentence (*Which chicken is kissing the mouse?*), (ii) a passive sentence (*Which chicken is being kissed by the mouse?*), (iii) a subject relative clause (*Where is the chicken that is kissing the mouse?*), and (iv) an object relative clause (*Where is the chicken that the mouse is kissing?*). Each child heard 32 sentences, 8 of each structure type, from a pool of 128 sentences distributed across 8 lists. The verbs were in the present progressive because the pictures denoted events that were currently occurring. There were eight verbs in total (*comb*, *feed, follow, hug, kiss, push, scare,* and *splash*), and 28 different animate characters that appeared across the sentences. The test sentences were controlled for length in syllables; all were between 10 and 12 syllables in length (mean length = 10.8 syllables). The length of active sentences is typically shorter than the other three sentence types because they require less grammatical function words (e.g., *is being, that*). In order to balance sentence length, the active sentences were made longer using adjectives that did not provide any clues to the identity of the target referent (e.g., *nice, tall*). Children were asked to point to the character that was identified in the question. They were given 1 point for identifying the correct character in each sentence. Their correct answers were pooled across all sentence types and converted to a proportion to determine an overall measure of their syntactic comprehension. The overall task internal reliability across all sentence types for the measure was acceptable (α = .75).

3.2.1.6. Non-Verbal IQ (Raven's Colored Progressive Matrices, RCPM, Raven et al., 1998). The RCPM served as a measure of nonverbal

IQ. In the task, children are asked to complete the visual patterns of varying complexity by selecting the correct missing piece, from an array of six. Cotton et al. (2005) reported good internal consistency and reliability in a large sample of Australian 6- to 11- year-olds (internal consistency [K–R formula 20] = .89; split-half reliability = .90). Children receive 1 point for every correct answer, with a ceiling of 36. We report raw scores here, which we used in our models.

3.2.1.7. Working memory. Children's verbal WM was measured because it has been implicated in language development (Baddeley, 2003; Montgomery et al., 2021). We used the Listening Span measure from Gathercole and Pickering's (2001) Working Memory Test Battery for Children, which is a measure of complex verbal working memory span. In the task, children listen to a series of sentences and are required to (a) verify the truth value of the sentence and (b) remember and recall the final word of each sentence. The test becomes progressively more difficult, as indexed by the number of sentences in each block. Gathercole and Pickering (2001) report very good test–retest reliability (r = .83). The children's raw scores were used in the analyses because there are no published Australian norms for the test. Scores range from 0 to 36.

3.2.1.8. Outcome variable: Clinical Evaluation of Language Fundamentals, 4th edition (CELF-4, Australian version, Semel et al., 2006). We used the Core Language scale of the CELF-4 to measure language proficiency. The test is a broad-spectrum standardised measure of children's language. The Core Language scale comprises four subtests: (i) Concepts and following directions, a measure of comprehension, (ii) Word structure, a measure of the production of English morphology (iii) Recalling sentences, a measure of sentence production that significantly implicates grammatical processing (see Boyle et al., 2013), and (iv) Formulating sentences, a measure of children's ability to generate sentences given prompt words. Performance was scored according to the test manual. Raw scores from each subtest were aggregated to produce one Core Language score, which was used in our analyses. The CELF-4 has good reliability: Cronbach's α range from .69 to.91 for subtests and from .87 to.95 for composite scores. The test-retest reliability of CELF-4 was evaluated in a study with 320 students. The stability coefficients range from .71 to.86 for subtests and from .88 to.92 for composite scores based on the standardization population.

3.2.2. Procedure

These data are drawn from a larger longitudinal study investigating individual differences in primary-school aged children's language. At time 1 the participants completed all the exogenous variables and the two SL tasks. At time 2 and 3 participants completed the SL tasks and some additional language tasks not reported on here. At time 4 they completed the SL tasks, the CELF-4, and some additional language tasks. Since the children were tested on a range of tasks and were tested during school time, the tasks were allocated to between 2 and 4 blocks, depending on how many tasks there were per time point. Within each block the order of tasks was always the same; however, block order varied to complete testing in each school in a timely manner. With only a few exceptions (1–2 occasions per time point), each child was only tested on one block per day. Thus, the tasks were administered in an as uniform manner as possible given the constraints of the school curriculum, and where it was not possible to test children on the exact same order of tasks, any influence a change in order may have had on the data was likely extinguished by the many hours in between session blocks. The two SL tasks were always in separate blocks.[3]

With the exception of the ASL task at time 1 (see explanation below), we used the same SL tasks at each time point. We made this decision primarily because changing the task would involve testing different triplets each time, which could have added additional noise to the data. Of course, this raises the possibility that children came to subsequent sessions with some memory of triplets in each task, a possibility we cannot rule out even though there was 6 months between each time point. We return to this point in the Discussion.

4. Results

The results section is structured as follows. First, we analyse the children's performance on the SL tasks, including an analysis of the reliability of the tasks. We then present the descriptives and correlations between all variables of interest. Finally, we present longitudinal path analyses that bear upon our aims and hypotheses. Our raw data and analysis scripts are available on the Open Science Framework (https://osf.io/5ad9v/?view). Our main analyses were conducted using the R statistical software (v. 4.1.2, R Core Team, 2021).

4.1. Reliability of SL tasks

Table 2 shows internal consistency (reported as Cronbach's α based on Pearson correlations) results for the 2AFC components of the ASL and VSL tasks and for the SICR repetition component of the ASL task across the four time points. Table 2 reports internal consistency for the VSL and ASL 2AFC tasks based on the raw data. We also computed the same statistics based on shrunken logit-transformed data. The logic of the logit transformation is that it brings proportions into the real line, thereby removing the

[3] An anonymous reviewer suggested that there may be an order effect in performance on SL tasks within sessions, such that children would perform better on the second task on which they were tested because they would be primed to look for triplets in the familiarisation stream. We did not find evidence of this (see OSF page for details).

Table 2
Internal consistency (Conbach's α) of SL measures Across Time.

	Time 1	Time 2	Time 3	Time 4
Measure				
VSL-2AFC	–	.45 (.56)	.60 (.69)	.59 (.64)
ASL-2AFC	–	–	.30 (.30)	.45 (.43)
ASL-SICR	.67	.79	.75	.76

Note: Figures in brackets denote log-transformed statistic based on shrunken data. VSL-2AFC = two alternative forced choice component of VSL task; ASL-2AFC = two alternative forced choice component of ASL task; ASL-SICR = test repetition component of ASL task.

influence of extreme scores (see OSF materials for further details).

What is clear is that the children's performance on the tasks becomes more internally consistent across time, but the change is task dependent. At time 1 none of the scores have acceptable reliability. In fact, the correlations between items for both 2AFC measures were close to zero and contained a mix of positive and negative values, which meant that computing α coefficients would not be informative (hence no coefficients are reported). The 2AFC component of the ASL task improved but never came close to acceptable levels (consistent with Arnon, 2020, Kidd et al., 2020, Siegelman et al., 2018), and so we did not use the task in our individual differences analyses. The reliability of the VSL task improved more than the ASL 2AFC measure, particularly at time 3. Note that Von Koss Torkildsen et al. (2019) used the same VSL task, with 64 test trials, and reported internal consistency of .81 in a sample of 7–12-year-olds (mean age = 10;3 years). Qi et al. (2019) reported high reliabilities between .79 and .88 using versions of a similar VSL task, with 32 test trials, in older children aged 8 – 16 years (mean age = 12; 2 years). This pattern of results is consistent with the suggestion that the VSL task becomes more reliable with age and with more test items.

The ASL-SICR task had the best reliability at time 1 and 2. At time 3 and 4 differences in reliability between VSL and ASL-SICR are smaller when considering the transformed data. Note, also, that the ASL-SICR has two layers of reliability compared with one layer for the VSL task, given the difficulties in coding speech data. Because not all tasks showed individual-level reliability at earlier ages, we did not analyse growth in SL across development, since poor reliability at earlier time points makes any observed development difficult to interpret.

4.2. SL tasks performance

Table 3 presents the descriptive statistics for children's performance on the VSL task at each time point, along with one-sample *t*-tests comparing performance to chance (= .50), and the effect size (Cohen's *d*). One univariate outlier was removed at time 3 (mean score = .13, $z = 2.97$). The results show that, as a group, the children scored significantly above chance at each time point, although the strength of the effect, as indicated by the effect sizes, was higher at times 3 and 4.

Table 4 presents the descriptive statistics of children's performance on the 2AFC and SICR components of the ASL task. The 2AFC data were analysed using one-sample *t*-tests comparing performance to chance (= .50); the SICR data were analysed using paired-samples *t*-tests, whereby the children's repetition of the target sequences was compared to their foil repetition.

Table 4 shows that the children exhibited no clear evidence of learning in the ASL task at time 1, based on both the 2AFC and SICR task. The modified task was used from time 2 onwards (see Appendix). On this task the children showed evidence of successful learning on the 2AFC task at time 3 and 4 only, whereas they showed successful learning with large effect sizes on the SICR measure from time 2 onwards.

4.3. Descriptives

Table 5 contains descriptive data for the time 1 exogenous variables at time 4 language proficiency. Table 6 shows the simple Pearson bivariate correlations between the exogenous time 1 variables, all SL measures, and the time 4 language outcome measure. We note several features of the correlations among the SL tasks and their relationship to other variables. For VSL, we see significant correlations in performance across all time points, although the strongest is between time 3 and time 4, when the internal consistency of the task was highest. In contrast, the time-lagged correlations among the 2AFC component of the ASL task were never significant, suggesting that performance was variable within individuals across the course of the study. In contrast, all ASL-SICR variables correlated with each other across time. There were no consistent correlations between the VSL an ASL tasks, which is consistent with similar past research (e.g., Arnon, 2020; Siegelman & Frost, 2015). Finally, all time 1 exogenous variables excluding age significantly correlated with time 4 language. Time 3 and 4 VSL significantly correlated with time 4 language, and all ASL-SICR measures significantly correlated with time 4 language.

4.4. Path analyses

Our original goal was to determine how individual differences in visual and auditory SL at time 1 related to our covariates and how variability in SL was subsequently related to language development across time. However, without reliable measurement of SL we reassessed our analysis plan. Since we wanted to determine the joint contribution of ASL and VSL to language proficiency, we used the most reliable measures of both constructs at time 3 and 4 (i.e., VSL & ASL-SICR), when the tests were most reliable and could both be

Table 3
Means (Standard Deviations), *t*-test, and Cohen's *d* Describing Children's Performance on the VSL Task at Each Time Point.

	Time 1	Time 2	Time 3	Time 4
Mean (SD)	.52 (.11)	.54 (.13)	.59 (.14)	.58 (.15)
t	2.19[a]	3.28[b]	6.29[c]	5.38[c]
d	0.20	0.30	0.58	0.52

[a] $p < .05$.
[b] $p < .01$.
[c] $p < .001$.

Table 4
Means (Standard Deviations), *t*-test, and Cohen's *d* describing children's performance on the ASL-2AFC and ASL-SICR Tasks at each time point.

	Time 1		Time 2		Time 3		Time 4	
ASL-2AFC								
M (SD)	.51 (.08)		.52 (.10)		.59 (.12)		.61 (.13)	
t	1.56		1.76		7.85***		8.52***	
d	0.14		0.16		0.73		0.82	
ASL-SICR	Target	Foil	Target	Foil	Target	Foil	Target	Foil
M (SD)	.27(.12)	.28(.13)	.38(.18)	.24(11)	.45(.17)	.23(.10)	.51(.18)	.26(.11)
t	-1.2		10.63***		16.7***		16.05***	
d	.09		.89		1.57		1.67	

*** p < .001. ASL-2AFC = two alternative forced choice component of ASL task; ASL-SICR = test repetition component of ASL task.

Table 5
Descriptive statistics for time 1 exogenous variables and the time 4 outcome variable.

	Mean	SD	Range	Skewness	Kurtosis
Time 1					
Vocabulary	125.9	16.97	[83,164]	-0.21	2.71
Grammar	0.65	0.16	[.25, 1.0]	0.31	2.56
Non-verbal IQ	22.18	4.49	[13,33]	0.29	2.61
Working Memory	8.25	2.96	[0,16]	0.22	3.14
Time 4					
Language	155.9	23.43	[96,206]	-0.08	2.43

included in our models. Thus, we ran path analyses where we used our covariates at time 1 as exogenous variables to predict VSL and ASL-SICR at time 3, which were then used to predict language proficiency at time 4 (called 'Model 1' in OSF materials). In a subsequent set of models, we added children's SL performance at time 4, which tested the stability of measurement of SL across time and how this relates to language (Models 2 – 3 in OSF materials).

All modelling was conducted in *lavaan* (v0.6–7, Rosseel, 2012). Following Kline (2005), we report the following fit indices: (i) model Chi-square statistic, with $p > 0.05$ indicating good fit between the model and the data, (ii) Comparative Fit Index (CFI), with a cut-off of CFI > = 0.90 to indicate good fit, (iii) Root Mean Square Error of Approximation (RMSEA), with a cut-off of RMSEA < 0.08, and (iv) Standardised Root Mean Square Residual (SRMSR), with a cut-off of SRMSR < 0.08. Model fits were compared using likelihood ratio tests and inspection of Akaike and Bayesian Information Criteria. All models can be viewed in the OSF materials; we only report our final models here.[4]

4.5. Predicting language proficiency from SL at time 3

A model containing all exogenous variables showed good fit to the data ($\chi^2 = 0.159$, df = 1, $p = .69$; CFI = 1.000; RMSEA < 0.001, 90 % CI = [0.000, 0.194]; SRMR = 0.006). This model is presented in Fig. 1. It shows that time 1 vocabulary, grammatical comprehension, and working memory significantly positively predict time 4 language, but age has a negative effect. Time 1 non-verbal IQ predicts time 3 VSL. No time 1 variables predict time 3 ASL-SICR. Together the exogenous variables explained 18.4 % of the variance in time 3 VSL ($R^2 = .184$), and 37.8 % of the variance in time 3 ASL-SICR ($R^2 = .378$), though this latter result was due to the variable's strong relationship with foil repetition. The path from time 3 VSL to time 4 language was not significant, whereas the path from time 3 ASL-SICR to time 4 language, controlling for foil repetition, was significant. The combination of all predictor variables

[4] During the review process we were asked to add gender as an exogenous variable, but it had no effect on either SL or on language outcomes at time 4. Since gender had no initial effect in the models, we did not test its role as a moderator effect. We refer the reader to our OSF page for the relevant details.

Table 6
Bivariate Correlations between Time 1 Exogenous Variables, SL Measures, and Time 4 Language (Bolded Correlations $p < .05$, uncorrected for multiple tests).

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Time 1																						
1. Age (months)	-																					
2. Vocabulary	**.47**	-																				
3. Grammar	**.32**	**.38**	-																			
4. Non-Verbal IQ	**.20**	**.29**	**.36**	-																		
5. Working Memory	**.29**	**.36**	**.33**	**.29**	-																	
SL variables																						
6. VSL T1	.07	.02	.02	.01	-.10	-																
7. VSL T2	-.05	.11	.12	.08	.18	**.22**	-															
8. VSL T3	.15	**.25**	**.29**	**.34**	**.30**	**.27**	**.38**	-														
9. VSL T4	.09	.15	.18	**.27**	**.30**	**.25**	**.38**	**.68**	-													
10. ASL 2AFC T1	.13	.02	.11	.04	-.07	.03	.06	.12	.16	-												
11. ASL 2AFC T2	**.38**	**.24**	.12	.04	.00	**.27**	.12	.08	.06	-.03	-											
12. ASL 2AFC T3	.08	.03	.03	.07	.11	-.05	.03	.06	.10	-.04	-.02	-										
13. ASL 2AFC T4	.05	.02	.17	**.26**	.15	.03	.01	.11	.13	-.14	.03	.09	-									
14. SICR Target T1	.07	**.22**	**.42**	**.24**	**.36**	-.07	.07	.12	.09	-.15	.05	.11	.12	-								
15. SICR Foil T1	.06	**.23**	**.35**	**.20**	**.42**	-.05	.07	.12	.11	**-.20**	-.07	-.02	.11	**.70**	-							
16. SICR Target T2	**.28**	**.25**	**.28**	**.25**	**.23**	-.02	.02	.12	.02	-.05	.15	.12	**.22**	**.40**	**.36**	-						
17. SICR Foil T2	.03	.14	**.25**	**.32**	**.23**	-.18	-.03	.07	.02	-.06	-.06	.07	.07	**.47**	**.45**	**.66**	-					
18. SICR Target T3	.10	.11	**.20**	.08	**.27**	.01	.01	.03	.03	**-.20**	.07	**.30**	**.32**	**.41**	**.40**	**.67**	**.46**	-				
19. SICR Foil T3	.00	**.19**	**.20**	.08	**.19**	.02	-.01	.08	.06	-.08	-.03	.17	**.22**	**.42**	**.34**	**.49**	**.48**	**.58**	-			
20. SICR Target T4	-.06	-.02	.02	.09	.13	.10	.03	.01	.14	-.19	-.06	**.20**	**.32**	**.33**	**.30**	**.53**	**.45**	**.62**	**.45**	-		
21. SICR Foil T4	-.03	.05	**.21**	.14	**.29**	-.12	.03	-.00	.04	-.11	-.09	.14	.09	**.53**	**.50**	**.49**	**.58**	**.46**	**.52**	**.47**	-	
Time 4																						
22. Language	.19	**.49**	**.47**	**.26**	**.54**	.02	.14	**.29**	**.35**	-03	.05	.12	**.19**	**.35**	**.41**	**.37**	**.39**	**.41**	**.29**	**.32**	**.33**	-

Note: VSL = Visual Statistical Learning; ASL = Auditory Statistical Learning; SICR Target = ASL SICR Target repetition; SICR Foil = ASL SICR Foil repetition; TX = (Time point, 1–4).

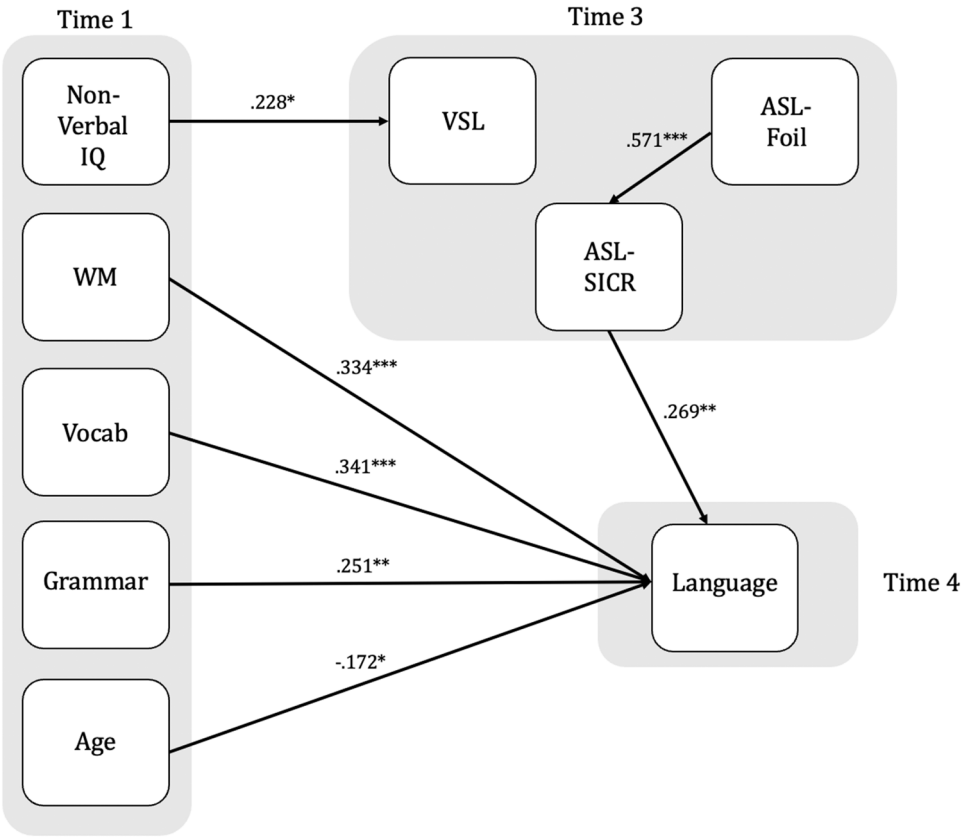

Fig. 1. Visualisation of final SEM predicting SL abilities at time 3 from covariates at time 1, with both predicting language at time 4. Lines denote significant paths (*$p < .05$, **$p < .01$, ***$p < .001$). N = 102, AIC = 666.66, BIC = 727.03). Note: VSL = Visual Statistical Learning; ASL-SICR = score on trained repetition component of ASL task; ASL-Foil = score on foil repetition in ASL task.

explained 54.4 % of the variance in language proficiency at time 4 ($R^2 = .544$).

4.6. *Predicting language proficiency from SL at times 3 and 4*

We added VSL and ASL-SICR at time 4 to the models. Since these SL tasks at time 3 and 4 had the best reliability observed in the study, we added a stationarity test of the SL task effects. That is, we restricted the coefficients for time 3 and 4 VSL and ASL-SICR to be the same. Stationarity tests are often used on time series and longitudinal data (see, e.g., Cole & Maxwell, 2003). In this particular context, the function of restricting coefficients for a variable serves to account for any time-based fluctuations in a variable that might influence an effect. Thus, by ensuring that SL coefficients are *stationary* we can infer that any relationship between SL and language is not attributable to the undue influence of performance at one time point.

The final models were a good fit to the data. A model without the stationarity test showed acceptable fit to the data ($\chi^2 = 6.567$, df = 7, $p = .475$; CFI = 1.000; RMSEA < 0.001, 90 % CI = [0.000, 0.118]; SRMR = 0.025). The model explained 18.4 % of the variance in time 3 VSL ($R^2 = .184$) and 48.3 % of the variance in time 4 VSL ($R^2 = .483$). The model explained 38.4 % of the variance in time 3 ASL-SICR ($R^2 = .384$) and 46 % of the variance in time 4 ASL-SICR ($R^2 = .460$). The combination of all predictor variables explained 58.4 % of the variance in language proficiency at time 4 ($R^2 = .584$).

The model containing the stationarity test also showed a good fit to the data ($\chi^2 = 21.327$, df = 17, $p = .212$; CFI = 0.983; RMSEA = 0.050, 90 % CI = [0.000, 0.109]; SRMR = 0.048). The model explained 7.5 % of the variance in time 3 VSL ($R^2 = .075$) and 45.6 % of the variance in time 4 VSL ($R^2 = .456$). The model explained 36.9 % of the variance in time 3 ASL-SICR ($R^2 = .369$) and 45.1 % of the variance in time 4 ASL-SICR ($R^2 = .451$). The combination of all predictor variables explained 59.5 % of the variance in language proficiency at time 4 ($R^2 = .595$).

A comparison of the two models using a log likelihood ratio test showed no reliable difference between them ($\chi^2 = 14.76$, df = 10, $p = .141$), with a numerical difference in favour of the non-stationary model. However, a comparison of AIC and BIC of the two models showed a numerical difference in favour of the stationarity model (Non-stationary model: AIC = 392.09, BIC = 504.54; Stationary model: AIC = 386.85, BIC = 473.15). In both models the paths from time 4 VSL and ASL-SICR to time 4 language were significant, with no qualitative difference in the pattern of significant paths to Language at time 4. Accordingly, we interpret the lack of difference between the models to mean that the stationarity test worked, in the sense that restricting the coefficients for the VSL and ASL-SICR

variables at time 3 and 4 did not result in significant loss of fit. Fig. 2 displays the significant paths for the stationarity model.

4.7. The independent contributions of ASL and VSL to language

In our final analyses we aimed to determine the source of the independent contributions of ASL-SICR and VSL to language proficiency. The CELF-4 Core Language Scale has four subscales. Even though they load onto the same factor, they may be differentially supported by ASL or VSL. Table 7 reports the correlations between time 3 and 4 ASL-SICR and VSL and the four CELF-4 subscales.

Table 7 shows that, in almost every case, both the ASL-SICR and VSL tasks were significantly and positively correlated with all subtests, with raw correlations ranging from .160 to .404 for ASL-SICR and from. 170 to .434 for VSL. However, the strength of the associations differs across subtests. Notably, the ASL-SICR is most strongly associated with the Recalling Sentences subtest, whereas the VSL task is most strongly correlated with the Concepts and Following Directions subtest, followed by the Word Structure subtest at time 4. To determine whether the variable pattern of correlations translated to the two SL tasks loading differentially on different subtests, we re-ran our path analyses using each subtest as outcome variables. Table 8 summarises the results (see OSF materials for code and output). Here we only report the results the paths from SL to each subtask, first for the time 3 SL tasks only, and then for models that added in the time 4 SL tasks, the latter including tests of stationarity, which in every case were the preferred models. The results showed that, consistent with the strength of the correlations, the ASL-SICR task and the VSL task differentially contributed to performance across the subtests. The ASL-SICR task best predicted the Recalling Sentences subtest at both time 3 and time 4. It significantly predicted the Formulated Sentences subtest at time 3. In contrast, the VSL task significantly predicted the Following Directions subtest and the Word Structure subtest at time 4.

5. Discussion

In this paper we presented longitudinal data bearing upon the relationship between SL and language proficiency. Our original aim of measuring both ASL and VSL at time 1 of our study was hampered by the fact that the measurements were unreliable in the children at this age (mean = 6;0). However, the reliability of SL measurement increased across time, and we determined (i) the relationship between our time 1 covariates (measuring language and cognitive skill) and later measured SL, and (ii) whether SL measured at time 3 and 4 was significantly related to language proficiency over and above our time 1 covariates. Our path analyses revealed evidence in favour a significant SL-language link. When only time 3 SL variables were included, we found a significant path between ASL-SICR and language proficiency at time 4, but the path between VSL and language proficiency was not significant (though the bivariate correlation between the two variables was significant). When we added time 4 SL variables into the path analysis, we found two notable

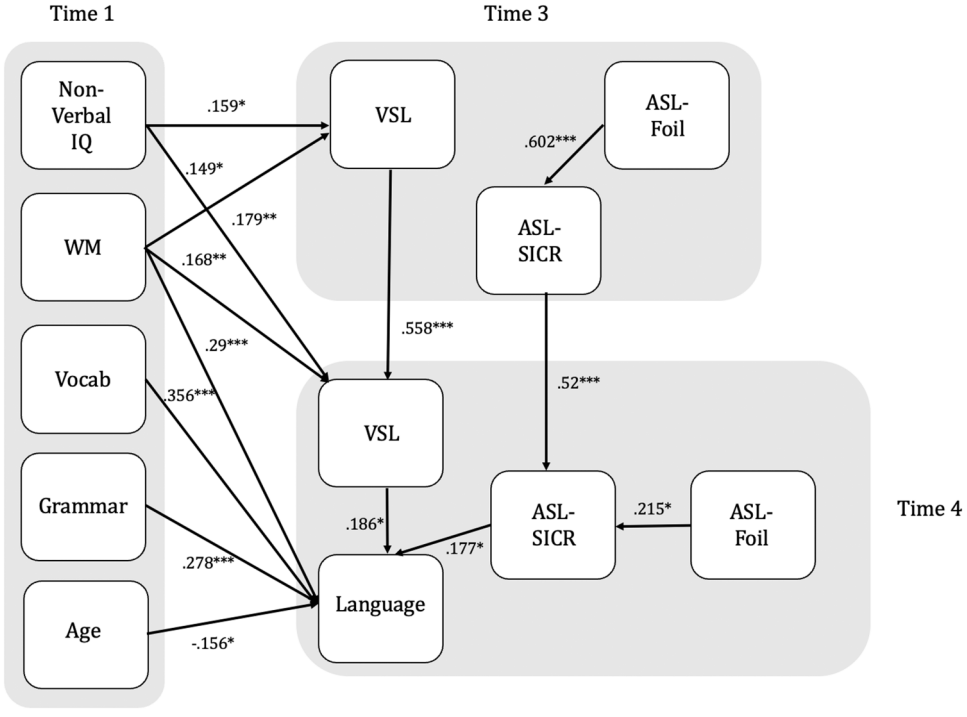

Fig. 2. Visualisation of final SEM predicting SL abilities at time 3 and 4 from covariates at time 1, with both predicting language at time 4. Lines denote significant paths (*$p < .05$, **$p < .01$, ***$p < .001$). N = 101, AIC = 386.85, BIC = 473.15). Note: VSL = Visual Statistical Learning; ASL-SICR = score on trained repetition component of ASL task; ASL-Foil = score on foil repetition in ASL task.

Table 7
Pearson Bivariate Correlations between Time 3 (T3) and 4 (T4) SL tasks and CELF Subtests and Scales (Bolded Correlations $p < .05$, uncorrected for multiple tests).

	1.	2.	3.	4.	5.	6.	7.	8.
SL tasks								
1. VSL T3	-							
2. VSL T4	**.682**	-						
3. ASL-SICR T3	.028	.003	-					
4. ASL-SICR T4	.005	.143	**.618**	-				
CELF-4 subscales								
5. Recalling Sentences	.170	**.206**	**.404**	**.360**	-			
6. Concepts and Following Directions	**.357**	**.434**	**.246**	.160	**.565**	-		
7. Word Structure	**.254**	**.352**	**.255**	**.218**	**.545**	**.476**	-	
8. Formulating Sentences	**.215**	**.235**	**.307**	**.195**	**.492**	**.564**	**.447**	-

Note: CELF-4 = Clinical Evaluation of Language Fundamentals (4th edition, Semel et al., 2006).

Table 8
Contributions from ASL-SICR and VSL Tasks to Subtests and Subscales of the CELF-4 from separate Path Analyses.

	β	p	R^2		β	p	R^2
Time 3				Time 4			
Subtest/scale							
Recalling Sentences							
ASL-SICR	.294	.001**	.456	ASL-SICR	.245	.010*	.518
VSL	-.023	.773		VSL	.063	.516	
Concepts and Following Directions							
ASL-SICR	.103	.298	.381	ASL-SICR	.043	.687	.425
VSL	.154	.074#		VSL	.254	.017*	
Word Structure							
ASL-SICR	.169	.084#	.394	ASL-SICR	.169	.091#	.470
VSL	.091	.282		VSL	.225	.026*	
Formulating Sentences							
ASL-SICR	.209	.047*	.299	ASL-SICR	.075	.510	.317
VSL	.044	.630		VSL	.129	.259	

** $p < .01$.
* $p < .05$.
# $p < .10$.

results. Firstly, the paths between time 3 and 4 SL were significant, showing that, importantly given past discussions about measurement (e.g., Arnon, 2020), we were (eventually) able to measure what appears to be a relatively stable trait in both modalities. Our path analyses involved a stationarity test, which revealed significant paths between time 4 VSL and ASL-SICR and language. Notably, the fit of the stationarity model suggested that the measurements of VSL and ASL-SICR at time 3 and 4 SL were interchangeable, such that we can tentatively suggest that the relationship we found between SL and language was stable in this sample at these time points. Our final analyses showed that the SL tasks across different modalities were differentially related to subtests of our language measure. In the remainder of the paper, we discuss these issues in greater detail.

5.1. Task reliability

First and foremost, the results suggest support for the argument that there are meaningful individual differences in both ASL and VSL, which are measureable and related to language proficiency. Our data add to recent debate concerning the existence of this link. Recent meta-analyses have found the SL-language link to be absent (Lammertink et al., 2020) or task dependent (West, Melby-Lervåg, et al., 2021). One persistent difficulty with much of the past research investigating the SL-language has been task choice, both of SL and language outcome measure. Notably, reliably measuring SL in young children using standard tasks has proven difficult (Arnon, 2020; West et al., 2018; though see Kidd et al., 2020; Von Koss Torkildsen et al., 2019; Qi et al., 2019), and using paradigms that make use of experimental designs originally designed to detect group effects means that individual differences, if they exist, are more difficult to detect (Arciuli & Conway, 2018; Hedge et al., 2018; Kidd et al., 2018; Von Koss Torkildsen et al., 2018). A less often made point in this literature is the importance of reliable outcome variables, but just as it is important to rank individuals reliably on their SL ability, it is also important to rank them reliably on their language proficiency.

In our study we found that the reliability of our SL tasks changed across time. For the ASL task, we never observed sufficient reliability in the 2AFC component to warrant use as an individual differences variable. It appears that ASL for linguistic stimuli is difficult to measure reliably using 2AFC tasks (Arnon, 2020; Kidd et al., 2020; Siegelman et al., 2018), which may at least in part may be due to the fact that existing linguistic knowledge interferes with the decision-making component of the test (though this does not happen for non-linguistic stimuli, Siegelman et al., 2018; Qi et al., 2019). By contrast, ASL measured via the SICR task showed

acceptable reliabilities from time 2 onwards (after we had ironed out features of the task, see Appendix). We recommend that researchers wanting to measure individual differences in ASL for linguistic stimuli consider using the repetition method utilised in SICR.

We found that the performance on the VSL task became more internally consistent over time, approaching comparable reliability to the ASL-SICR task by time 3 and 4 of the study, when children were, on average 7 years and older. We can only speculate on why this might be the case. The growing reliability of the VSL task reported here may be related to the reflection-based processing of 2AFC trials, where children make explicit decisions about implicitly learnt sequences (Christiansen, 2019). When children begin school they may struggle with reflection-based processing, although the evidence is consistent with them becoming more adept with age and school experience, with Von Koss Torkildsen et al. (2019) reporting a Cronbach's α of .81 for the same VSL task in older children using a greater number of test trials. In their discussion of the developmental trajectory of performance on the VSL task, Arciuli and Simpson (2011) suggested that development may be precipitated by the maturing of multiple component skills that contribute to learning. Two cognitive skills that we found were developmentally related to performance on the VSL task were working memory and non-verbal IQ, raising the possibility that reliable performance within individual children is contingent on the development of underlying skills shared across these tasks. One possible candidate is sustained attention (Conway, 2020; Turk-Browne et al., 2005), which may serve a gating function that supports reliable learning of statistical regularities and, ultimately, test performance on the VSL task (see also West, Shanks, et al., 2021). Note that we are not arguing that VSL is contingent on these processes in general, since even neonates and young infants are capable of SL in the visual domain (Bulk et al., 2011; Kirkham et al., 2002). Rather, the suggestion is that reliable performance at the level of the individual may depend upon important maturing skills that allow sustained and controlled attention to stimuli during the task. Sustained attention may be more necessary for statistical learning involving visual stimuli compared with auditory stimuli.

5.2. SL-language link

Against the backdrop of questions concerning the existence of the SL-language effect, it is notable that we observed significant correlations between SL and our language outcome measure only when we had some degree of (though still not perfect) internal consistency in our SL measurement. Our relatively large sample for this literature means that we can also estimate the SL-language relationship as small-to-medium, given the magnitude of the correlations we observed. At the same time, while we observed significant independent contributions of both ASL and VSL to language proficiency, our follow-up analyses showed that SL in the two modalities was differentially related to different sub-components of the language proficiency measure.

The ASL-SICR task significantly predicted Recalling Sentences at time 3 and 4, and significantly predicted Formulated Sentences at time 3. The commonality between these subtests is that they all test language production, as does the ASL-SICR test. However, the notably robust relationship between ASL-SICR and Recalling Sentences suggests a common underlying mechanism supports performance in both tasks. One possibility is that this is the implicit sequencing of linguistic information in the speech production system (Chang et al., 2006). That is, the same process that enables children to acquire and subsequently produce syllable transitions in the SICR task is also implicated in sentence production as measured through recall. This result is consistent with recent work with adults by Isbilen et al. (2022), who showed that performance on an adult version of the SICR task significantly correlated with the recall of lists of highly frequent three-word phrases (e.g., *had a dream*, *in the mailbox*). Their analyses, as did ours, rule out the possibility that the association is due simply to the recall component of the task, since each controlled for short-term recall by incorporating foil repetition into their analyses. The suggestion is that the data from both studies are consistent with the possibility that the ASL-SICR task captures an individual's ability to chunk statistically contiguous linguistic units into larger units (e.g., Christiansen, 2019; Perruchet, 2019; McCauley & Christiansen, 2019), and that these local computations are also relevant for sequencing words during sentence production.

The VSL task was significantly related to the Concepts and Following Directions and Word Structure subtests. One commonality across these CELF subtasks is that they rely heavily on language comprehension, with the Word Structure subtest also involving language production. In addition, we note that each of these CELF subtasks implicate both linguistic and visual processing, requiring children to map language onto visual scenes in order to implement a required behaviour on the basis of language comprehension (Concepts and Following Directions), or to use visual cues to comprehend language and then draw on lexico-grammatical knowledge to produce an appropriate spoken response (Word Structure). Past demonstrations of a link between the *Aliens Attack* VSL task and language have also involved language comprehension tasks using visually presented materials (Kidd & Arciuli, 2016). The associations between VSL and working memory and non-verbal IQ cannot explain fully the result, since the direct paths from VSL to performance on the subtests were significant. Thus, we suggest that the association between VSL and language is contingent upon but not due to the attentional gating we invoked to explain the developmental increase in internal consistency on the task. The source of the association likely lies in mechanisms within and outside the visual system. Performance on VSL tasks is influenced by how efficiently individuals can rapidly encode visual objects, in line with a multi-componential framework of SL (Arciuli & Simpson, 2011; Perfors & Kidd, 2022). This ability may explain overlapping variance in the relationship between VSL and the language tasks, in addition to the ability to identify statistical relations over visual percepts and those that bind visual and linguistic information.

Overall, we have found that ASL and VSL contribute independently to children's language proficiency, but it is important to acknowledge that, depending on the analysis, these relationships are often smaller in magnitude in comparison to both linguistic knowledge and working memory at time 1. This is not too surprising. The nature of most behavioural SL tasks is that they measure very simple statistical distributions using very few stimulus items. This is different from the task facing a learner of a natural language, who is confronted with very different distributions (e.g., Lavi-Rotbain & Arnon, 2022; Stärk et al., 2022a; see Arciuli & Conway, 2018 and Bogaerts et al., 2022, for a discussion) and a multitude of diverse and complicated form-meaning mappings they must master. In our case, we tested our participants' capacities to learn four tri-segment sequences that were statistically defined by perfect (or almost

perfect) within-sequence predictability. Thus, with the current, highly simplified SL paradigms, it is sensible to expect a moderate upper limit on the degree to which any SL task will be related to natural language, if at all. This is not to say that SL is not involved in the acquisition of many of the complexities of natural languages, just that capturing how SL operates across all corners of language learning in one or two standard tasks aimed at measuring individual differences is no doubt impossible. This places important constraints on any study like ours. Understanding the role of SL in language acquisition requires a multi-method approach that not only employs a wider variety of SL tasks that are more closely tailored to the specific distributional patterns relevant for developing specific language skills, but also incorporates statistical analyses of natural language to pinpoint specific distributional patterns to target with these SL tasks. A combination of corpus analyses, computational modelling, and behavioural and neuroscientific investigations is required (e.g., for work on SL and written language using this approach see Arciuli, 2018; Arciuli et al., 2010). For further discussion on some of these points see Arciuli and Conway (2018), Bogaerts et al. (2022), and Isbilen et al. (2022).

One final unexpected result requires explanation. We found that all of our time 1 covariates except age positively predicted time 4 language proficiency, which was a negative predictor. This an interesting though ambiguous result, likely reflecting the presence of unmeasured individual differences in language learning capacity in the cohort. The children spanned a two-year age range, having come from the first two grades of primary school. Thus, it is inevitable that children who differ in age score similarly on other variables like language at the beginning of the study. If this was the case, it would mean that those children who were younger at time 1 but had similar language (or working memory) to older children came into the study learning language at a faster rate. We can only speculate what the cause of this might be, but the end result would be that these younger but faster learners would overtake the older children in linguistic proficiency across the following 18 months, thus leading to the negative association between age and language once the other covariates were held constant.

5.3. Limitations and future directions

Our study has several limitations. Notably, despite our best efforts to measure SL reliably at the beginning of our study, we did not. This result is perhaps not surprising in light of past studies investigating SL task reliability in children (Arnon, 2020; West et al., 2018), but it also meant that we had to update our analysis strategy. This means that, while we have presented evidence in favour of a positive SL-language link, our analyses should be considered exploratory. At the same time, we also take from our data a cautionary note for future work in this space – the establishment of SL task reliability is an essential precondition for any investigation of the SL-language link. With this is mind, it is important to point out that at time 3 and 4 our measurements of SL were hovering around or just below what is be considered acceptable reliability (i.e., $\alpha > .70$), although our good test-retest reliability for both VSL and ASL-SICR across time points 3 and 4 of the study the makes us confident we are capturing something intrinsic to individual performance on these tests. One way to improve internal consistency is to add more test trials. A second limitation is common to any correlational study, longitudinal or otherwise – the 'third variable' problem. We did attempt to control for both existing linguistic knowledge and prominent cognitive skills that may be related to both SL and language at time 1, but it is virtually impossible to control for every possible variable in studies like this. In this vein, we note the relatively high SES of our participants. SES has long been linked to language outcomes (e.g., Hoff, 2003; Justice et al., 2019), but the relatively homogenous nature of the sample may have reduced any effect SES may have had on the data. Interestingly, there is some evidence to suggest that higher SL ability buffers children against the negative effect of low SES on language outcomes (Eghbalzad et al., 2021), and thus SL-language links may be stronger in low SES samples. Additionally, we did not control for visual working memory, which may explain some variance shared between VSL and language proficiency (see Pickering et al., 2022).

We also acknowledge there is ambiguity in our data due to repeat testing of the SL tasks. Notably, it is possible that the increase in reliability we observed across the study is at least in part due to the children's increased familiarity with the SL tasks and, in particular, the trained triplets. Future studies could overcome this problem by testing different triplets at each point, which would be best done using different visual objects and syllables each time. Although we cannot rule out the possibility that repeat testing played a role in the age-related increases in task reliability, we do suspect that the general conclusion that performance does become more reliable with age is correct, since at least for the VSL task we used, greater reliability in older children has been reported (Qi et al., 2019; Von Koss Torkildsen et al., 2019). Stärk et al. (2022b) also reported higher reliability than we reported here using a modified version of the SICR task in 7 – 9-year olds. The reasons why this is the case remain unknown because the relevant studies have yet to be conducted, but are no doubt driven by developmental changes in underlying perceptual and cognitive abilities across the visual and linguistic domains, in addition to reflecting brain maturation in those neural substrates devoted to SL. In addition, as children gain more experience with school their ability to understand and implement task demands improves.

One major development in our sample's cognitive lives was the beginning of reading instruction. Our focus was not on reading, but it is conceivable that the change in children's exposure to language via print significantly alters how children process auditory and visual input (e.g., Dehaene et al., 2010; Favier et al., 2021; Lange-Küttner and Martin, 1999; Van Paridon et al., 2021). This could change how children learn on SL tasks in the auditory and visual domains, an intriguing question that awaits further research. On a more basic level, while SL in the auditory and visual domains has been demonstrated even in newborn infants (Bulk et al., 2011; Teinonen et al., 2009), the pre-existing biases that children bring to SL across perceptual domains and how they change across development is largely unknown. Biases about syllable-groupings (and other linguistic features) in hearing infants no doubt develop early under the sheer weight of linguistic exposure (for evidence these biases influence SL see Stärk et al., 2022b). SL in the visual

domain has similar natural analogues (e.g., face perception, Altvater-Mackensen et al., 2017, with similar critical periods for development, McKone et al., 2019), although it is highly likely that differences in how information is processed in each domain influence any initial conditions and subsequent development (e.g., with auditory learning having a highly temporal component, see Goswami, 2019; Kalashnikova et al., 2021, and visual learning having a both a temporal component and a spatial component). Indeed, Emberson et al. (2019) found that ASL with speech stimuli preceded the development of VSL with faces in 8–10-month-olds.

An additional question concerns whether the repeat testing threatens the reliability of the relationship between both SL tasks and language. We cannot be sure, although at face value repeat testing was only giving children more exposure to the task, and no feedback was given on learning. However, as with the discussion of possible third variables, we cannot rule out that repeated testing did not lead to the introduction of an additional extraneous influence on our results.

6. Conclusion

In this paper we presented, to our knowledge, the first longitudinal study investigating the dual influence of ASL and VSL on children's language development. While SL proved difficult to measure reliably at earlier ages, we found evidence to suggest that SL in both the visual and auditory domains contributes to language development in primary school-age children.

Data Availability

Our data and code are available on the Open Science Framework (https://osf.io/5ad9v/?view).

Acknowledgements

We thank Katherine Revius, Shanthi Kumarage, and Tanya Price for help with testing, Erin Isbilen for her work in developing the ASL-SICR task, Seamus Donnelly for helpful discussion on the data, and Chris Langer-Kuettner and three anonymous reviewers This research was supported by the Australian Research Council awarded to Kidd, Arciuli and Smithson (DP160101917). Order of authorship following the first author is alphabetical.

Appendix. Explanation of change in ASL materials after time 1

At time 1 we used four trisyllabic words taken from Isbilen et al. (2020) (*guliti, dipapu, tagalu, latugi*); however, the foils consisted of scrambled syllable sequences form the same word. For instance, the foil for *guliti* was *gutili* and the foil for *dipapu* was *pupidi.* This made the test component of both the ASL 2AFC and the SICR task more difficult, which meant we did not find significant group effects for learning on the task, and now did we find good internal consistency with either measure. We note that the SICR scores are significantly associated with all Time 1 covariates and with Time 4 language, although we resist interpreting these effects due to the lower reliability of the task when compared to later time points. Before time 2 testing began we changed the nature of the foils in the task so that they came from the same syllable inventory as the target words, but drew separate syllables from each target word (for full details see Kidd et al., 2020). This made the task somewhat easier, and resulted in both group learning effects and more reliable measurement of individual SL ability. The target language used a different set of 'words' in the new task, to avoid any potential interference effects from time 1 testing.

References

Altvater-Mackensen, N., Jessen, S., & Grossman, T. (2017). Brain responses reveal that infants' face discrimination is guided by statistical learning from distributional information. *Developmental Science, 20*, Article e12393. https://doi.org/10.1111/desc.12393

Ambridge, B., & Lieven, E. V. M. (2015). A constructivist account of child language acquisition. In B. MacWhinney, & W. O'Grady (Eds.), *Handbook of language emergence* (pp. 478–510). Hoboken, NJ: Wiley Blackwell.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B, 372*(1711), 1–9. https://doi.org/10.1098/rstb.2016.0058

Arciuli, J. (2018). Reading as statistical learning. *Language, Speech, and Hearing Services in Schools, 49*, 634–643. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0135

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., & Von Koss Torkildsen, J. (2012). Advancing our understanding of the link between statistical learning and language acquisition: The need for longitudinal data. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00324. Article 324.

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. https://doi.org/10.1177/0963721418779977

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language, 63*, 180–196. https://doi.org/10.1016/j.jml.2010.03.005

Arciuli, J., von Koss Torkildsen, J., Stevens, D. J., & Simpson, I. C. (2014). Statistical learning under incidental versus intentional conditions. *Front. Psychol., 5*, 747. https://doi.org/10.3389/fpsyg.2014.00747

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. https://doi.org/10.3758/s13428-019-01205-5

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-olds. *Psychological Science, 9*, 321–324. https://doi.org/10.1111/1467-9280.00063

Australian Bureau of Statistics (2016). Socioeconomic Indexes for Areas, Australia. (https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa). Accessed 26th October 2021.
Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4
Bogaerts, L., Siegelman, N., Christiansen, M. H., & Frost, R. (2022). Is there such a thing as a 'good statistical learner'? *Trends in Cognitive Science, 26*, 25–37. https://doi.org/10.1016/j.tics.2021.10.012
Boyle, W., Lindell, A. K., & Kidd, E. (2013). Investigating the role of verbal working memory in young children's sentence comprehension. *Language Learning, 63*, 211–242. https://doi.org/10.1111/lang.12003
Bulk, 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
Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234–272. https://doi.org/10.1037/0033-295X.113.2.234
Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11*, 468–481. https://doi.org/10.1111/tops.12332
Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences, 39*, Article e62. https://doi.org/10.1017/S0140525X1500031X
Clark, G., & Lum, J. A. G. (2017). Procedural memory and speed of grammatical processing: comparison between typically developing and language impaired children. *Research in Developmental Disabilities, 71*, 237–247. https://doi.org/10.1016/j.ridd.2017.10.015
Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577. https://doi.org/10.1037/0021-843X.112.4.558
Conti-Ramsden, G., Ullman, M. T., & Lum, J. A. G. (2015). The relation between receptive grammar and procedural, declarative, and working memory in specific language impairment. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01090
Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience and Biobehavioral Reviews, 112*, 279–299. https://doi.org/10.1016/j.neubiorev.2020.01.032
Conway, C. M., Arciuli, J., Lum, J. A. G., & Ullman, M. T. (2019). Seeing a problem that may not exist: A reply to West et al.'s (2018) questioning of the procedural deficit hypothesis. *Developmental Science, 22*, Article e12814. https://doi.org/10.1111/desc.12814
Cotton, S., Kiely, P., Crewther, D., Thompson, B., Lay- cock, R., & Crewther, S. (2005). A normative reliability study for the Raven's Coloured Progressive Matrices for elementary school aged children from Victoria, Aus- tralia. *Pers. Individ. Differ., 39*, 647–659. https://doi.org/10.1016/j.paid.2005.02.015
Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 330*, 1359–1364. https://doi.org/10.1126/science.1194140
Dunn, D. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (4th ed.). Minneapolis, MN: Pearson.
Eghbalzad, L., Deocampo, J. A., & Conway, C. M. (2021). How statistical learning interacts with the socioeconomic environment to shape children's language development. *PLoS One, 16*(1). https://doi.org/10.1371/journal.pone.0244954. Article e0244954.
Ellis, E. M., Gonzalez, M. R., & Déak, G. O. (2014). Visual prediction in infancy: What is the association with later language? *Language Learning and Development, 10*, 36–50. https://doi.org/10.1080/15475441.2013.799988
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*, Article 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. https://doi.org/10.1044/1092-4388
Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in spoken language processing. *Journal of Experimental Psychology: General, 150*, 2167–2174. https://doi.org/10.1037/xge0001042
Frost, R., Armstrong, B., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible directions. *Psychological Bulletin, 145*, 1128–1153. https://doi.org/10.1037/bul0000210
Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19*, 117–125. https://doi.org/10.1016/j.tics.2014.12.010
Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C. F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology, 120*, Article 101291. https://doi.org/10.1016/j.cogpsych.2020.101291
Gathercole, S., & Pickering, S. (2001). *Working memory test battery for children (WMTB-C)*. Sydney, Australia: Pearson Assessment.
Gathercole, S. E. (2006). Nonword repetition and vocabulary: The nature of the relationship. *Applied Psycholinguistics, 27*, 513–543. https://doi.org/10.1017/S0142716406060383
Gerbrand, A., Gredebäck, G., Hedenius, M., Forsman, L., & Lindskog, M. (2022). Statistical learning in infancy predicts vocabulary size in toddlerhood. *Infancy, 27*, 700–719. https://doi.org/10.1111/infa.12471
Goswami, U. (2019). Speech and rhythm in language acquisition: An amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences, 1453*, 67–78. https://doi.org/10.1111/nyas.14137
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research Methods, 50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
Hoareau, M., Yeung, H. H., & Nazzi, T. (2019). Infants' statistical word segmentation in an artificial language is linked to both parental speech input and reported speech production abilities. *Developmental Science.* , Article e12803. https://doi.org/10.1111/desc.12803
Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*, 1368–1378. https://doi.org/10.1111/1467-8624.00612
Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition, 225*, Article 105123. https://doi.org/10.1016/j.cognition.2022.105123
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*, Article e12848. https://doi.org/10.1111/cogs.12848
Justice, L., M., Jiang, H., Purtell, K. M., Schmeer, K., Boone, K., et al. (2019). Conditions of poverty, parent–child interactions, and toddlers' early language skills in low-income families. *Maternal and Child Health Journal, 23*, 971–978. https://doi.org/10.1007/s10995-018-02726-9
Kalashnikova, M., Burnham, D., & Goswami, U. (2021). Rhythm discrimination and metronome tapping in 4-year-old children at risk for developmental dyslexia. *Cognitive Development, 60*, Article 101129. https://doi.org/10.1016/j.cogdev.2021.101129
Kalra, P. B., Gabrieli, J. D., & Finn, A. (2019). Evidence of stable individual differences in implicit learning. *Cognition, 190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007
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
Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*(*2*), 154–169. https://doi.org/10.1016/j.tics.2017.11.006
Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E., Revius, K., & Smithson, M. (2020). Measuring auditory statistical learning via serial recall. *Journal of Experimental Child Psychology, 200*, Article 104964. https://doi.org/10.1016/j.jecp.2020.104964
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. https://doi.org/10.1016/S0010-0277(02)00004-5

Kline, R. B. (2005). *Principles and practices of structural equation modelling* (2nd edition). New York: The Guildford Press.
Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2020). Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without Developmental Language Disorder: A conceptual replication and meta- analysis. *Language Learning and Development, 16*, 426–450. https://doi.org/10.1080/15475441.2020.1820340
Lange-Küttner, C., & Martin, S. (1999). Reversed effects of familiarity and novelty in visual and auditory working memory for words. *Brain and Cognition, 40*, 159–162. https://doi.org/10.1006/brcg.1999.1066
Lany, J. (2014). Judging words by their cover and the company they keep: Probabilistic cues support word learning. *Child Development, 85*, 1727–1739. https://doi.org/10.1111/cdev.12199
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
Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition, 223*, Article 105038. https://doi.org/10.1016/j.cognition.2022.105038
Lidz, J., & Gagliardi, A. (2015). When nature meets nurture: Universal Grammar and statistical learning. *Annual Review of Linguistics, 1*, 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236
McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126*, 1–51. https://doi.org/10.1037/rev0000126
McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., Kidd, E., & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports, 9*, 12820. https://doi.org/10.1038/s41598-019-49202-0
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
Montgomery, J. M., Gillam, R. B., & Evans, J. L. (2021). A new memory perspective on sentence comprehension deficits of school-age children with developmental language disorder: Implications for theory, assessment, and intervention. *Language, Speech, and Hearing Services in Schools, 52*, 449–466. https://doi.org/10.1044/2021_LSHSS-20-00128
Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cogn. Psychol., 19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8
Perfors, A., & Kidd, E. (2022). The role of stimulus-specific perceptual fluency in statistical learning. *Cognitive Science, 46*(2), Article e13100. https://doi.org/10.1111/cogs.13100
Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunking in language learning. *Top. Cogn. Sci., 11*, 520–535. https://doi.org/10.1111/tops.12403
Pickering, H. E., Peters, J. L., & Crewther, S. G. (2022). A role for visual working memory in vocabulary development: A systematic review and meta-analysis. *Neuropsychology Review*. https://doi.org/10.1007/s11065-022-09561-4
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*(1), 101–115. https://doi.org/10.1080/10888438.2018.1485680
R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ⟨https://www.R-project.org/⟩.
Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for the Raven's progressive matrices and vocabulary scales*. UK: Oxford Psychologist Press.
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*(4), Article e12593. https://doi.org/10.1111/desc.12593
Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.
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
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805
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
Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime (Version 2.0). [Computer software and manual]*. Pittsburgh, PA: Psychology Software Tools Inc.
Semel, E., Secord, W.A., & Wiig, E.H. (2006). *Clinical Evaluation of Language Fundamentals* (Australian 4th edition). Marrackville, NSW: Harcourt Assessment.
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
Siegelman, N. (2020). Statistical learning theories and their relationship to language. *Language and Linguistics Compass, 14*, Article e12365. https://doi.org/10.1111/lnc3.12365
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., & Frost, R. (2017). 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., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B, 327*(1711). https://doi.org/10.1098/rstb.2016.0059
Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011
Stärk, K., Kidd, E., & Frost, R. L. A. (2022a). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech, 65*, 3–27. https://doi.org/10.1177/0023830920979016
Stärk, K., Kidd, E., & Frost, R. L. A. (2022b). The effect of children's prior knowledge and language abilities on their statistical learning. *Applied Psycholinguistics, 43*, 1045–1071. https://doi.org/10.1017/S0142716422000273
Stojanovik, V., Zimmerer, V., Setter, J., Hudson, K., Poyraz-Bilgin, I., & Saddy, D. (2018). Artificial grammar learning in Williams syndrome and in typical development: The role of rules, familiarity, and prosodic cues. *Applied Psycholinguistics, 32*, 327–353. https://doi.org/10.1017/S0142716417000212
Teinonen, T., Felham, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical learning in neonates revealed by event-related potentials. *BMC Neuroscience, 10*. https://doi.org/10.1186/1471-2202-10-21
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
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
Van Paridon, J., Ostarek, M., Arunkumar, M., & Huettig, F. (2021). Does neuronal recycling result in destructive competition? The influence of learning to read on the recognition of faces. *Psychological Science, 32*, 459–465. https://doi.org/10.1177/0956797620971652
Von Koss 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. https://doi.org/10.1016/j.lindif.2018.11.003
West, G., Melby-Lervåg, M., & Hulme, C. (2021). Is a procedural learning deficit a causal risk factor for developmental language disorder or dyslexia? A meta-analytic review. *Developmental Psychology, 57*, 749–770. https://doi.org/10.1037/dev0001172

West, G., Shanks, D. R., & Hulme, C. (2021). Sustained attention, not procedural learning, is a predictor of reading, language and arithmetic skills in children. *Scientific Studies of Reading, 25*(1), 47–63. https://doi.org/10.1080/10888438.2020.1750618

West, G., Vadillo, M. A., Shanks, R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: we see some problems. *Developmental Science, 21*, Article e12552. https://doi.org/10.1111/desc.12552

Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences of the United States of America, 119*, Article e2021865119. https://doi.org/10.1073/pnas.2021865119