*COGNITION*, in press

# Does short-term memory develop?

Gary Jones[1]

Lucy V. Justice[1]

Francesco Cabiddu[1]

Bethany J. Lee[1]

Lai-Sang Iao[1]

Natalie Harrison[1]

Bill Macken[2]

[1] Nottingham Trent University (UK)

[2] Cardiff University (UK)

Corresponding author: Gary Jones, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ (UK). Telephone: +44 (115) 848 2422; E-mail: gary.jones@ntu.ac.uk.

GJ & BM contributed to the conceptual development of the work and writing of the article. GJ, BM, LJ, LI, & FC contributed to the methodological development of the study design. GJ performed the model simulation. Testing and data collection were performed by LJ, FC, BL & NH. LJ & FC performed the data analysis. All authors approved the final version of the manuscript for submission.

Word count: 9,457

## Abstract

Such is the consistency by which performance on measures of short-term memory (STM) increase with age that developmental increases in STM capacity are largely accepted as fact. However, our analysis of a robust but almost ignored finding – that span for digit sequences (the traditional measure of STM) increases at a far greater rate than span for other verbal material – fundamentally undermines the assumption that increased performance in STM tasks is underpinned by developmental increases in capacity. We show that this digit superiority with age effect is explained by the relatively greater linguistic exposure to random sequences of digits versus other stimuli such as words. A simple associative learning process that learns incrementally from exposure to language accounts for the effect, without any need to invoke an STM mechanism, much less one that increases in capacity with age. By extension, using corpus data directed at 2-3 year old children, 4-6 year old children, and adults, we show that age-related performance increases with other types of verbal material are equally driven by the same basic associative learning process operating on the expanding exposure to language experienced by the child. Our results question the idea that tests such as digit span are measuring a dedicated system for the temporary maintenance and manipulation of verbal material, and as such have implications for our understanding of those aspects of typical and atypical development that are usually accounted for with respect to the operation of such a system.

Keywords: Short-term memory; digit span; child development; associative learning; capacity.

## Introduction

“The notion that age differences in [verbal short-term memory] reflect structural changes in capacity has for many the status of an established fact that does not require formal endorsement” (Dempster, 1981, p.85).

“Individuals differ in capacity, which ranges from about two to six items in adults (and fewer in children)” (Cowan, 2008, p.330).

Ingrained within the developmental psyche are limits to cognitive performance in a given setting that are almost universally argued to be fundamentally underpinned by a limited capacity short-term memory (STM) system. Changes in STM performance across individuals and time are typically argued to be due, at least in part, to changes in that limited processing capacity. Empirically, such explanations are informed by the range of associations between performance on typical STM tasks and performance on higher level cognitive functions. Such is the consistency by which performance on STM tasks increases with age that it is assumed to play a developmental role in almost every domain that involves some form of information processing. This view sees STM as a processing *primitive*: a fundamental cognitive system whose operation constrains and determines higher functions built upon its operations.

Studies of STM have largely focused on verbal STM (vSTM), where a short list of verbal material (e.g., syllables, words, digits) is presented and must be reproduced in its original order. There is a consistent pattern to every developmental study involving vSTM measures: increases in recall accuracy until adulthood (e.g., Bopp & Verhaeghen, 2005, Nettelbeck & Burns, 2010). Whether the focus is within or across ages, a wealth of studies shows associations between vSTM and a wide range of developmental tasks, implicating STM in domains ranging from language learning (Gathercole, 2006), mathematics (Stipek & Valentino, 2015), and general intelligence (Hornung, Brunner, Reuter & Martin, 2011), together with involvement in atypical development (Helland & Asbjørnsen, 2004) and neuropsychological conditions (Hinton, De Vivo, Nereo, Goldstein & Stern, 2001).

In this paper we present a parsimonious yet contentious explanation for observed developmental increases in vSTM tasks: associative learning that takes place within the linguistic environment of the rememberer. Associative learning is a process by which new knowledge is gained based on perceived associations between two or more stimuli or events. Contrary to traditional explanations of age-related increases in vSTM outlined above, an associative learning account implies that developmental changes in performance on vSTM tasks do not arise from increasing capacity of dedicated vSTM mechanisms, but rather from increasing knowledge of the sequential structure of language that is acquired via the expanding linguistic experience of the child. Associative learning is therefore heavily embedded in familiarity – items and item sequences that are often encountered in the language environment will be more amenable to learning than those that are not. The potential breadth of this learning is vast: children as young as 9 to 15 months hear around half a million words in a three-week period (Swingley, 2007) and so rather than linguistic stimuli being familiar or unfamiliar, familiarity is a question of degree based on the rememberer's experience with their native language. Moreover, since the linguistic domain involves sound sequences (sublexical knowledge), words (lexical knowledge), and sentences/utterances (word sequences), familiarity will play a role across a range of scales of knowledge.

Examining the influence of familiarity with the sequential structure of STM test stimuli requires an estimate of exposure to the particular stimuli. Typically in vSTM research, participants' familiarity with the verbal material is characterized by establishing parameters for individual items or item sets (as opposed to sequences) such as frequency of occurrence (Turner, Henry, Smith, & Brown, 2004), phonological neighborhood density (Roodenrys & Hinton, 2002), or wordlikeness of nonwords (Archibald & Gathercole, 2006). However, large-scale samples of child- and adult-directed speech are now available that provide opportunity for more ecologically grounded estimates of the linguistic experience of the rememberer. We use such large-scale corpora here, enabling investigation of the relationship between linguistic knowledge and the development of vSTM performance in a novel way.

The main task we use is digit span, the archetypal measure of vSTM in children, appearing in almost every standardized test battery that includes a vSTM measure (e.g., Elliot & Smith, 2011; Kaufman & Kaufman, 2004; Wechsler, 2004, 2008, 2009) and more generally outnumbering other vSTM measures by a factor of at least 16:1 (G. Jones & Macken, 2015). The task involves lists of digits being read aloud for subsequent recall by participants. Studies universally show that digit span increases with age until adulthood (Bopp & Verhaeghen, 2005; Hale, Bronik & Fry, 1997; Salthouse, Mitchell, Skovronek & Babcock, 1989) and since this performance also correlates with an assortment of other areas of cognitive function (Stipek & Valentino, 2015; Helland & Asbjørnsen, 2004), vSTM is ascribed a fundamental role in child development.

Digit span exhibits an interesting - and, it turns out, theoretically significant - phenomenon when considered developmentally: digit span increases with age at a far greater rate than span for other stimuli such as words or nonwords. For example, in Dempster’s (1981)[1] meta-analysis involving studies that have used digit span and word span, digit span increased from 4.3 to 6.5 between the ages of five and twelve, whereas word span increased from 4.0 to 4.6. We call this age-related performance advantage for digits over other stimuli ‘digit superiority with age’ and clearly such a phenomenon poses an issue for capacity-based explanations of development given the increasing discrepancy in the estimate of capacity depending on the stimulus type. Since Dempster's (1981) meta-analysis, developmental

[1] Note that we could not find more recent meta-analyses that directly compared digit span and word span across developmental ages.

studies using digit span have increased exponentially whereas those using word span are comparatively rare.

Typically, digit superiority with age has been explained with reference to factors such as digits forming a homogenous set and occurring with greater token frequency than words within natural language. In this way the typical account focuses on mechanisms operating on the processing of the individual lexical item, in much the same way as typical explanations of, for example, lexicality and word frequency are based on more robust processing of lexico-phonological representations of individual items as a function of long-term language knowledge (e.g., Gathercole, Pickering, Hall & Peaker, 2001; Roodenrys, Hulme & Brown, 1993; Stuart & Hulme, 2000. But see also Macken, Taylor & Jones, 2014). Regardless of the veracity of these explanations, the fact that the typical measure of STM capacity is one that shows a different developmental trajectory than would be found should another stimulus type be used at least raises a cautionary note about how such capacity increases should be interpreted.

However, there is also a more fundamental theoretical issue at stake here: namely the role that development of a capacity-limited STM system (as indexed by span) plays in the development of cognitive function. A growing body of research suggests that, rather than construing STM as a primitive processing *system*, it is better thought of as a particular *setting* within which participants must deal with a relatively novel situation by flexibly co-opting whatever skills and knowledge they have to accomplish the particular task goals (e.g., G. Jones & Macken, 2015; D. M. Jones, Macken & Nicholls, 2004; Macken & D. M. Jones, 2003; Macken, Taylor & D. M. Jones, 2015; see also, e.g., Craik & Lockhart, 1972; Ericsson & Kintsch, 1995). Such an approach turns the theoretical focus away from considerations of assessing the limited capacity of putative primitive processing systems and towards a consideration of the degree of convergence or divergence between the participants' skills and knowledge and the parameters (task requirements, material, etc.) of the particular STM setting.

Our hypothesis is that associative learning may account for the digit superiority with age effect based solely on knowledge gained from the expanding experience of sequential structure inherent within language as a function of age. This approach differs from the typical focus in vSTM theory on merely item-level factors, as mentioned above (e.g., Gathercole, et al., 2001; Roodenrys et al., 1993; Stuart & Hulme, 2000). An account that focuses on experience (for vSTM, that experience represents associative learning taking place in the linguistic environment) has no need to assume existence of any bespoke vSTM mechanism, nor any need to invoke additional mechanisms to explain phenomena such as digit superiority with age. Under this view, digit span increases at a greater rate than span for other stimuli because, since such sequences are unconstrained by the syntax that typically constrains word order, experience of random sequences of digits (e.g., from numeracy lessons, soccer scores, dates, addresses, etc.) increases at a greater rate with age than does experience of random sequences of other stimuli. This would also explain more muted though robust increases in span for other stimuli; for example, while random sequences of words may not appear often in natural language, their occurrence inevitably increases with age. Along with digit span, therefore, we also examine word span, where an associative learning account predicts a shallower developmental trajectory to that of digit span.

Some types of evidence have been used to make the claim that changes in capacity per se, rather than mere increases in task relevant knowledge, are needed to account for improved performance in STM tasks. For example, Cowan, Ricker, Clark, Henrichs and Glass (2015) compared probed-remembering of matrices of either ‘familiar’ (i.e., letters) or ‘unfamiliar’ (i.e., abstract shapes) items. They revealed a developmental increase in performance across four age groups (7 years to 24 years) for both types of item with the rate of increase for the familiar items being greater than that for the unfamiliar. However, when comparing normalized capacity estimates based on this performance data, the capacity estimates increased at the same rate for both letter and shape stimuli. Cowan et al. concluded that familiarity of the items – particularly in the case of familiar letters – could not explain these developmental increases, supporting a view of age-related increases in STM capacity. One questionable assumption underlying this conclusion is that changes in performance for the abstract shape stimuli should have little or no contribution from increasing knowledge with age. However, familiarity versus novelty is always a matter of degree, and it seems at least possible, contrary to the assumption of Cowan et al. (2015), that knowledge of nonverbal forms will increase over the examined age span, albeit at a slower rate than that for verbal stimuli. For example, increasing experience of lines, angled lines and circles (as used for 'unfamiliar' shapes by Cowan et al., 2015) as well as increasing knowledge of more objects with more and varied visual forms coupled with an increasing fluency in naming visuo-spatial stimuli, will lead to efficiencies in the representation of such items over developmental time. From this point of view, the increase in performance with both types of stimuli may well be due to increasing knowledge, but knowledge that is increasing at different rates for different stimuli. Thus the alignment of normalized development trajectories across stimuli, rather than revealing an underlying increase in capacity, may in fact reveal similar processes of knowledge acquisition that is merely accruing at different rates for different materials, depending on the extent to which those materials are encountered in the child's environment. Here we will be able to evaluate which of these possibilities is most likely since the model we present below quantifies the graded levels of familiarity with different types of verbal stimuli, based solely on their presence within the linguistic experience of the child, without any supplementary assumptions about changes in underlying mechanisms such as STM capacity.

Our purpose is threefold: first, to investigate an associative learning account of increases in span performance with age without recourse to the idea of a limited (and developmentally increasing) STM capacity; second, to apply this account to differences in the developmental trajectories of digit and word span; and third, to manipulate the digit span task to more directly test whether such differences arise because of inherent properties of the stimulus items (e.g., digits versus words) or more plausibly arise from our associative learning account. To accomplish this, we use a computational model of associative learning that is trained on language samples that estimate the linguistic input that 5-10 year old children receive. In Study 1a, we compare digit and word span in children and model and in Study 1b we test whether performance is governed by inherent characteristics of digits by using mixed lists (lists containing a mixture of digits and words). In Study 2, we show that our model results appear in every language sample that we used. Before detailing the studies, we outline both the language samples used for our modeling environment and the modeling environment itself.

## Language samples and modeling environment

### *Language samples used*

In order to have a developmentally appropriate sample of language, we used child-directed speech from the Manchester corpus aimed at 2-3 year old children (306,831 utterances, Theakston, Lieven, Pine & Rowland, 2001) and child-directed speech and children’s literature aimed at 4-6 year old children (ratio spoken:written is 2:1, 75,981 utterances, see G. Jones, 2016). However, we could not find large quantities of child-directed speech for 7-10 year old children and hence used adult-directed speech and literature from the British National Corpus (BNC, 2007). Approximately 9 of the 10 million BNC utterances are taken from textual sources, so we pseudo-randomly sampled 2 million utterances to ensure a spoken:written ratio of 1:1 to reflect the likelihood that a significant amount of older children's language input will be speech-based. For ease of reference, we use 'utterance' to refer to both a spoken utterance and a written sentence. All utterances were converted to their phonemic equivalent using the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). In all cases, phonemic utterances were word-delimited on the basis that children above the age of 2 years can readily identify word boundaries (see Rowland, 2014, for a review).

### *Computational modeling environment*

The computational modeling environment is based on CLASSIC (Chunking Lexical and Sub-lexical Sequences In Children; G. Jones, 2016; G. Jones *et al.*, 2014; G. Jones & Macken, 2015, 2018; G. Jones & Rowland, 2017) that incrementally learns phoneme and word sequences from the type of language samples described above. For ease of reference, we will call any knowledge in the model (e.g., a learned phoneme sequence) a 'chunk'. Our assumption is that sequence learning in the model reflects associative learning that may arise from a range of linguistic experience, including perceptual learning of sound sequences on the basis of their frequent occurrence in natural language as well as increasing proficiency in the execution of the sequences of articulatory gestures required for production of extended utterances. CLASSIC has been focused on the language domain because language offers the opportunity to estimate the type of sequence knowledge that one may experience (e.g., using child-directed language samples).

At the outset, the only chunks in the model are the 44 phonemes of standard British English. Each input utterance is presented to the model one at a time, in the order that utterances appear in the language sample. The model processes each utterance as follows: (1) encode the utterance into the fewest chunks possible based on current (chunk) knowledge; (2)

learn new chunks by joining sequentially adjacent chunks from the encoded utterance, with the proviso that learning across word boundaries only occurs if the adjacent chunks are themselves either words or word sequences. The model therefore learns incrementally longer chunked sequences (initially involving phonemes, but later involving words) as it progresses through the input, developing an expanding repertoire of linguistic knowledge in the form of an increase in both the number and size of chunks as exposure to the language increases.

As learning progresses, the model is likely to be able to encode each utterance more efficiently (i.e., needing fewer chunks to represent the whole utterance). Therefore, the model represents language *knowledge*, rather than being, for example, a model of language production or perception, per se. It is this characteristic that allows us to investigate our question – broadly, to what extent can the development of vSTM be accounted for with reference to growth of language knowledge, without invoking developmental changes in the processing capacity of an underlying STM system. As such, for any given input, the measure of *performance* of the model is the number of chunks it needs to encode that input, with fewer chunks corresponding to better performance, as a function of increasing knowledge[2]. Since our model learns incrementally, STM test stimuli can be presented to it for every 5% increment in the corpus input so that developmental progression can be examined and compared to 5-10 year old children.

Starting from afresh, if the model was presented with the utterance ‘what’s this?’ (coded phonemically as W, AH, T, S / DH, IH, S), it would encode the input as seven chunks (one for each phoneme) and would learn five chunks (W AH, AH T, T S, DH IH, IH S). If presented a second time with exactly the same input, it would now be encoded using only

[2] In this respect, our modeling application deviates substantially from the typical approach to modeling vSTM which has sought to simulate a variety of factors affecting patterns of STM, especially serial recall performance (see Hurlstone, Baddeley & Hitch, 2014 for an overview). Clearly, a key area for extension of our approach will require modeling of the process whereby the knowledge acquired by the model is converted into output/production within the task setting, which, as well as motivating further theoretical development, will also enable more fine-grained comparison of model and participant performance.

four chunks (W AH, T S / DH IH, S) and the model would subsequently learn two new chunks, both corresponding to words (W AH T S, DH IH S). Presenting the utterance a third time would lead to the utterance being encoded using only two chunks (one for each word), and learning would now be allowed to cross the word boundary such that a word sequence is learned (W AH T S DH IH S). CLASSIC learns at every opportunity; one would not expect children to do so, but this is a practical consequence of the model receiving a very small amount of language as input compared to children. In Table 1 we show how learning progresses when the first three utterances are different to one another.

Table 1. How learning progresses in the model when presented with three different utterances, one after the other. The model begins with the phonemic inventory for standard British English (for ease of exposition we use the CMU Pronouncing dictionary phonemes). Word boundaries are represented using ‘/’ and chunks are separated using ‘,’.The model only learns across word boundaries if the chunks involved are words or word sequences.

| Utterance | Encoded sequences | Chunks learned |
|---|---|---|
| [1] Where is she? | W, EH, R / IH, Z / SH, IY | W EH, EH R, IH Z, SH IY |
| (W EH R / IH Z / SH IY) | Encoded using 7 chunks | 4 chunks learned |
| [2] Is she there? | IH Z / SH IY / DH, EH R | IH Z SH IY, DH EH R |
| (IH Z / SH IY / DH EH R) | Encoded using 4 chunks | 2 chunks learned |
| [3] No she isn’t | N, OW / SH IY / IH Z, N, T | N OW, IH Z N, N T |
| (N OW / SH IY / IH Z N T) | Encoded using 6 chunks | 3 chunks learned |

CLASSIC has previously been applied to a number of language-related domains. G. Jones (2016) simulated children’s nonword repetition performance using the 2-3 year old and 4-6 year old inputs in the current paper. Across six different studies involving six different nonword sets and children of two to six years of age, the model was able to fit over 85% of child datapoints purely as a function of increased exposure to the linguistic input. G. Jones and Macken (2018) again used the same input sets, this time showing how the number of chunks required by the model to encode digit lists, word lists, nonsense words, and sentences predicted the performance of six year old children for the same stimuli. Finally, G. Jones and Rowland (2017) used the modeling environment and the 2-3 year old input to investigate differences in quantity and diversity of language exposure, showing how quantity was important early in learning with diversity important thereafter, consistent with findings from children (Rowe, 2012). CLASSIC has therefore successfully been applied across a range of language-related domains, representing a good foundation for the examination of age-related increases in span performance.

## Study 1: vSTM performance in the model and in 5-10 year old children

Both study 1a and 1b were approved by the College Research Ethics Committee of Nottingham Trent University. All analyses were run in R Version 3.5.2 (R Core Team, 2018), using the packages “lme4” (1.1-21; Bates, Mächler, Bolker, & Walker, 2015) and “ordinal” (version 2019.12-10; Christensen, 2019) for statistical analyses.

### *Model language samples, study 1a and study 1b*

30 input files were produced to ensure that any model effects were not because of a fortuitous random sample. Each input file contained 75,981 randomly sampled utterances from the 2-3 year old input, a random ordering of the 75,981 utterances in the 4-6 year old input, and 75,981 randomly sampled utterances from the 2 million BNC input. Each input file was ordered so that the beginning contained 80% of 2-3 year old input, the middle contained 80% of 4-6 year old input, and the remainder contained 80% of the adult input. The remaining 20% in each third was made up from the other samples (10% each). The rationale

here is simply to reflect the language input of the particular children. For example, a child of 2-3 years of age will not solely hear linguistic input directed at 2-3 year old children. The model was run individually for each of the 30 input files.

*Study 1a: Model and child performance for digit and word lists*

*Child participants*

143 5-10 year old children (*M* = 105.29 months [range 66 - 128], 63 boys) were recruited from five schools in the Nottinghamshire area using UK year 1 (5-6 years of age), year 3 (7-8 years of age) and year 5 (9-10 years of age) classes.

*Digit and word lists*

The basis for the digit and word lists was the digit span task from the Wechsler Intelligence Scale for Children, 4th Edition (WISC4, Wechsler, 2004) since these dominate assessment of vSTM. Participants hear increasingly longer lists of digits (e.g., *nine-five-two-seven*) that they must repeat back in their correct order, with the longest list length accurately recalled being their ‘span’. There are two digit lists at each of eight lengths (i.e., two to nine digits).

Word lists were produced by yoking a word to each digit and substituting digits for their corresponding words in all of the WISC4 digit lists (e.g., *three-eight-two* becomes *door-boat-water*). Words were all highly imageable object nouns to mitigate any effect of digits forming a homogenous group and of digits existing in both verbal and numeral forms. Additionally, the average frequency of the words was substantially greater than that of the digits, while phonemic and syllabic word length was equal across digits and words. Note that no attempt was made to separate semantically associated words that may facilitate recall (e.g., *boat-water* in the example). Such associations act *against* our hypothesis and therefore any superiority for digits over words is over and above semantic relations that may exist between words.

Table 2 shows the key characteristics of the digits and words used in span lists. Note that although digits and words were comparable on key characteristics, this was not on a one-to-one basis but as an average across the whole set of items due to the digit ‘one’ being of such high frequency that no highly imageable object noun of similar frequency was available.

Table 2. Characteristics of the digits and words that comprise digit lists, word lists, and mixed lists (frequencies taken from the Children’s Printed Word Database that contains frequencies of words aimed at five to nine year old children, Masterson, Stuart, Dixon & Lovejoy, 2010).

| Digit | Syllables/Phonemes/ Frequency | Word Match | Syllables/Phonemes/ Frequency |
|---|---|---|---|
| one | 1/3/3069 | House | 1/3/1880 |
| two | 1/2/1114 | Water | 2/4/1525 |
| three | 1/3/706 | Door | 1/3/857 |
| four | 1/3/276 | School | 1/4/1393 |
| five | 1/3/173 | Bear | 1/2/1214 |
| six | 1/4/103 | Bed | 1/3/771 |
| seven | 2/4/70 | Food | 1/3/925 |
| eight | 1/2/41 | Boat | 1/3/563 |
| nine | 1/3/38 | Cup | 1/3/216 |
| | *M* = 1.1/3.0/621.1 | | *M* = 1.1/3.1/1038.2 |

*Design*

For the model, the independent variables were stimulus type (digits or words) and amount of input seen (5% increments), with the dependent measure being the average number of chunks required to encode digit and word lists.

For the children, the independent variables were stimulus type (digits or words) and age (in months); the dependent variable was span size (longest list length accurately recalled) since this was used by Dempster (1981). The same results apply for total span (number of lists accurately recalled; see Appendix: Analyses - Study 1a).

*Procedure*

For children, the WISC4 instructions were followed for both digit lists and word lists so that the study was as similar as possible to previous studies involving digit span. Presentation of digit lists and word lists was counterbalanced. Each of the two lists of a particular length was spoken aloud separately (with a short pause between each item), starting at list length two. Children were asked to repeat the list immediately after presentation, with responses being recorded manually. Children only proceeded to the next list length if they correctly recalled one of the two lists correctly. Span size was recorded as the longest list length accurately recalled (e.g., a span size of 5 if at least one list of length 5 was recalled correctly but no lists of length 6 were recalled correctly).

*Results*[3]

As Figure 1 shows, while the number of chunks that are required by the model to encode digit lists and word lists reduces as the amount of input to the model increases, this is more pronounced for digit lists than word lists. The associative learning account therefore suggests: (a) span performance will increase with age; (b) digit span performance will be

[3] Our statistical models could be improved by including item as a random effect. These analyses do not change the digit superiority with age effect and are included in the repository where our data reside.

superior to word span performance; and (c) a digit superiority with age effect whereby digit span will increase with age at a far greater rate than word span.

The child data in Figure 1 were examined using two ordered categorical mixed-effects models to analyze span size[4]. We did this because our focus was on the digit superiority with age effect (i.e., the interaction) and we therefore wanted to examine the additional model fit that the interaction term supplies. These models were run using the R package “ordinal” (Christensen, 2019). Model 1 included stimulus type (digits, words) and age (months) as fixed effects. Model 2 additionally included their interaction as a fixed effect. A maximal random effects structure was included for participant for both models. Model 1 showed that span increased with age ($p$ = .001) and was greater for digit lists over word lists ($p$ < .001). The effect of stimulus type disappeared in Model 2 ($p$ = .461), however age remained significant ($p$ = .026) and the interaction was also significant ($p$ = .039). Model 2 provided a significantly better fit to the data than Model 1 ($p$ = .036) indicating a digit superiority with age effect within the children’s performance that supports the associative learning account (see Appendix: Analyses - Study 1a).

However, while the model explains performance differences in terms of greater experience with random sequences of digits rather than inherent characteristics of digits themselves, the latter could still explain children’s performance. While we made every effort to match digits to words on key criteria, it may still be plausible that digits differ from words on some other criteria that we have not considered. We therefore detail Study 1b that shows how these explanations can be teased apart by using mixed lists. Mixed lists will allow us to compare recall of individual digits versus individual words (i.e., digits and words that are not part of a digit sequence or word sequence) and to compare recall of digit sequences versus

[4] Categorical rather than linear models were used because span size is bounded, with scores either being 0 or from 2 to 9 inclusive (none of our participants scored 0). The same effects are seen when treating span as continuous.

word sequences. If digit superiority with age is due to inherent characteristics of digits over words as individual lexical items, then performance should be superior for both individual digits over individual words and for digit sequences over word sequences.

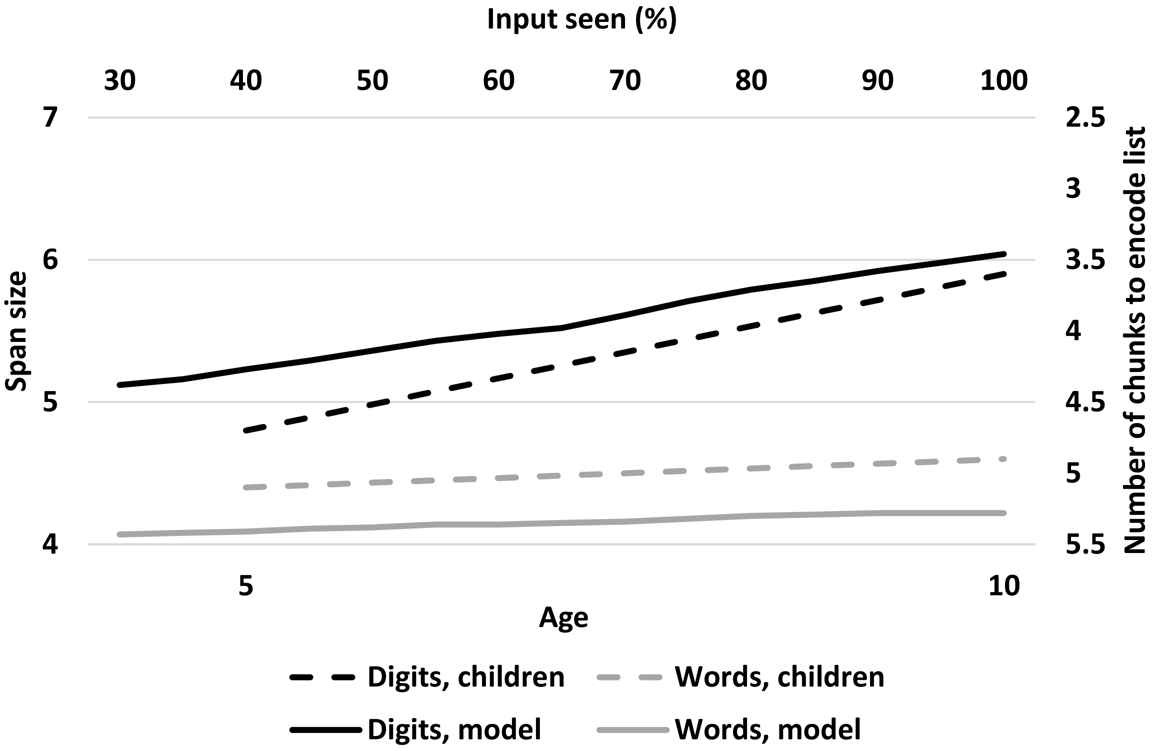

Figure 1. Number of chunks required to encode digit lists and word lists as the model's input increases (from 30%-100%), together with span size (longest list length accurately recalled) for the 5-10 year old children in study 1a (N = 143) based on the regression line (see Appendix: Analysis - Study 1a for a more in-depth view of the child data including variance across the regression line).

*Study 1b: Model and child performance for mixed lists*

*Child participants*

70 5-10 year old children (*M* = 92.40 months [range 60-131], 34 boys) were recruited from two schools in the Nottinghamshire area using classes ranging from year 1 to year 5.

Note: Three additional children were tested but excluded from analyses because they were 11 years of age (age in months 138, 140 and 143).

*Mixed lists*

Four novel digit lists were created (two of five items, two of six items). These were then converted into mixed lists by substituting particular digits for their corresponding word such that the lists contained isolated digits and isolated words – digits or words that are not flanked by another item of the same type – together with digit sequences and word sequences. Mixed lists were in pairs such that one list contained isolated digits or digit sequences at the same serial position as the isolated words or word sequences in the other list (e.g., *three-two-five-nine-four-eight* became *door-two-five-cup-four-eight* and *three-water-bear-nine-school-boat*). We did this to ensure that any effects seen are not because of serial position effects that are renowned in serial recall tasks (Grenfell-Essam & Ward, 2012). There were therefore eight mixed lists (see Table 3), four of five items and four of six. These lengths were chosen in order to both be challenging for children and to allow sufficient numbers of isolated digits (e.g., *three*), isolated words (e.g., *door*), digit sequences (e.g., *two-five*) and word sequences (e.g., *water-bear*).

In total there were seven isolated digits and words, and eight digit and word pairs. Although the digit lists in the WISC4 have 72 possible pairs of digit sequences (2 lists at each of 8 list lengths from 2 to 9), we managed to ensure that the majority (5 of 8) digit pairs in the mixed lists did not occur in the WISC4 digit lists while not forming sequential runs (e.g., *one-two*) or familiar triplets (e.g., *two-four-six*). Each digit and word appeared at least twice and at most three times in the mixed lists. There were equal numbers of each digit and its corresponding word, and an equal number of digits and words appeared in each serial position.

If digit superiority with age is due to inherent characteristics of digits over words, then digit recall should increase at a faster rate than word recall, regardless of whether it is assessed as recall of items in isolation or sequence. In the above example, this would mean that both *three* and *two-five* would be expected to be recalled more accurately with age than *door* and *water-bear*. If, however, the steeper developmental trajectory for digit span is due to associative learning based on greater age-related exposure to random *sequences* of digits than random *sequences* of words, the digit superiority with age effect should only be found for sequence recall.

Table 3. Mixed list stimuli comprising paired lists whereby one list in the pair contains isolated digits and/or digit pairs at the same serial position as isolated words and/or word pairs in the second pair. There are four lists of five items in length and four lists of six items in length.

| Pair | Mixed lists |
|---|---|
| 1 | seven two house door five |
| 1 | food water one three bear |
| 2 | boat one six nine food |
| 2 | eight house bed cup seven |
| 3 | four nine boat three six food |
| 3 | school cup eight door bed seven |
| 4 | door two five cup four eight |
| 4 | three water bear nine school boat |

*Design*

For the model, the independent variables were stimulus type (digits or words) and amount of input seen (5% increments), with the dependent measures being the average number of chunks required to encode mixed lists plus the average number of chunks required to encode isolated digits, isolated words, digit pairs and word pairs. Recall of sequences is assessed at the level of item pairs, since this is the smallest (supra-lexical) sequential unit (i.e., *one-six-nine* is two paired sequences, with *one-six* as the first pair and *six-nine* as the second pair).

For children, the independent variables were stimulus type (digits or words) and age (in months). The dependent variables were the number of mixed lists accurately recalled plus the number of isolated digits, isolated words, digit pairs, and word pairs accurately recalled.

*Procedure*

For children, each digit and word was recorded individually as an MP3 file and spliced together (with a short pause between each item) to produce the mixed lists. Children were placed in a quiet room near their classroom and were presented with all lists one at a time via headphones. Lists were presented in a randomized order for each child, with the child being asked to repeat each list immediately after they heard it. Responses were recorded manually.

*Results*

The model data in Figure 2 show that as the linguistic input increased, the model was able to encode mixed lists using fewer chunks. The child data in Figure 2 mirror those of the model, with accurate recall of mixed lists increasing with age (*p* = .001, see Appendix: Analyses - Study 1b).

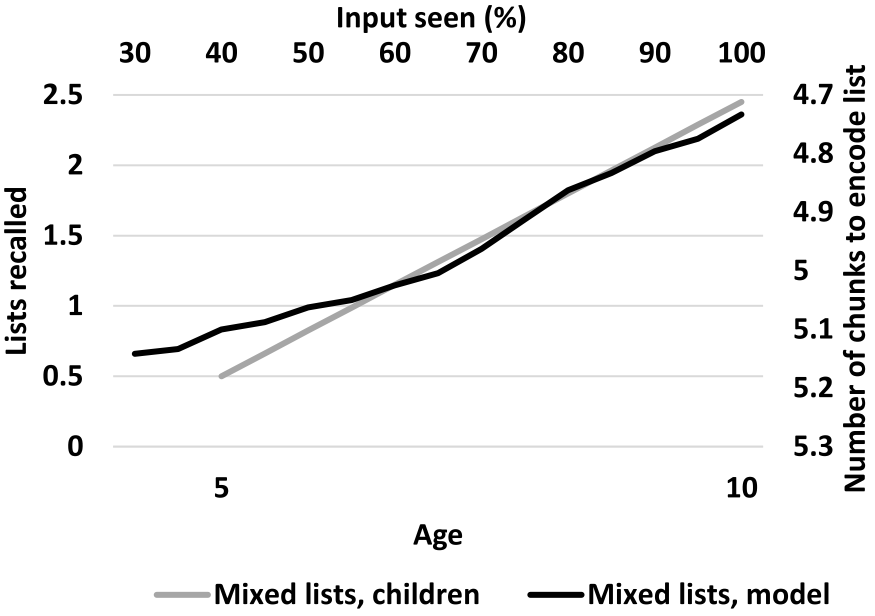

Figure 2. Number of chunks required to encode mixed lists as the model's input increases (from 30%-100%) together with the number of lists accurately recalled by the 5-10 year old children in Study 1b (N = 70) based on the regression line (see Appendix: Analysis - Study 1b for a more in-depth view of the child data including variance across the regression line).

The item-level model performance (see Figure 3) shows that marginally fewer chunks are required to encode isolated words than isolated digits across the whole input[5]. The model therefore suggests that the digit superiority with age effect is not caused by inherent characteristics of digits themselves. However, digit pairs are encoded using fewer chunks than word pairs across the whole input and more importantly, there is a pronounced reduction over time in the number of chunks required to encode digit pairs compared to word pairs. Put simply, over time the model learns more chunks containing digit sequences than word sequences.

[5] The marginally greater reduction in encoding of isolated words over isolated digits is because some words must exist within a chunk that includes one or both digits in a digit pair e.g., *one cup* (otherwise there would be no difference between isolated words and isolated digits).

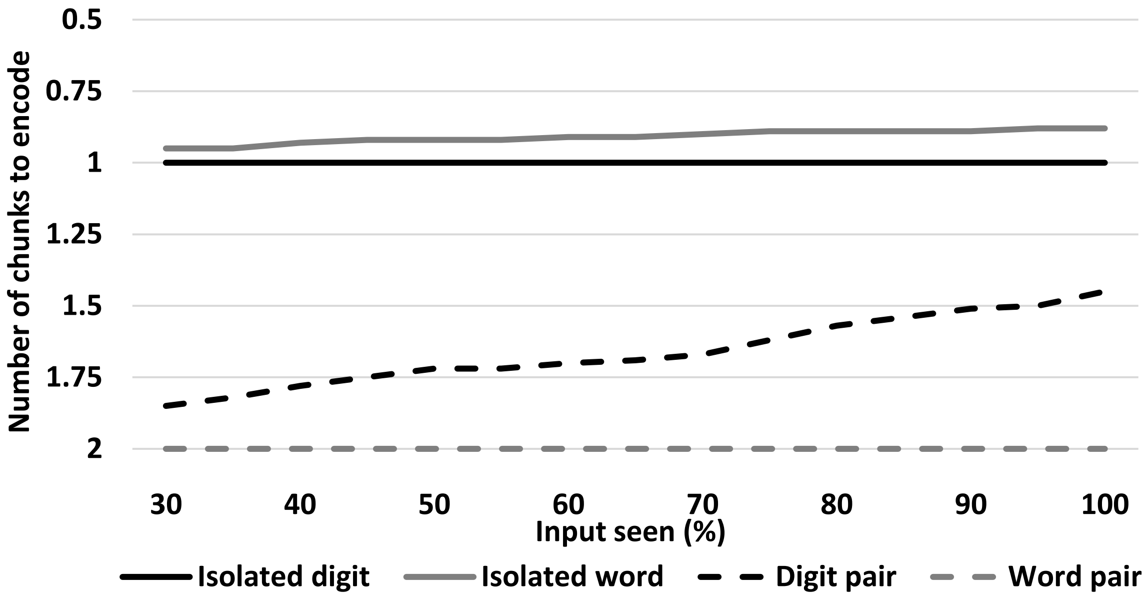

Figure 3. Number of chunks required to encode isolated digits, isolated words, digit pairs, and word pairs as the model's input increases (from 30%-100%).

Figure 4 shows the child data for the component items within the mixed lists. We used the same analysis format as per the digit and word lists, with analyses carried out using lme4 (Bates et al., 2015) and outlined in full in the Appendix: Analyses - Study 1b.

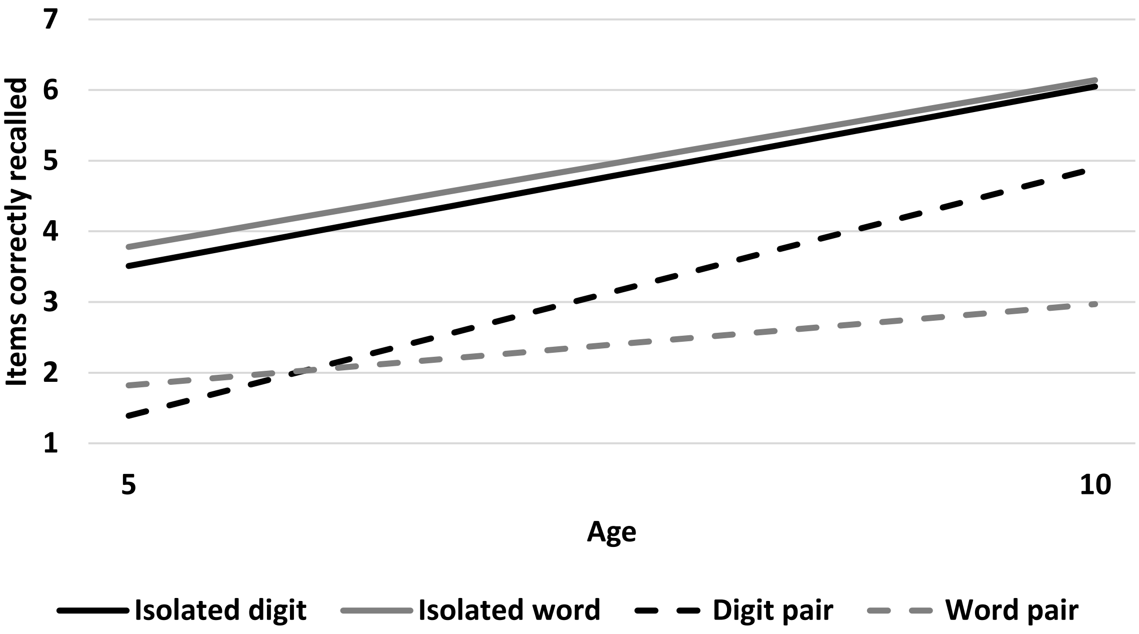

Figure 4. Number of items correctly recalled for isolated digits, isolated words, digit pairs, and word pairs for the 5-10 year old children in Study 1b based on the regression line (see Appendix: Analysis - Study 1b for a more in-depth view of the child data including variance across the regression line). Maximum scores: isolated items = 7, item pairs = 8.

*Isolated items*

Model 1 showed an increase in recall of isolated items with age (*p* < .001) but no difference in recall between isolated words and isolated digits (*p* = .192). Adding the interaction term in Model 2 did not change the effect of stimulus type (*p* = .880), but age remained significant (*p* < .001) and there was no interaction between age and stimulus type (*p* = .908). Model 2 did not provide a significantly better fit to the data than Model 1 (*p* = .908).

*Item pairs*

Model 1 showed that recall of item pairs increased with age (*p* = .001) and was greater for digit pairs over word pairs (*p* = .005). When we added the interaction term in Model 2, the effect of stimulus type (*p* = .040) and age remained (*p* < .001). More importantly, the interaction term was significant (*p* = .008) and Model 2 provided a significantly better fit to the data than Model 1 (*p* = .008).

The children's mixed list data are largely consistent with the model. Isolated words were recalled marginally though non-significantly better than isolated digits, and digit pairs were recalled better than word pairs. More importantly, this latter effect interacted with age, with the developmental increase in digit pair recall being far greater than that of word pairs, while there was no interaction between age and stimulus type for isolated items. We note that, unlike the model, for which there was no effect of age on performance for isolated items, the empirical data does show increased recall of isolated items with age. We will return to this finding in the general discussion, noting here that this effect is not the source of performance differences across digit and word lists since both isolated digits and isolated words increase at exactly the same rate with age.

The effects seen in the model – and borne out in children – arise because of the greater presence in natural language of random sequences of digits compared to random sequences of words, and those digit sequences increase in number and diversity as the age at which the input is directed increases. That said, the model's performance was based on random samples of both 2-3 year old child-directed speech and adult-directed language in order to match the 4-6 year old child-directed speech samples. We therefore examine each of the full samples of data to show how the same digit superiority with age effect is seen irrespective of the speech sample.

## Study 2: Model span performance for three different language samples

### *Language samples used*

Each language sample was considered separately. For the 2-3 year old input, ten input files were created each containing 200,000 utterances randomly sampled from the full set of utterances directed at 2-3 year old children. For the 4-6 year old input, ten input files were created each containing the full set of 75,981 utterances directed at 4-6 year old children but with each file having the utterances randomly ordered. For the 2 million BNC, ten input files were created each containing 200,000 randomly sampled utterances directed at adults from the 2 million BNC. The model was run individually for each of the ten input files from each source (2-3, 4-6, BNC).

*Design, materials, and procedure*

The digit and word lists from Study 1a were used as stimuli. The independent variables was stimulus type (digits or words), with the dependent measure being the average number of chunks required to encode digit and word lists.

*Results*

Since each input file corresponds to a particular age group (2-3, 4-6, adults) we plot digit span and word span performance after all input is seen rather than developmentally. Figure 5 shows that regardless of which sample is used – and therefore regardless of the age of the person at which the speech is directed – there is a pattern of digit span exceeding word span (i.e., fewer chunks required to encode digit lists than word lists). Moreover, the *magnitude* of the increase is greater as the age at which the input is directed increases. Digit span is more superior to word span for the input aimed at 4-6 year old children than for the input aimed at 2-3 year old children, and in turn, digit span is more superior to word span for the input aimed at adults than for the input aimed at 4-6 year old children. Although we know of no study that has examined digit span and word span from two years of age to adulthood, these results are consistent with the 4 to 15 year old children of Gathercole, Pickering, Ambridge and Wearing (2004): digit recall exceeded word recall by 5.85 at the youngest age and 10.70 at the oldest age[6]. Notable also in Figure 5 is that word span increases with the age at which the input is directed, purely on the basis of an increasing number of instances of particular word sequences within the experimental stimuli. This is again consistent with the Gathercole et al. (2004) data. Put simply, as the age at which the input is directed increases, there are more instances of random sequences of words but - crucially - substantially more instances of random sequences of digits. This holds even when the input aimed at 2-3 year old children contains a total of 306,831 phonemic utterances whereas that aimed at 4-6 year old children contains only 75,981 phonemic utterances.

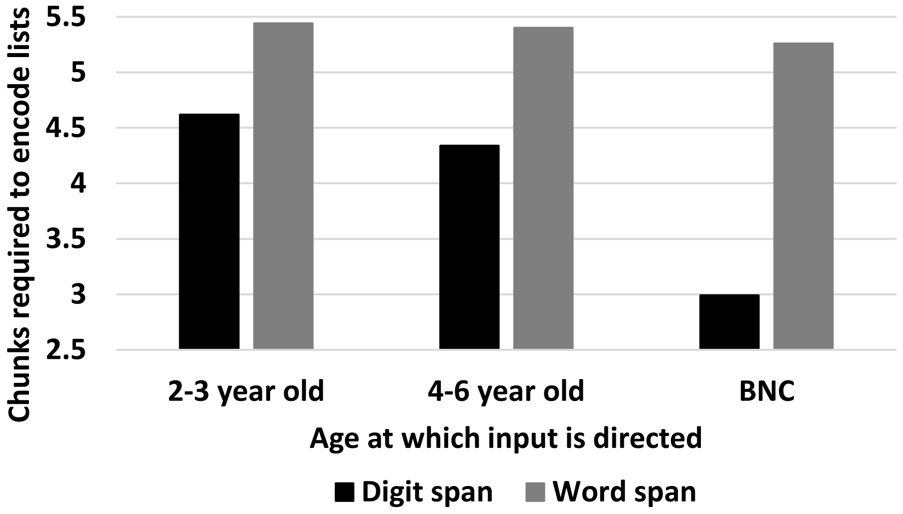

Figure 5. Number of chunks required to encode lists relating to digit span (DS) and word span (WS) based on language inputs aimed at 2-3 year old children (2-3), the 4-6 year old children (4-6), and adults (British National Corpus, BNC). Note that fewer chunks to encode lists equates to superior performance.

## Discussion

[6] Digit recall and word recall were assessed using the Working Memory Test Battery for Children (Pickering & Gathercole, 2001).

The capacity of STM, as a basic cognitive system or mechanism, is almost universally argued to increase with age – a developmental process that purportedly plays a determining role in a variety of developmental outcomes. Performance on measures of vSTM show individual differences and developmental increases to such an extent that performance on these measures have placed STM at the heart of explanations of age-related differences in performance and implicated vSTM and STM more generally across a broad range of cognitive functions. We have shown how associative learning is able to account for age-related performance increases in the archetypal measure of vSTM (digit span) and variants (word span, mixed span) together with the digit superiority with age effect whereby span size for digits increases with age at a greater rate than span size for other stimuli. Our explanation is simple: performance on vSTM measures such as span recall increases with age because exposure to language increases with age; the reason for digit superiority over other stimuli is that natural language has a much higher prevalence of random sequences of digits than random sequences of other stimuli. By extension, improvements in vSTM performance for any verbal material as a function of age may be parsimoniously accounted for in terms of increasing exposure to, and learning about, language, rather than by the joint outcome of increasing knowledge plus an increase in the capacity of an underlying STM mechanism. An important corollary to this conclusion is that claims based on contrasts between performance with ostensibly novel and patently familiar materials should be treated with caution. Not only are such distinctions always a matter of degree, the idea that any stimuli can be constructed for which there is not at least some accumulation over time of some potentially useful knowledge is doubtful. As such, the acquisition of knowledge, rather than increases in underlying capacity (cf. Cowan et al., 2015) provides a theoretically more promising mechanism giving rise to increases in STM performance.

Our theoretical position is not unique. Indeed, it bears great resemblance to Case's (1978) notion of automaticity whereby increased exposure to the same task or stimulus leads to processing efficiencies. The approach we present here represents an implementation of such a broad approach, whereby the processing efficiencies associated with automaticity arise due to the knowledge acquired via associative learning of the sequential structure of language. It also captures the graded nature of such automaticity/processing efficiency in terms of the degree of correspondence between the participant's knowledge and the precise content of the vSTM material. In this way, our approach is also commensurate with those usage-based approaches to language in general (e.g., Bybee, 2010; Ellis, 2002) and more formal specifications of such approaches (e.g., Christiansen & Chater, 2016; McCauley & Christiansen, 2019), where child performance in a range of productive linguistic settings is argued to be directly related to the extent to which the child has been exposed to similar settings that may serve as analogies for dealing with the current, relatively novel material. In our model, we interpret the creation of larger chunks arising from increased linguistic input as being a proxy for a multitude of performance gains that arise from increased exposure to language (e.g., perceptual, speech-motor). This position not only fits with the ideas above but also our model and child data show how such a position is able to capture not only developmental increases with age but also differences in those increases on the basis of the stimuli involved.

Related to the above theoretical viewpoints, we did find that recall of isolated items (both digits and words) increased with age. While the precise reason for the model's failure to capture the child data here remains moot, it seems plausible that it relates to the nature of lexical knowledge that is acquired by the model compared to children. In the model, the only effect of greater linguistic experience is the extent to which a chunked lexical item becomes *sequentially* associated with other units, whereas one could parsimoniously account for the effects seen by assuming, for example, that processing a chunk becomes more efficient as its frequency of encounter increases (e.g., the availability for encoding of chunks in the model's repertoire varies probabilistically as a function of frequency of occurrence, rather than, as is currently modeled, being either present in the model's knowledge, or not). For children, once a lexical item has been learned, increasing experience will also increase the richness of that lexical representation in itself. For example, richer semantic and episodic associations will be elaborated for that item of the type that will sustain better memory performance (see Jefferies & Lambon Ralph, 2006, for discussion of how lexico-semantic networks support STM). Indeed, increased experience with stimuli results in faster recognition and production, even for 5-10 year old children (Ojima, Matsuba-Kurita, Nakamura & Hagiwara, 2011), while for 7-12 year old children, naming accuracy steadily increases with age as the frequency of encounter of a word increases (Newman & German, 2002). Another possibility is that the increase in item recall with age is an outcome of the increase in sequence recall; since both words and digits here form closed sets, the successful recall of any part of a given sequence will increase the likelihood that another element will be correctly recalled by chance. Alternatively, the divergence here between model and data might reflect a degree of insensitivity in the model to picking up knowledge of specific digit-plus-word chunks (such as *two socks, five fingers*) that have been acquired by the child with respect to the actual stimuli used in these experiments and that may therefore give rise to improved child performance that, given our stimulus structure and scoring method, will appear as increases in item recall. It is important to remember here that despite the very large corpus on which the model is trained, it nonetheless represents a mere fragment of a developing child's encounter with language. For present purposes, the important point to note about this divergence is that the age-related increase in recall of isolated items in the child data cannot account for the digit superiority with age effect found with the digit and word lists in Study

1a. The Study 1b data show no interaction between (isolated) item type and age while in Study 1a, a performance advantage for digits over words that interacts with age is clearly present in both model and child performance.

From a practitioner's point of view, the usefulness of tests of vSTM and their value within assessment are not directly called into question by our argument; the extent to which they may serve as predictive or diagnostic measures derives, at least in part, from their statistical properties within such diagnostic and predictive settings. So, although our findings do not undermine the use of simple STM tasks as a means of predicting outcomes or identifying atypicalities, they do implicate a reappraisal of what such tasks measure, and therefore the sorts of possible underlying mechanisms that might give rise to those diagnostic or predictive properties. It has long been known that performance in vSTM tasks is affected by a variety of aspects of the long-term linguistic knowledge of the participant, but the implications of this association remain controversial. Traditional approaches interpret such effects as pointing to long-term memory and STM as distinct but interacting modes of processing, either in terms of distinct systems (Baddeley, 2012) or STM being the limited-capacity activated state of long-term knowledge (MacDonald & Christiansen, 2002). Since our account captures the empirical pattern without invoking either temporary storage or any other limited-capacity process, the implication is that while measures of vSTM performance are undoubtedly predictive of higher cognitive functions, their predictive power does not stem from them measuring the functioning of a primitive STM system upon which those higher functions are built. Rather, STM is that setting within which participants deploy their relevant knowledge that may be utilized to accomplish a given task. Rather than concluding that researchers should simply take into account the interaction between long-term knowledge and STM in order to qualify their estimates of STM per se, instead we would conclude that the theoretical approach be recast to place the long-term learning process itself as the focus of investigation.

This approach may be fruitfully applied to understanding cases of substantial individual differences in the linguistic performance across children of the same age. For example, children from low socioeconomic status (SES) families are widely reported to have poorer language outcomes than same-age children from high SES families (Fernald, Marchman & Weisleder, 2013; Huttenlocher, Waterfall, Vasilyeva, Vevea & Hedges, 2010). A robust finding is that language exposure for low SES children is impoverished relative to high SES children. Under an associative learning account, this would lead to poorer language outcomes and poorer performance on language-related tasks such as vSTM (Fernald, Marchman & Weisleder, 2013; Ottem, Lian & Karlsen, 2007; Parra, Hoff & Core, 2011). Generally, our account predicts differences in outcomes due to opportunities for learning from linguistic experience rather than any age-related maturational factors. As such, if the opportunities for learning are themselves held constant, language differences across children would chiefly be expected where the linguistic experience is different. This is the case even in children from low SES families (Weisleder & Fernald, 2013). Therefore, striking language differences across same-age children do not necessarily point to an underlying impairment, since deficiencies may derive instead from the quality of the language experience encountered by children.

Undoubtedly, individual differences will occur across children, even when language input is similar, but our account suggests that seeking explanation for this in individual differences in STM function will not be fruitful[7]. Instead, while the opportunity for learning is in part constrained by the linguistic environment of the child, it will also be constrained by

[7] Note that vSTM in atypical populations is usually assessed using digit span or nonword repetition i.e., tasks that we have shown can be accounted for by associative learning based on linguistic experience (see G. Jones & Macken, 2018).

those low-level processes that transfer information to any learning mechanism (e.g., auditory perception) and by the child’s ability to learn from their linguistic experience, while performance is also influenced by low-level output processes (e.g., speech/motor mechanisms). For the model, the elements are already coded in sequential form, since they are presented as phonetic transcriptions; the key learning that occurs in the model relates to the *sequences* within language; and performance is measured as the number of chunks required to represent a given input. However, individual differences may occur within the basic auditory processes that underpin sequential perception (Bregman, 1990; Warren, 1999) as well as the basic motor processes that control the execution of sequences of gestures (Lenneberg, 1967; Siegel *et al.*, 1982). Such perceptual and motor processing is associated with individual differences in a variety of language tasks in both impaired and unimpaired populations (e.g., Carello, LaVasseur & Schmidt, 2002; Krishnan, Alcock, Mercure, Leech, Barker, Karmiloff-Smith & Dick, 2013; Goswami, 2012). Indeed, many of the patterns of performance traditionally explained in terms of the operation of an underlying vSTM system in adults have been shown to be due to domain-general perceptual and motor functions, rather than STM processes, or even specifically verbal ones (D. M. Jones, Macken & Nicholls, 2004; Macken, Taylor & D. M. Jones, 2014; 2015). Research has also shown that learning ability differs across typical and atypical children (Hsu & Bishop, 2014) while more generally there is large variability in learning associations amongst stimuli (Frost, Armstrong, Siegelman & Christiansen, 2015). In this sense, individual differences in low-level input/output processes and individual differences in one’s ability to learn will all result in performance differences in language-related tasks such as vSTM.

Our results speak to the numerous studies that have drawn implications based on digit span and similar measures of vSTM. We have shown that such measures are inextricably linked to the linguistic experience of the rememberer. If it were possible to assume that linguistic experience is broadly similar *within* the same age group, this would not be a problem. However, many factors affect the linguistic exposure of individuals at any given age. Even more problematic are those studies that draw conclusions on the basis of vSTM performance *between* ages where greater differences in linguistic experience are expected. To add further concern, language differences follow through into differences in reading ability (Lervåg & Aukrust, 2010), which in turn, may provide richer language experience than natural speech (Montag, M. N. Jones & Smith, 2015). Caution must therefore be applied in evaluating claims about development based on vSTM measures.

All told, then, while the predictive power of tests of vSTM and their value within assessment are not in doubt, our findings suggest that how such effects are understood needs to be radically re-evaluated. We have shown that performance in such tasks can be modeled in terms of the linguistic experience of the child and the sequential learning about language that such experience affords, without any need for bespoke STM processes. Not only does this mean that the pivotal causal role in development which STM has been awarded should be reviewed, it also means that the quest for those causal mechanisms underlying the development of higher cognitive function (such as language) needs to be directed instead to basic associative learning processes operating within perception and production.

## References

Archibald, L. M. D., & Gathercole, S. E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research, 49*, 970-983.

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

BNC (2007). The British National Corpus, version 3 (BNC XML Edition) [on-line database]. http://www.natcorp.ox.ac.uk/.

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 60,* P223-P233.

Bregman, A. (1990). *Auditory scene analysis*. Cambridge, MA.: MIT Press.

Bybee, J. (2010). *Language, usage and cognition.* Cambridge: Cambridge University Press

Carello, C., LeVasseur, V. M., & Schmidt, R. C. (2002). Movement sequencing and phonological fluency in (putatively) nonimpaired readers. *Psychological Science, 13,* 375-379.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39,* 1-72.

Christensen, R. H. B. (2019). Ordinal-regression models for ordinal data. https://CRAN.R-project.org/package=ordinal

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169,* 323-338.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Dempster, F. N. (1981). Memory span : Sources of individual and developmental differences. *Psychological Bulletin, 89,* 63-100.

Elliot, C. D., & Smith, P. (2011). *British Abilities Scales (3rd ed.).* London, UK: GL assessment.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211-245.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234-248.

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

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 54*, 1–30.

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

Goswami, U. (2012). Language, music, and children's brains: a rhythmic timing perspective on language and music as cognitive systems. In P. Rebuschat, M. Rohrmeier, J. Hawkins and I. Cross (Eds.). *Language and music as cognitive systems.* Oxford, UK: Oxford University Press.

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language*, *67*, 106-148.

Hale, S., Bronik, M. D., & Fry, A. F. (1997). Verbal and spatial working memory in school-age children: Developmental differences in susceptibility to interference. *Developmental Psychology*, *33*, 364-371.

Helland, T., & Asbjørnsen, A. (2004). Digit span in dyslexia: Variations according to language comprehension and mathematics skills. *Journal of Clinical and Experimental Neuropsychology, 26,* 31-42.

Hinton, V. J., De Vivo, D. C., Nereo, N. E., Goldstein, E., & Stern, Y. (2001). Selective deficits in verbal working memory associated with a known genetic etiology: The neuropsychological profile of Duchenne muscular dystrophy. *Journal of the International Neuropsychological Society, 7,* 45-54.

Hornung, C., Brunner, M., Reuter, R. A. P., & Martin, R. (2011). Children's working memory: Its structure and relationship to fluid intelligence. *Intelligence, 39,* 210-221.

Hsu, H. J., & Bishop, D. V. M. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science, 17,* 352–365.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*, 343-365.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140*, 339-373.

Jefferies E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain, 129,* 2132-2147.

Jones, D. M., Macken, W. J., & Nicholls, A. P. (2004). The phonological store of working memory: Is it phonological and is it a store? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 656-674

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition, 153*, 79-88.

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

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

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory and Cognition, 46,* 216-229.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology, 98,* 1-21.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children (2nd ed.).* Circle Pines, MN: American Guidance Service.

Krishnan, S., Alcock, K., Mercure, E., Leech, R., Barker, E, Karmiloff-Smith, A., & Dick, F. (2013). Articulating novel words: Children's oromotor skills predict nonword repetition abilities. *Journal of Speech, Language and Hearing Research, 56*, 1800-1812.

Lenneberg, E. H. (1967). *Biological foundations of language.* New York: John Wiley & Sons.

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 51,* 612-620.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109,* 35-54.

Macken, W. J., & Jones, D. M. (2003). Reification of phonological storage. *Quarterly Journal of Experimental Psychology Section A, 56*, 1279-1288.

Macken, B., Taylor, J. C., & Jones, D. M. (2014). Language and short-term memory: The role of perceptual-motor affordance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1257-1270.

Macken, B., Taylor, J., & Jones, D. M. (2015). Limitless capacity: A dynamic object-oriented approach to short-term memory. *Frontiers in Psychology, 6,* 293.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children’s printed word database: Continuities and changes over time in children’s early reading vocabulary. *British Journal of Psychology, 101*, 221-242.

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.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*, 1489-1496.

Nettelbeck, T., & Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Personality and Individual Differences*, *48*, 379-384.

Newman, R. S., & German, D. J. (2002). Effects of lexical factors on lexical access among typical language-learning children and children with word-finding difficulties. *Language and Speech, 45*, 285–317.

Ojima, S., Matsuba-Kurita, H., Nakamura, N., & Hagiwara, H. (2011). The acceleration of spoken-word processing in children’s native-language acquisition: An ERP cohort study. *Neuropsychologia, 49*, 790–799.

Ottem, E. J., Lian, A., & Karlsen, P. J. (2007). Reasons for the growth of traditional memory span across age. *European Journal of Cognitive Psychology, 19*, 233-270.

Parra, M., Hoff, E., & Core, C. (2011). Relations among language exposure, phonological memory, and language development in Spanish-English bilingually developing 2-year-olds. *Journal of Experimental Child Psychology*, *108,* 113-125.

Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 29-33.

Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: Separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology*, *56*, 431–442.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development*, *83,* 1762-1774.

Rowland, C. (2014). *Understanding child language acquisition*. Abingdon: Routledge.

R Core Team (2018). *R: A language and environment for statistical computing. R foundation for statistical computing*. Vienna, Austria. https://www.R-project.org/

Salthouse, T. A., Mitchell, D. R., Skovronek, E., & Babcock, R. L. (1989). Effects of adult age and working memory on reasoning and spatial abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15,* 507-516.

Siegel, L. S., Saigal, S., Rosenbaum, P., Morton, R. A., Young, A., Berenbaum, S., & Stoskopf, B. (1982). Predictors of development in preterm and full-term infants: A model for detecting the at risk child. *Journal of Pediatric Psychology*, *7*, 135-148.

Stipek, D., & Valentino, R. A. (2015). Early childhood memory and attention as predictors of academic growth trajectories. *Journal of Educational Psychology, 107*, 771-788.

Stuart, G., & Hulme, C. (2000). The effects of word co-occurrence on short-term memory: Associative links in long-term memory affect short-term memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 796–802.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*, 454-464.

Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of Verb-Argument structure: An alternative account. *Journal of Child Language*, *28*, 127-152.

Turner, J. E., Henry, L. A, Smith, P. T., & Brown, P. A. (2004). Redintegration and lexicality effects in children: Do they depend upon the demands of the memory task? *Memory & Cognition*, *32*, 501-510.

Warren, R. M. (1999). *Auditory perception: A new analysis and synthesis*. Cambridge, UK: Cambridge University Press.

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children - Fourth UK Edition (WISC-IV UK).* San Antonio, TX: Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale (4th ed.).* San Antonio, TX: Pearson.

Wechsler, D. (2009). *Wechsler Memory Scale (4th ed.).* San Antonio, TX: Pearson.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143-2152.

## Appendix

*Analyses - Study 1a*

For each dependent variable, we ran two ordered categorical logistic mixed-effects models. Model 1 used stimulus type (digits or words) and age (in months) as fixed effects, while Model 2 additionally included the interaction between stimulus type and age as a fixed effect. A maximal random effects structure was included for participant.

Span size

Table A1 shows the analyses relating to span size and Figure A1 shows the plotted data. Model 1 showed that span increased with age ($p$ = .001) and was greater for digit lists over word lists ($p$ < .001). The effect of stimuli type disappeared in Model 2 ($p$ = .461), the effect of age remained ($p$ = .026), and the interaction was significant ($p$ = .039). Table A2 illustrates the model comparisons, showing that Model 2 provided a significantly better fit to the data than Model 1 ($p$ = .036).

Table A1. Estimates, 95% CIs and *p* values for Model 1 and Model 2 with span size as the dependent variable.

| | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* | *Odds Ratios* | *CI* | *p* |
| (Intercept: 2\|3) | .22 | .01 – 4.08 | .309 | .08 | .00 – 1.72 | .106 |
| (Intercept: 3\|4) | 1.18 | .08 – 16.76 | .901 | .41 | .02 – 7.12 | .544 |
| (Intercept: 4\|5) | 34.98 | 2.47 – 495.76 | **.009** | 12.04 | .69 – 208.63 | .087 |
| (Intercept: 5\|6) | 1153.00 | 69.52 – 19121.43 | **<.001** | 428.59 | 20.94 – 8771.43 | **<.001** |
| (Intercept: 6\|7) | 7362.41 | 391.45 – 138472.33 | **<.001** | 2962.08 | 124.01 – 70750.04 | **<.001** |
| (Intercept: 7\|8) | 159386.83 | 5711.90 – 4447585.88 | **<.001** | 71488.05 | 1911.85 – 2673088.72 | **<.001** |

| | | | | | | |
|---|---|---|---|---|---|---|
| (Intercept: 8\|9) | 2200398.17 | 29640.44 – 163349530.21 | **<.001** | 1200426.14 | 10901.45 – 132186325.76 | **<.001** |
| Stimulus Type | 9.90 | 5.17 – 18.97 | **<.001** | .28 | .01 – 8.45 | .461 |
| Age | 1.04 | 1.02 – 1.07 | **.001** | 1.03 | 1.00 – 1.06 | **.026** |
| Stimulus Type * Age | | | | 1.04 | 1.00 – 1.07 | **.039** |
| Observations | 286 | | | 286 | | |

Table A2. Comparisons between Model 1 and Model 2 for the span size data, total span data, isolated item data, and item pair data. Df = degrees of freedom; logLik = log likelihood; Chisq = chi square.

| | Model comparisons - Span size | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Df | AIC | BIC | logLik | deviance | Chisq | Chi Df | Pr(>Chisq) |
| Model 1 | 12 | 758 | - | -367 | - | – | – | – |
| Model 2 | 13 | 756 | - | -365 | - | 4.42 | 1 | **.036** |
| | Model comparisons - Total span | | | | | | | |
| Model 1 | 16 | 1073 | - | -520 | - | – | – | – |
| Model 2 | 17 | 1071 | - | -518 | - | 4.08 | 1 | **.043** |
| | Model comparisons - Isolated item data | | | | | | | |
| Model 1 | 6 | 1168 | 1197 | -578 | 1156 | – | – | – |
| Model 2 | 7 | 1170 | 1204 | -578 | 1156 | .01 | 1 | .908 |
| | Model comparisons - Item pair data | | | | | | | |
| Model 1 | 6 | 1371 | 1401 | -679 | 1359 | – | – | – |
| Model 2 | 7 | 1365 | 1401 | -676 | 1351 | 7.07 | 1 | **.008** |

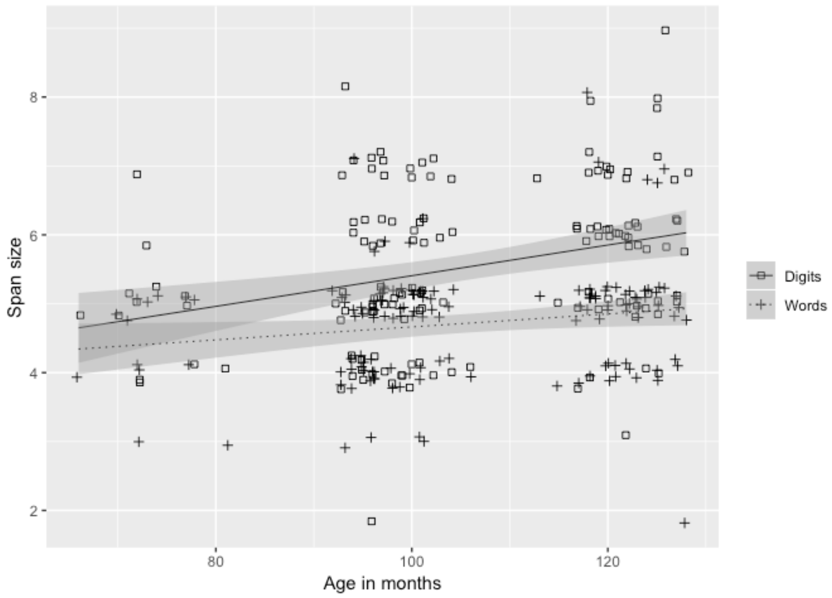

Figure A1. Span size by age (N = 143) for digit and word lists (25% of random noise was added to each y axis tick value to avoid datapoints sitting on top of each other). A regression line with 95% confidence interval was added.

Total span

Table A3 shows the analyses relating to total span and Figure A2 shows the plotted data. Model 1 showed that span increased with age (*p* = .001) and was greater for digit lists over word lists (*p* < .001). Adding the interaction term in Model 2 meant the effect of stimulus type disappeared (*p* = .694) but age was significant (*p* = .005) together with the interaction (*p* = .045). Table A2 illustrates the model comparisons, showing that Model 2 provided a significantly better fit to the data than Model 1 (*p* = .043).

Table A3. Estimates, 95% CIs and *p* values for Model 1 and Model 2 with total span as the dependent variable.

| | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* | *Odds Ratios* | *CI* | *p* |
| (Intercept: 2\|3) | .31 | .01 – 7.96 | .479 | .13 | .00 – 3.87 | .241 |
| (Intercept: 3\|4) | .48 | .02 – 11.07 | .647 | .21 | .01 – 5.43 | .345 |
| (Intercept: 4\|5) | 4.34 | .22 – 84.65 | .333 | 1.81 | .08 – 40.45 | .707 |

| | | | | | | |
|---|---|---|---|---|---|---|
| (Intercept: 5\|6) | 21.97 | 1.12 – 432.01 | **.042** | 8.94 | .40 – 200.81 | .168 |
| (Intercept: 6\|7) | 130.76 | 6.34 – 2695.82 | **.002** | 53.23 | 2.27 – 1248.67 | **.014** |
| (Intercept: 7\|8) | 762.12 | 34.52 – 16825.94 | **<.001** | 320.24 | 12.78 – 8021.63 | **<.001** |
| (Intercept: 8\|9) | 6820.58 | 282.12 – 164892.92 | **<.001** | 3077.17 | 111.06 – 85259.94 | **<.001** |
| (Intercept: 9\|10) | 22307.31 | 866.84 – 574054.91 | **<.001** | 10528.03 | 352.26 – 314649.33 | **<.001** |
| (Intercept: 10\|11) | 179619.45 | 5951.78 – 5420752.99 | **<.001** | 89117.55 | 2462.09 – 3225695.36 | **<.001** |
| (Intercept: 11\|12) | 526822.81 | 15440.70 – 17974722.76 | **<.001** | 269840.95 | 6494.33 – 11211955.49 | **<.001** |
| (Intercept: 12\|13) | 1694550.83 | 39722.43 – 72289199.94 | **<.001** | 932764.14 | 17563.04 – 49538641.91 | **<.001** |
| Stimulus Type | 11.56 | 6.47 – 20.65 | **<.001** | .54 | .03 – 11.28 | .694 |
| Age | 1.05 | 1.02 – 1.08 | **.001** | 1.04 | 1.01 – 1.07 | **.005** |
| Stimulus Type * Age | | | | 1.03 | 1.00 – 1.06 | **.045** |
| Observations | 286 | | | 286 | | |

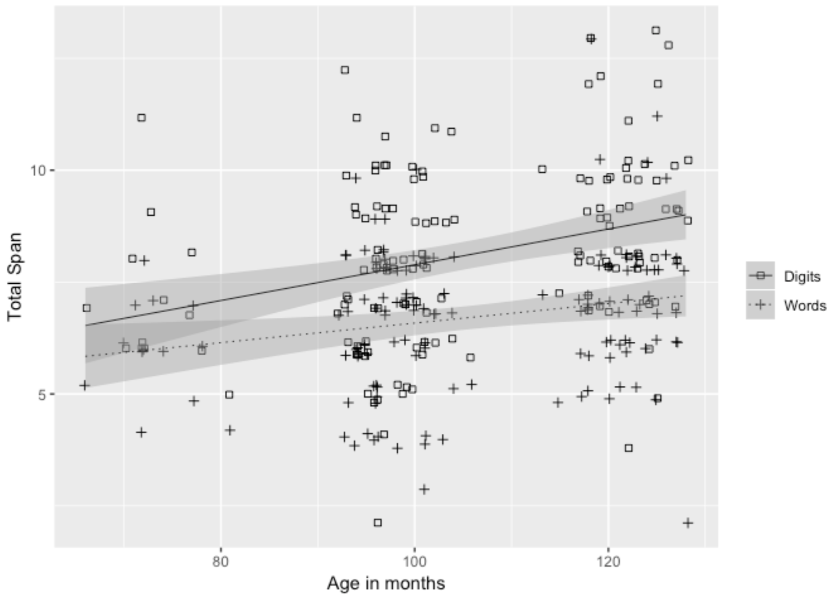

Figure A2. Total span by age (N = 143) for digit and word lists (25% of random noise was added to each y axis tick value to avoid datapoints sitting on top of each other). A regression line with 95% confidence interval was added.

*Analyses - Study 1b*

Figure A3 shows the mixed span performance. We ran an ordered categorical logistic model on the mixed span data using age as the independent variable (see Table A4). There was a clear effect of age (*p* = .001) showing that as age increased, the number of mixed lists recalled also increased.

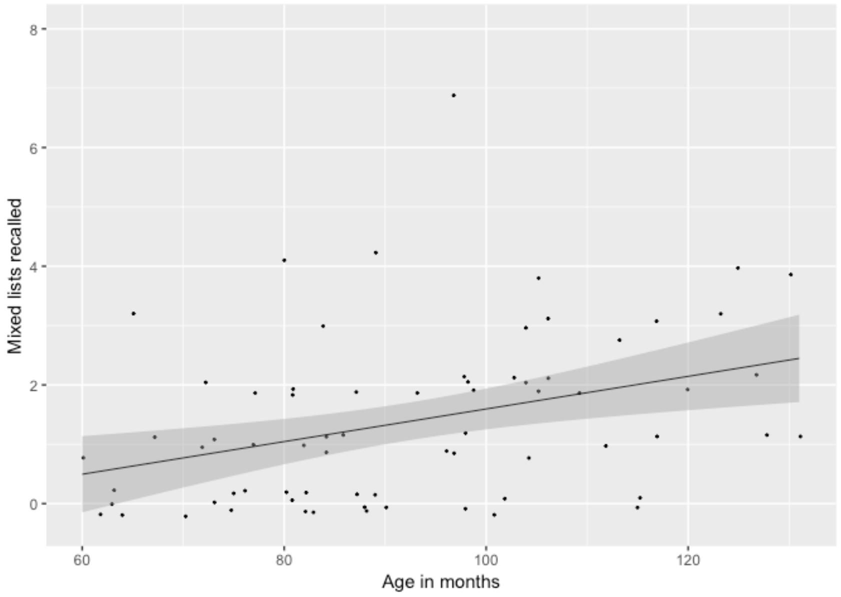

Figure A3. Number of mixed lists accurately recalled (maximum = 8) by age (N = 70). 25% of random noise was added to each y axis tick value to avoid datapoints sitting on top of each other. A regression line with 95% confidence interval was added.

Table A4. Estimates, 95% CIs and *p* values with number of mixed lists accurately recalled as the dependent variable.

| **Mixed lists recalled** | | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept: 0\|1) | 22.04 | 2.25 – 216.33 | **.008** |
| (Intercept: 1\|2) | 68.70 | 6.31 – 748.02 | **.001** |
| (Intercept: 2\|3) | 242.49 | 19.48 – 3018.22 | **<.001** |
| (Intercept: 3\|4) | 631.69 | 44.88 – 8892.11 | **<.001** |
| (Intercept: 4\|7) | 4284.12 | 169.87 – 108044.32 | **<.001** |
| Age | 1.04 | 1.02 – 1.07 | **.001** |
| Observations | 70 | | |
| Cox & Snell's $R^2$ / Nagelkerke's $R^2$ | 0.153 / 0.161 | | |

We then ran binomial logistic mixed effects models separately for isolated items and item pairs using the same analyses technique as Study 1a and the same fixed and random effects that were used in Models 1 and 2. However, the dependent variable in this case was whether each isolated item or item pair was recalled correctly (0 = incorrect, 1 = correct). Since the mixed lists were intentionally challenging for the children, we considered an isolated item or item pair to be correctly recalled if it occurred in any serial position (i.e., free recall rather than serial recall; there is little difference between the two, Grenfell-Essam & Ward, 2012).

Isolated items

Table A5 shows the analyses relating to isolated items and Figure A4 shows the plotted data. Model 1 showed an increase in accurate recall of isolated items with age (*p* < .001) but no difference in recall accuracy for isolated words over isolated digits (*p* = .192). Adding the interaction term in Model 2 did not change the effect of stimulus type (*p* = .880), however age remained significant (*p* < .001). Crucially, there was no interaction between age and stimulus type (*p* = .908) and Model 2 did not provide a significantly better fit to the data than Model 1 (*p* = .908, see Table A2).

Table A5. Odds ratios, 95% CIs and *p* values for Model 1 and Model 2 for isolated items.

| | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | .20 | .07 – .61 | **.005** | .21 | .06 – .71 | **.012** |
| Stimulus Type | 1.24 | .90 – 1.71 | .192 | 1.13 | .23 – 5.58 | .880 |
| Age | 1.03 | 1.01 – 1.04 | **<.001** | 1.03 | 1.01 – 1.04 | **<.001** |
| Stimulus Type * Age | | | | 1.00 | .98 – 1.02 | .908 |
| **Random Effects** | | | | | | |
| $\sigma^2$ | 3.29 | | | 3.29 | | |
| $\tau_{00}$ | .31 $_{\text{Id}}$ | | | .30 $_{\text{Id}}$ | | |
| $\tau_{11}$ | .23 $_{\text{Id, Stimulus Type}}$ | | | .23 $_{\text{Id, Stimulus Type}}$ | | |
| $\rho_{01}$ | .54 $_{\text{Id}}$ | | | .54 $_{\text{Id}}$ | | |
| ICC | .08 $_{\text{Id}}$ | | | .08 $_{\text{Id}}$ | | |
| Observations | 980 | | | 980 | | |
| Marginal $R^2$ / Conditional $R^2$ | .06 / .20 | | | .06 / .20 | | |

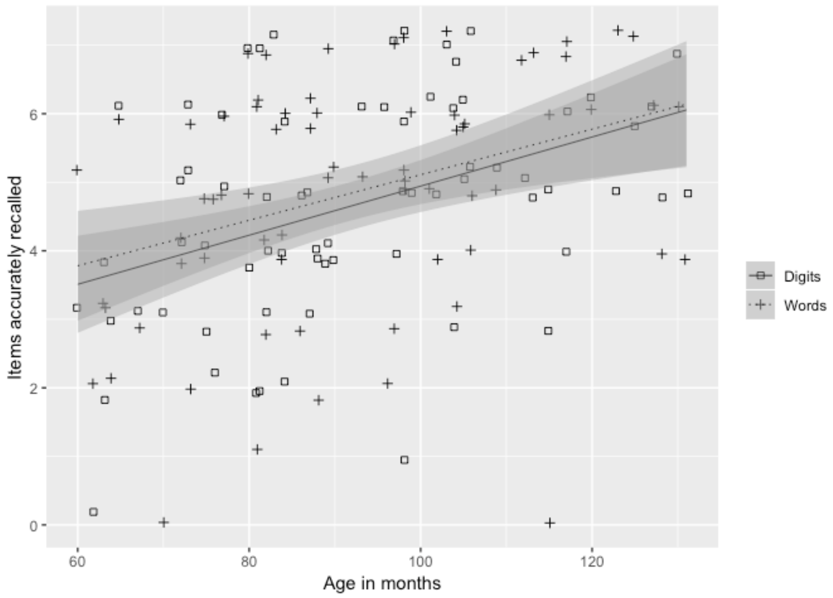

Figure A4. Recall of isolated items (maximum score = 7) by age (N = 70) for mixed lists. 25% of random noise was added to each y axis tick value to avoid datapoints sitting on top of each other. A regression line with 95% confidence interval was added.

Item pairs

Table A6 shows the analyses relating to item pairs and Figure A5 shows the plotted data. Model 1 showed that recall increased with age (*p* = .001) and was greater for digit pairs over word pairs (*p* = .005). When we added the interaction term in Model 2, the effects of stimulus type (p = .040) and age remained (*p* <. 001). Crucially, the interaction term (*p* = .008) was significant and Model 2 provided a significantly better fit to the data than Model 1 (*p* = .008, see Table A2).

Table A6. Odds ratios, 95% CIs and *p* values for Model 1 and Model 2 for item pairs.

| | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | .09 | .03 – .26 | **<.001** | .03 | .01 – .12 | **<.001** |
| Stimulus Type | .67 | .51 – .89 | **.005** | 4.44 | 1.07 – 18.37 | **.040** |
| Age | 1.02 | 1.01 – 1.03 | **.001** | 1.03 | 1.02 – 1.05 | **<.001** |
| Stimulus Type * Age | | | | .98 | .97 – .99 | **.008** |

| Random Effects | | |
|---|---|---|
| $\sigma^2$ | 3.29 | 3.29 |
| $\tau_{00}$ | .45 $_{\text{Id}}$ | .50 $_{\text{Id}}$ |
| $\tau_{11}$ | .00 $_{\text{Id, Stimulus Type}}$ | .01 $_{\text{Id, Stimulus Type}}$ |
| $\rho_{01}$ | -1.00 $_{\text{Id}}$ | -1.00 $_{\text{Id}}$ |
| ICC | .12 $_{\text{Id}}$ | .13 $_{\text{Id}}$ |
| Observations | 1120 | 1120 |
| Marginal $R^2$ / Conditional $R^2$ | .05 / .15 | .05 / .16 |

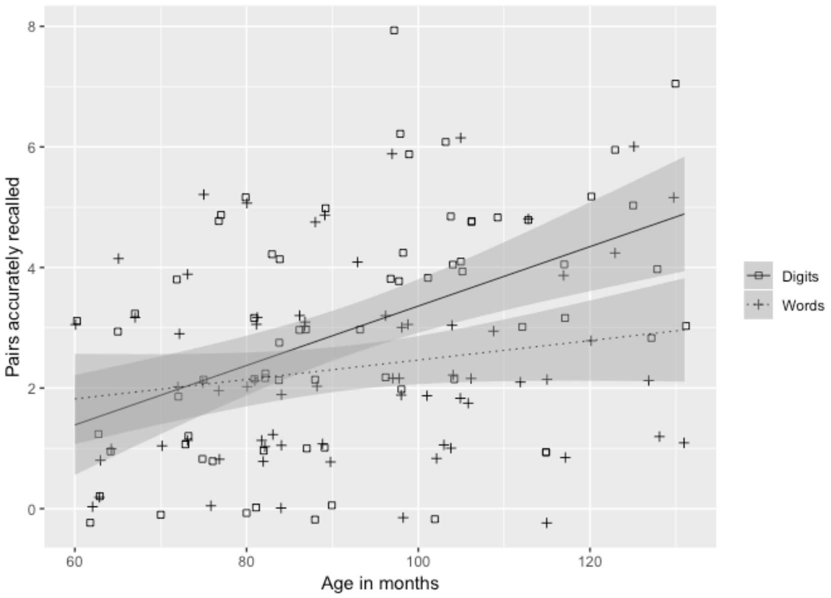

Figure A5. Recall of item pairs (maximum score = 8) by age (N = 70) for mixed lists. 25% of random noise was added to each y axis tick value to avoid datapoints sitting on top of each other. A regression line with 95% confidence interval was added.