LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Simulating the Relationship Between Nonword Repetition Performance and Vocabulary Growth in 2-Year-Olds: Evidence From the Language 0–5 Project

Caroline F. Rowland [a,b,c], Amy Bidgood [d], Gary Jones [e], Andrew Jessop,[b] Paula Stinson,[b] Julian M. Pine,[b] Samantha Durrant,[f] and Michelle S. Peter[g]

CRediT author statement: **Caroline F. Rowland**: conceptualization; methodology; software; formal analysis; resources; data curation; writing – original draft; writing – review and editing; supervision; project administration; funding acquisition. **Amy Bidgood**: conceptualization; methodology; software; formal analysis; investigation; resources; data curation; writing – review and editing. **Gary Jones**: software; formal analysis; investigation; resources; writing – review and editing; visualization. **Paula Stinson**: conceptualization; methodology; investigation; resources; data curation. **Julian M. Pine**: conceptualization; methodology; writing – review and editing; funding acquisition. **Samantha Durrant**: conceptualization; methodology; formal analysis; investigation; resources; data curation; writing – review and editing. **Michelle S. Peter**: conceptualization; methodology; formal analysis; investigation; resources; data curation; writing – review and editing. **Andrew Jessop**: software; formal analysis; investigation; resources; data curation; writing – original draft; writing – review and editing; visualization.

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org

We would like to thank all of the families who participated in the Language 0–5 Project. This work was supported by the ESRC International Centre for Language and Communicative Development (LuCiD), funded by the UK Economic and Social Research Council (ES/L008955/1). All supplementary materials are available in an Open Science Framework repository (https://osf.io/9s4d7).

Correspondence concerning this article should be addressed to Caroline Rowland, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands. Email: Caroline.Rowland@mpi.nl

The handling editor for this article was Kristopher Kyle.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Language Learning* published by Wiley Periodicals LLC on behalf of Language Learning Research Club, University of Michigan.
DOI: 10.1111/lang.12671

[a]Max Planck Institute for Psycholinguistics [b]University of Liverpool [c]Donders Institute for Brain, Cognition and Behaviour [d]Liverpool John Moores University [e]Nottingham Trent University [f]University of Manchester [g]Great Ormond Street Hospital for Children NHS Foundation Trust

Abstract: A strong predictor of children's language is performance on non-word repetition (NWR) tasks. However, the basis of this relationship remains unknown. Some suggest that NWR tasks measure phonological working memory, which then affects language growth. Others argue that children's knowledge of language/language experience affects NWR performance. A complicating factor is that most studies focus on school-aged children, who have already mastered key language skills. Here, we present a new NWR task for English-learning 2-year-olds, use it to assess the effect of NWR performance on concurrent and later vocabulary development, and compare the children's performance with that of an experience-based computational model (CLASSIC). The new NWR task produced reliable results; replicating wordlikeness effects, word-length effects, and the relationship with concurrent and later language ability we see in older children. The model also simulated all effects, suggesting that the relationship between vocabulary and NWR performance can be explained by language experience-/knowledge-based theories.

Keywords vocabulary development; non-word repetition; syntax development; phonological working memory; computational modelling; CLASSIC

Introduction

A strong predictor of children's language ability is performance on the non-word repetition (NWR) task, in which children are asked to repeat nonwords of varying lengths (e.g., *pomaguv*). Despite its simplicity, performance on this task robustly correlates with language knowledge. The relationship holds across the whole of middle childhood (Adams & Gathercole, 1995, 2000; Gathercole et al., 1992; Gathercole & Adams, 1993; Gathercole & Baddeley, 1989; Roy & Chiat, 2004), in multiple languages (see Coady & Evans, 2008, for a review), and in children with and without language disorders (see Graf Estes et al., 2007, for a meta-analysis). It also holds across different language measures, such that NWR performance both correlates with vocabulary and predicts children's ability to learn novel words in the lab (Bowey, 2001; Chiat & Roy, 2008; Gathercole & Baddeley, 1989, 1990). The relationship cannot be explained by variables such as age, reading ability, or nonverbal intelligence, or by appealing to children's articulation skills (Adams & Gathercole, 1995, 2000; Baddeley, 1986a; Gathercole et al., 1992, 1999).

When Gathercole and Baddeley first assessed the relationship between NWR performance and language, they assumed that NWR tasks measured the

capacity of phonological working memory (Baddeley, 1986b, 2000; Baddeley & Hitch, 1974; Gathercole, 2006; Gathercole & Baddeley, 1989). However, we now know that successful performance calls on a number of different skills (Coady & Evans, 2008; Snowling et al., 1991). Most pertinently, although the task uses nonwords to minimize the effect of known words (Gathercole et al., 1992), there is still a strong influence of prior linguistic knowledge on performance. For example, there are strong wordlikeness effects, such that nonwords that contain phoneme sequences that occur in the child's native language are repeated more accurately (Dollaghan et al., 1993, 1995; Edwards et al., 2004; Gathercole, 1995; Gathercole et al., 1991; Keren-Portnoy et al., 2010; Munson et al., 2005; Szewczyk et al., 2018), and there is evidence that performance correlates with increases in linguistic knowledge in monolingual and bilingual children (Messer et al., 2015). Because of this, some recommend that NWR tasks be "used as a method to assess the structural organisation of the phonological lexicon," rather than as a measure of phonological working memory capacity (Coady & Evans, 2008, p. 3), and some modeling work provides evidence supporting this suggestion (see below).

One influential idea is that differences in NWR performance reflect not just linguistic knowledge in general but, more specifically, differences in children's knowledge of sublexical representations (representations corresponding to parts of words, e.g., the phoneme [d] or the phoneme sequence [dɒ]). For example, Szewczyk et al. (2018) investigated a range of potential predictors including length, phonotactic probability, lexical neighborhood, and phonological complexity. Performance across the 150 different nonwords tested was best explained by the amount of support that a nonword received from sublexical representations of all grain sizes, which led the authors to propose a new index of sublexical support (average phonemic ngram frequency) to measure these representations. Additional evidence comes from a series of simulations by Jones and colleagues (see, e.g., Jones, 2016; Jones et al., 2007; Jones & Macken, 2018), who demonstrated that an experience-based model with a fixed processing limit that learns both lexical and sublexical representations can model developmental changes in children's NWR performance as well as the relationship with vocabulary.

The assumption behind Jones and colleagues's model (originally EPAM-VOC, now CLASSIC) is that, although there is a limit on the amount of information that we can process at any one time (three to five chunks or meaningful items in young adults; Cowan, 2010), this limit does not vary across individuals or development.[1] Rather, the driver of individual and developmental differences is the amount of linguistic material already stored in the child's

lexicon. Experience with language leads children to build both lexical (word-level) and sublexical chunks of linguistic information in long-term memory. Children who hear more input build more lexical chunks (i.e., have bigger vocabularies) and, at the same time, build more and longer chunks of sublexical knowledge (which they can use to solve NWR tasks). In other words, different amounts of linguistic experience explain individual differences and developmental increases in NWR performance, and also explain why there are strong correlations between NWR performance and vocabulary.

The CLASSIC model predicts, in particular, that we should find strong associations between NWR performance and vocabulary development in 2-year-olds (i.e., children who have already learned to segment speech into words and are now rapidly expanding their vocabulary), since it suggests that, with exposure to language, a larger repertoire of both sublexical chunks (used in NWR tasks) and lexical chunks (vocabulary) is acquired. This means that high scores on NWR tasks should be related to larger vocabularies. However, although there is substantial evidence for a link between NWR performance and vocabulary, the majority of both experimental and modeling work on this topic has been with school-aged children, who have already mastered many key language skills. There are only a handful of studies with preschool-aged children, very few of which assess development over time (e.g., Chiat & Roy, 2008; Gathercole & Adams, 1993; Hoff et al., 2008; Jones et al., 2007; Newbury et al., 2016; Roy & Chiat, 2004; Stokes et al., 2013; Stokes & Klee, 2009; Torrington Eaton et al., 2015; Verhagen et al., 2019). Given the substantial cognitive and linguistic advances made during the first years of life, we cannot assume that effects found in school-aged children will be replicated at younger ages. In fact, there is some evidence that, although the direction of the correlation remains the same, the direction of the causal relationship changes with age; for example, that NWR ability plays a causal role in vocabulary growth before, but not after, age 5;0, perhaps because it is only in younger children that the constraint measured by NWR performance is limited enough to affect everyday speech processing (Cheung, 1996; Gathercole et al., 1992; Gathercole & Adams, 1993). This means that it is not possible to draw conclusions about the size or direction of the causal relationship between NWR performance and language development in preschoolers from work on school-aged children.

The first goal of the present study was to create a new NWR task specifically for 2-year-olds and to use this to assess the effect of NWR performance on concurrent and later vocabulary development. The second goal was to then compare the performance of the children with that of an experience-based model (CLASSIC), to determine if the model can simulate NWR

performance and the relationship with vocabulary in the preschool years. Our overarching aim was to test whether the relationship between language and NWR performance in English speaking 2-year-olds can be explained by language-experience-based theories.

Testing the Relationship Between Nonword Repetition Performance and Language Growth in 2-Year-Olds with a New Age-Appropriate Nonword Repetition Task

Studies of NWR performance and its relation with vocabulary in preschool children are scarce and sometimes contradictory. For example, of the three studies that measure whether 2-year-olds' NWR performance predicts later language growth, one reported a clear relationship over time (Chiat & Roy, 2008), one reported no evidence of NWR performance predicting unique variance (Newbury et al., 2016), and one reported a reciprocal relationship, but with stronger predictions from early vocabulary to later NWR performance than vice versa (Verhagen et al., 2019). One of the reasons for the scarcity of studies is the difficulty of administering experimental tasks to children this young. For example, it is difficult to encourage 2-year-olds to repeat nonwords in a laboratory situation, which means that, even if the number of children recruited to a study is initially large, the number of children who contribute to the final analysis can be small (see, e.g., Gathercole & Adams, 1993). Then, when young children do try to repeat the nonword, it is difficult to determine whether errors simply reflect the articulation problems that are common at this age (e.g., producing *chor* as *tor*; Keren-Portnoy et al., 2010; Krishnan, 2017). In fact, there is some evidence that substantially smaller correlations between 2-year olds' NWR performance and vocabulary are observed when the effects of articulatory difficulty are controlled for (e.g., when clusters are removed to avoid cluster simplification; Torrington Eaton et al., 2015).

To address these issues, we created a new NWR task designed to be suitable for children as young as 24 months. We evaluated the robustness of the new task by investigating whether the results were reliable over time and whether we could replicate two classic effects from the literature: the effect of wordlikeness (that wordlike nonwords will be repeated correctly more easily than nonword-like nonwords) and that of word length (that nonwords with fewer syllables will be repeated correctly more easily than longer nonwords). We then assessed the relationship between NWR performance and vocabulary not just concurrently but also longitudinally, using data from parent report checklists (Communicative Development Inventories [CDIs]) of children's vocabulary completed at multiple time-points between 27 and 37 months and from a standardized test

of receptive vocabulary (the British Picture Vocabulary Scale, Third Edition [BPVS-3]) administered at 36 and 42 months.[2] If NWR tasks are capturing individual differences that have an effect on language acquisition, we expect that NWR performance will not only correlate with concurrent linguistic knowledge, but also predict subsequent language growth in 2-year-olds.

Simulating the Relationship Between Nonword Repetition Performance and Vocabulary Growth in 2-Year-Olds

Our second goal was to compare the performance of the children with that of an experience-based computational model fitted with a fixed processing limit. This allowed us to test whether we can explain individual differences in NWR performance, and its relationship with vocabulary, in terms of different amounts of linguistic experience. This model is CLASSIC (Chunking Lexical and Sublexical Sequences in Children), a computational cognitive model that has been used to simulate the learning and processing of lexical content in child-directed speech and experimental tasks (Jones, 2016; Jones et al., 2007, 2014; Jones & Macken, 2018; Jones & Rowland, 2017).

CLASSIC is a derivative of the EPAM/CHREST[3] architecture (Gobet et al., 2001) that uses an associative chunking mechanism to both process incoming utterances and expand the current knowledge base. Chunking is a process where multiple individual elements are compressed and recoded into a single perceptual unit (Cowan, 2010; Gobet et al., 2001; Miller, 1956), such as when strings of phonemes are grouped to form a word. This process of building chunked representations through association is a fundamental part of human learning and processing, explaining both why experience leads to more efficient processing of a stimulus within the confines of a limited processing window (Cowan et al., 2004; Gobet & Clarkson, 2004; Miller, 1956) and why experience influences acquisition (Christiansen, 2019; Jones & Rowland, 2017; McCauley & Christiansen, 2011, 2017; Perruchet, 2019; Perruchet & Pacton, 2006).

Figure 1 provides a schematic representation of how CLASSIC learns. At the outset of learning, CLASSIC has no knowledge of the target language (here, British English) other than an inventory of phonemes.[4] Learning progresses in a very simple way: First, an utterance is coded into as few chunks as possible based on whatever chunks CLASSIC has learned (the Processing layer in Figure 1), and second, adjacent chunks are grouped to form a new chunk (the Learning layer in Figure 1). Initially, CLASSIC represents an entire utterance phonemically; for example, it will require nine chunks to fully comprehend a nine-phoneme utterance such as *Where's the dog* ([w/e/ə/z] [ð/ə] [d/ɒ/g]; in this notation, chunks are separated by "/"). Learning will form new chunks

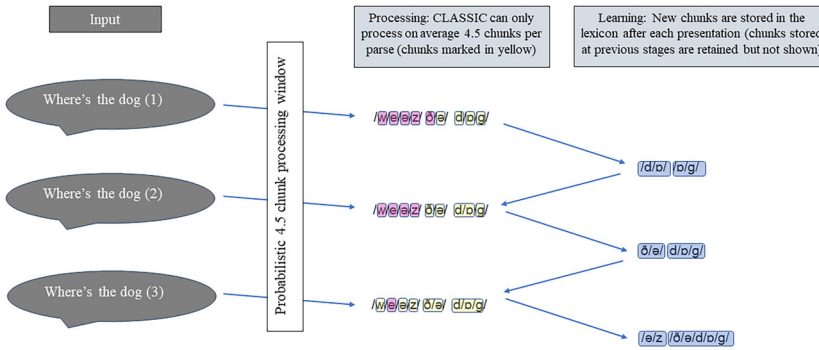

Figure 1 Learning in CLASSIC after three presentations of *Where's the dog*. Pink and yellow squares represent how CLASSIC chunks the input. Yellow squares indicate the chunks that CLASSIC has accessed in its processing window (on average 4.5 chunks per parse) which probabilistically favour the ends of utterances (here, the model processes 4, 4, and 5 chunks at each presentation respectively). Blue squares represent chunks learned from the input after each presentation. CLASSIC starts with only knowledge of phonemes. **1st presentation**: Processing: CLASSIC is limited to processing 4 chunks. Since it starts out with only knowledge of phonemes, it parses the input as 9 one-phoneme chunks, only 4 of which are accessed for learning ([ə],[d],[ɒ] and [g]). Learning: From the 4 one-phoneme chunks it has accessed, it creates (learns) two new, bigger chunks by combining adjacent accessed chunks and storing them in the lexicon ([d/ɒ/] and [ɒ/g/]; note that CLASSIC does not chunk phonemes across different words unless the words themselves are full chunks. **2nd presentation**: Processing: CLASSIC is limited to processing 4 chunks. It has already chunked [d/ɒ/], which enables it to process this two phoneme sequence as one chunk. This means that it can process more of the utterance than it did at the 1st presentation (5 phonemes as 4 chunks: [ð], [ə], [d/ɒ/], and [/g/]). Learning: From the 4 chunks it has accessed, it again combines adjacent chunks, resulting in two new, bigger, chunks: [ð/ə/] and [d/ɒ/g/]. Note that CLASSIC has now learned two complete words, each represented as 1 chunk: *the* ([ð/ə/]) and *dog* ([d/ɒ/g/]). **3rd presentation**: Processing: CLASSIC is limited to processing 5 chunks: the newly learned words [ð/ə/] and [d/ɒ/g/] together with 3 more chunks [w], [ə] and [z]. Learning: Adjacent chunks are then chunked again: [ə/z], [ð/ə/ d/ɒ/g/]. Note that because [ð/ə/] and [d/ɒ/g/] are whole words. (*the, dog*), CLASSIC chunks them into a phrase.

that pair adjacent chunks such as /dɒ/ and /ɒg./[5] Later, after some experience with the language, the model starts to learn word-level chunks, representing *the dog* using two single lexical chunks: /ðə/ and /dɒg/. Still later in development, phrases are learned, such as /ðədɒg/.

CLASSIC has a processing constraint set to an average of 4.5 chunks,[6] which probabilistically limits the number of chunks it can parse and access for learning. For example, if *the* is coded as two phonemes (/ð/, /ə/), the /ðə/ chunk

will only be learned if both /ð/ and /ə/ are accessed. However, the processing constraint does not vary across individual models or across development. This means that if CLASSIC simulates differences in NWR performance, it does so because of differences in the amount and type of linguistic knowledge stored in long-term memory, not intrinsic differences in processing capacity. For example, early in learning, a three-syllable nonword such as *doitervab* will need to be coded using many chunks (e.g., six: /d/, /ɔɪ/, /tə/, /v/, /æ/, /b/), so is unlikely to be processed in full or repeated correctly. However, later in learning it will be coded using a smaller number of chunks (e.g., four: /dɔɪ/, /tə/, /væ/, /b/), at which point there is a strong probability that the whole utterance will fit within the 4.5-chunk processing window, and be repeated correctly. In other words, "young" CLASSIC models, and CLASSIC models that receive less language input, perform less well on NWR tasks because they know less language. The more language CLASSIC is exposed to, the better it performs.

CLASSIC has been used previously to simulate NWR performance and its relationship with vocabulary development (e.g., Jones, 2016; Jones et al., 2007, 2014; Jones & Macken, 2018). Most relevant are Jones et al. (2007) and Jones (2016), both of which report successful simulations of 2–3- and 4–5-year-old children's NWR performance and its relationship with concurrent vocabulary. In the research presented here, we built on these studies in a number of ways. First and perhaps most importantly, we collected extensive longitudinal data on children's NWR performance and growth in vocabulary knowledge between 25 and 42 months of age. This enabled us to investigate whether the model's NWR performance predicted the relationship between NWR performance and vocabulary not just concurrently but also longitudinally over the next 2 years. We also widened the range of language experience levels tested by combining all utterances in a number of corpora of child-directed speech and extracting varying quantities to use as input. Finally, we created a novel method of matching children and simulations on vocabulary knowledge, which allowed us to make direct comparisons between them. To do this, we estimated the point in learning at which the simulations were equivalent to 25- and 31-month-old children in terms of vocabulary scores (these being the ages at which the NWR tests were administered), then compared the NWR performance of children and simulations at those ages, and assessed the effect of NWR performance at those points in learning on the simulation's concurrent and subsequent vocabulary growth.

To sum up, the present study had two goals. First, we created a new NWR task that was designed to be engaging for children as young as 24 months, validated it by determining whether the results were reliable over time (by comparing performance at 25 and 31 months), and tested whether it yielded

wordlikeness and word-length effects. We then tested whether performance on this new task predicted vocabulary growth over the next 2 years. Second, we simulated the data using a computational model to determine whether the relationship between NWR performance and vocabulary could be explained as emerging from differences in the amount and/or type of knowledge (particularly sublexical knowledge) stored in the mental lexicon.

Method

Participants

This study formed part of a longitudinal project (the Language 0–5 Project) approved by the University of Liverpool Research Ethics Committee. Ninety-five monolingual British English-speaking families were recruited: an initial sample of 89 families when the child was 6 months of age and an additional six families at 15 months of age. One child was excluded because of a persistent ear infection, and four families did not continue after the initial visit. By the end of data collection (at age 4;6), a further 13 had dropped out. All infants were born at full term, none were of low birth weight, and all were typically developing when recruited.

NWR tasks were administered at the 25- and 31-month age-points, and data on vocabulary were collected from parent report instruments (Lincoln CDI, CDI-3) at regular intervals between 25 and 37 months (seven age-points in total) and from a standardized test (BPVS-3) at 36 and 42 months. At 25 months, 75 children took part in the NWR task, 67 of whom (89%, 35 female) provided usable data (i.e., at least one valid response; mean age 25.86 months, range = 25.17–26.73). At 31 months, 74 children took part in the NWR tasks, 71 of whom (95%, 37 female) provided usable data (mean age 31.96 months, range = 31.03–32.93). CDI scores for vocabulary were available from all participants for at least one age-point; 85.3% provided data at all three Lincoln CDI age-points (25, 27, and 30 months), 71.6% provided data at the four CDI-3 age-points (31, 34, 36, and 37 months), and 94% provided data at both BPVS-3 age-points (36 and 42 months).

Materials

Nonword Repetition Test

We created a NWR task for 2-year-olds that was designed to be engaging, and that avoided, as much as possible, nonwords containing phonemes that are commonly misarticulated by 2-year-olds. Full details of the stimulus creation procedure and scoring, including which nonwords were included or excluded and why, can be found in Appendix S1 in the Supporting Information

online and in the Open Science Framework (OSF) repository. We started with a large list ($n = 123$) of one-, two- and three-syllable nonwords that had been used in previous tasks (Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1996; Jones et al., 2007, 2010, 2014; Roy & Chiat, 2004; Stokes & Klee, 2009; Tamburelli et al., 2012). We first coded the nonwords for articulatory ease based on work by Grunwell (1981). We then discarded, as much as possible, nonwords containing phonemes or phoneme combinations with which 2-year-old English-learning children are likely to make articulation errors, such as consonant clusters (e.g., *bl*) that are likely to be simplified and late-acquired sounds (e.g., coronal fricatives and liquids) that are likely to elicit stopping- and gliding-type substitutions (see also Keren-Portnoy et al., 2010). We coded items for word length in syllables and for wordlikeness (defined in terms of biphone probability and neighborhood density) and then chose and adapted items to reduce the possibility that regional accent differences could be misinterpreted as errors.

This resulted in 36 suitable nonwords divided into two lists of 18 nonwords for the 25-month and 31-month age-point respectively, which each included six 1-syllable, six 2-syllable, and six 3-syllable nonwords. Half of the nonwords at each word length were wordlike and half were not wordlike. The nonwords were placed in semirandom order within the lists (order of presentation: 1-syllable – 2-syllable – 3-syllable). We added natural English stress patterns (strong–weak for two syllables, strong–weak–weak for three syllables) and created phonetic transcriptions for training purposes.

Communicative Development Inventories

The Lincoln CDI (Meints et al., 2017) was administered at the 25-, 27-, and 30-month age-points. It is a British English adaption of the MacArthur-Bates CDI Words & Sentences, and comprises a parent report checklist that contains, among other sections, a vocabulary scale of the most common vocabulary items in UK children's vocabulary between 18 and 30 months of age (total possible vocabulary score = 689). The CDI-3 (Dale, 2007) was administered at the 31-, 34-, 36-, and 37-month age-points. It is a British English CDI created for the UK-based Twins Early Development Study and comprises a brief parent report checklist for children between 30 and 37 months that contains a short vocabulary scale (total possible vocabulary score = 100) and a short syntactic complexity section (not reported). Information about CDI construction, validity, and reliability is provided by Fenson et al. (2007).

British Picture Vocabulary Scale, Third Edition

The BPVS-3 was administered at 36 and 42 months of age. It is a standardized measure of receptive vocabulary for 3- to 16-year-old British English children, with good reliability and validity (Dunn et al., 2009). Children are asked to point to the picture that best matches a word's meaning, choosing from an array of four images. A stopping rule is applied when children respond incorrectly to eight or more target items in a set (14 sets, each containing 12 target items; total possible score = 168).

Other Materials

A puppet (Franklin the Frog), a Fuzzy Felt picture board and stickers, a video camera to record the session for offline coding, and a score sheet were also used.

Procedure

The NWR task was embedded in a longer lab session, including additional tasks not reported here. The experimenter explained, "We are going to play a copying game; Franklin [the frog hand puppet] will say some words and we have to copy them. Some of the words sound funny, but we'll try and copy them anyway." The session began with three real-word practice items (*cow*, *button*, *elephant*) followed by the 18 nonwords. Parents could help the children repeat the practice items only. Each time the child attempted to copy the nonword, they received a Fuzzy Felt sticker to place on a board. By the end of the study, the child had created a full picture on the board.

All nonwords were produced live by the experimenter using a natural prosodic pattern (see also Keren-Portnoy et al., 2010; Roy & Chiat, 2004; Stokes & Klee, 2009; Torrington Eaton et al., 2015). Each nonword was modeled a maximum of twice, always in a carrier phrase (e.g., "Can you say … ?"). If the child failed to respond to the second iteration, the experimenter encouraged the child once more but without repeating the word (e.g., "Can you say it? What did he say?"). If there was still no response, they moved on to the next item.

Simulations

The CLASSIC model was rebuilt using the Python 3.10 programming language. As input, we used the childesr package (Sanchez et al., 2019) to extract from the CHILDES database (MacWhinney, 2000) all fully intelligible utterances addressed by British English caregivers to children aged 1;10 to 4;10. This resulted in 712,441 utterances directed to 2-year-olds (1;10–2;10),

294,670 utterances directed to 34-year-olds (2;10–3;10), and 67,130 utterances directed to 4-year-olds (3;10–4;10). Each word in the transcripts was converted to its constituent phonemes using a lexicon (https://github.com/cmusphinx/cmudict), with phonemic transcripts for unknown words added manually for all words occurring with a frequency of 100 or more across all transcripts. This enabled us to retain 96% of the original orthographic utterances in their phonemic form. The boundaries between words were also retained.

Our prediction was that variance in language experience would predict individual differences in NWR task performance and its relationship with vocabulary. Thus, the number of utterances presented to CLASSIC was manipulated to capture the variance in language input that different children hear in their learning environments. Data from all CHILDES transcripts were pooled, and novel input samples were generated by randomly selecting utterances from the pooled data. Temporal characteristics of the utterances were retained by creating 10 sample bins to randomly select utterances from, each reflecting the time-point at which the utterance occurred; for example, Bin 1 contained utterances to children aged 1;10∼2;2, bin 2 from 2;2∼2;6 etc. The number of utterances in the resampled transcripts ranged from 1,500 to 120,000, increasing in equal increments of 1,500 utterances and retaining the temporal characteristics of the original transcripts; for example, the 1,500-utterance samples contained 150 utterances from Bin 1, then 150 utterances from Bin 2, and so on. Thus, we simulated learning at 80 unique input quantity levels from low (1,500 utterances) to high (120,000). Note that all inputs captured utterances that appeared across the whole age range and captured individual differences in exposure to language by varying from 1,500 utterances to 120,000 utterances. Utterances aimed at 2–3-year-old children constituted one third of each input, which is consistent with previous simulations (e.g., Jones et al., 2007).

Since learning in the model is dependent on a processing constraint that averages 4.5 chunks, the results were slightly different each time the models were run. Thus, five separate samples were generated for each quantity level and were individually presented to the model. In total, the results of the present work were based on 400 (5×80 input levels) simulations.

Comparing Children and Simulations

To compare child and simulation performance, we identified the stages in the learning cycle at which the simulations could be considered to be equivalent to a 25- and 31-month-old child in terms of vocabulary. We estimated this based on the median number of words on the MacArthur CDI Words & Sentences known by English-learning children according to the data from the CDI

available on Wordbank (Frank et al., 2016; $N = 4{,}868$). To match to the 25-month age-point, we extracted from Wordbank the median number of words that were produced by 25-month-old English-learning children: $Mdn = 373$. We then identified the point in the learning cycle at which the median model also knew approximately 373 words on the CDI, which was after 13% of input had been seen (see below for how we calculated how many words the simulations knew). We used this point in the learning cycle (13%) as the 25-month age-point for all models. To match to the 31-month age-point, we extracted from Wordbank the median number of words that were produced by 30-month-old English-learning children: $Mdn = 558$ (no 31-month data are available on Wordbank). The median model knew approximately 558 words after 48% of input had been seen, so we used this point in the learning cycle (48%) as the 31-month age-point for all models.

Coding the Nonword Repetition Task

Children

Responses were phonemically transcribed in ELAN (https://archive.mpi.nl/tla/elan) and exported into a CSV file for coding. The following were excluded: null responses, unclear responses (e.g., a mumbled response or a response obscured by something in the child's mouth), responses to targets in which the experimenter's voice was obscured (e.g., by background noise), and responses in which the experimenter made a pronunciation error or exceeded the number of presentations allowed. We also excluded items repeated spontaneously by the child (e.g., not immediately preceded by the target), and those for which the parent accidentally prompted the child. If a child repeated a nonword more than once, we scored their first scorable attempt. There were, on average, 16.43 valid responses per child at 25 months (range = 8–18, $SD = 2.56$) and 17.23 at 31 months (range = 7–18, $SD = 1.68$). Most invalid responses were unclear or null responses (67% of invalid responses at 25 months, 62% at 31 months).

Although we had excluded most phonemes or phoneme combinations that young children find difficult to articulate, this had not always been possible, so we created two coding schemes, one that allowed for common articulatory substitutions and one that did not. We also coded for errors at different levels: at the word level (1 = all consonants correct; 0 = at least one error) and the phoneme level (proportion of consonants correct[7]), since there is no coding consensus on this in the literature. We ran assumption tests for normality and chose the scheme that best fitted the assumptions for statistical analysis: the word-level coding scheme that allowed for common articulatory errors. However, note that the pattern of results reported below was replicated across

all four schemes. Interrater agreement, calculated on the 25-month data from eight participants (items = 276) by two coders trained in phonemic transcription, was 93.8% (Cohen's kappa = .86). The outcome variable for item-level analysis was simply whether the item had been repeated correctly or not (0/1). Participants' total NWR scores were calculated as a proportion of trials out of the total number of trials where the item was successfully repeated at 25 months and 31 months.

Simulations

The simulations were given the same NWR items as the children at the 25-month age-point (after 13% of training input) and 31-month age-point (after 48%). No learning took place during this test battery. Each nonword was presented 100 times to gain a reliable estimate of nonword accuracy because the probabilistic processing constraint accesses 4.5 chunks *on average*. The outcome variable for item-level analysis was the proportion of times all the chunks that were needed to process the nonword fitted within the 4.5-chunk processing constraint (i.e., the proportion of times out of 100 that the nonword could be considered to have been "repeated" correctly; from 0 to 1), which was then averaged across the five simulations at each input level. The simulations' total NWR score was calculated by averaging across the scores for each nonword to yield a score out of 1 at the equivalent of 25 months and 31 months.

Coding: Vocabulary

Children

Expressive vocabulary CDI scores were calculated according to the instructions in the manual. BPVS scores were calculated by summing the total number of correct picture choices made by the child before the stopping rule was applied.

Simulations

CDI scores were calcuated by identifying how many vocabulary items from the Lincoln CDI and CDI-3 existed as a single chunk in the model's learned chunks (i.e., had been learned as one whole entity; this included phrases such as *thank you*). Where the CDI offered alternatives (e.g., *telly*/*TV*/*television*), knowledge of any of the items as a single chunk was taken as knowledge of the vocabulary item. BPVS scores for individual items were assessed as correct if the model knew the word as a single chunk. The score was the total number of correct items identified before the stopping rule was applied. When children do not know the correct answer in the BPVS-3 test, they tend to point to a picture

at random, so we enabled the simulations to do the same (i.e., if the simulation did not know a word, it guessed at random). A similar pattern of results was obtained whether guessing was allowed or not.

Results

Analyses were conducted in R (Version 4.3.2; R Core Team, 2023). Models were fitted using the lme4_1.1-35-1 package (Bates et al., 2015). The level of significance was set at .05 (two tailed). Glmers were fitted using maximum likelihood estimation and lmers using restricted maximum likelihood estimation. Both used the nloptwrap optimization algorithm unless otherwise specified. All categorical variables were effect coded (1, −1), and all continuous predictors were centered and standardized on a standard deviation scale unless otherwise specified. Confidence intervals were computed using parametric bootstrapping (1,000 iterations), and p values were obtained via t tests with Satterthwaite's method for lmers and Wald's tests for glmers. Maximal models were reduced systematically if they led to convergence errors or a singular fit, first by removing the random correlation parameter and then by systematically removing random slopes. The assumptions of the relevant test were met for all final models used. Output, including assumption testing, is available at the project's OSF site.[8]

Reliability and the Effects of Wordlikeness and Word Length

Children

There was a moderate, significant correlation between the participants' NWR task scores at 25 and 31 months (Pearson's $r = .49$, bootstrapped 95% CI [.28, .76], $df = 60$, $p < .001$), suggesting that the task captured individual differences in performance that remained stable over the 6-month interval.

To test for wordlikeness and word-length effects, binomial generalized linear mixed-effects models were fitted to the 25- and 31-month data. The outcome measure was NWR score, binomially coded (1 = correct, 0 = incorrect), and the fixed-effects structure consisted of wordlikeness (sum coded; high = 1, low = −1) crossed with nonword length (in syllables) as a continuous predictor. The maximal models supported by the data included random intercepts for subjects and items and by-subject random slopes for wordlikeness and length but not their interaction. The model for 31 months also included the correlation parameter between random effects.

Descriptive statistics are illustrated in Figure 2, and parameter estimates and fit metrics are provided in Table 1. As predicted, the children were

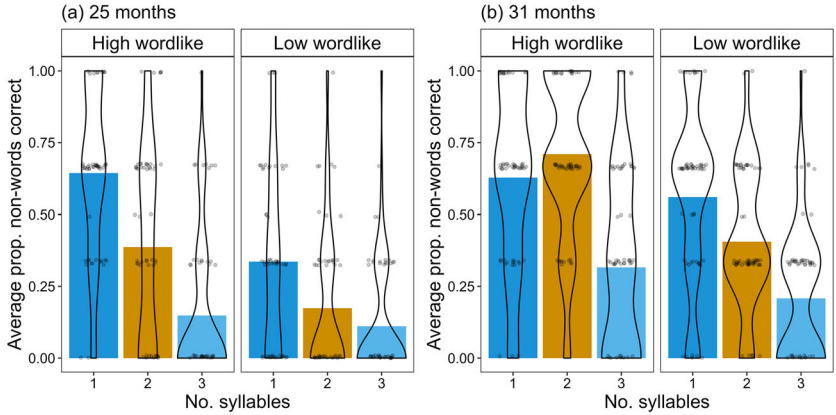

Figure 2 Children: Effect of wordlikeness and length on nonword repetition scores.

significantly better at repeating the highly wordlike nonwords and the shorter nonwords at both 25 and 31 months.

Simulations

There was a strong correlation between NWR scores at 25 and 31 months (Pearson's $r = .89$, bootstrapped 95% CI [.83, .90], $df = 78$, $p < .001$), suggesting that, as with the children, the task was capturing differences that remained stable over time. To test for wordlikeness and word-length effects, linear mixed-effects models were fitted separately to the 25- and 31-month data. The outcome measure was the proportion of correct repetitions for each item (see the section on coding above). The fixed-effects structure consisted of wordlikeness (high $= 1$, low $= -1$) crossed with nonword length (in syllables) as a continuous predictor. For both models, the maximal random-effects structure supported by the data included random intercepts for subjects and items. At 25 months it also included by-subject random slopes for length (but not wordlikeness) but no correlation between random effects. At 31 months it included by-subjects random slopes for length and wordlikeness, and their interaction, but no correlation between random effects.

Descriptive statistics are provided in Figure 3, and parameter estimates and fit metrics are provided in Table 2. Like the children, the models showed significantly better performance with the highly wordlike nonwords and shorter nonwords at both 25- and 31-month age-points.

Table 1 Children: Results of statistical models estimating the effect of wordlikeness and nonword length (in syllables) on nonword repetition

Age-point	Term	*b* [95% CI]	*SE*	*z*	*p*
25 months	(Intercept)	−1.20 [−1.53, −0.80]	0.19	−6.35	< .001
	Wordlikeness	0.64 [0.31, 0.93]	0.14	4.45	< .001
	Nonword length	−0.88 [−1.13, −0.57]	0.14	−6.20	< .001
	Wordlikeness × Length	−0.21 [−0.53, 0.09]	0.14	−1.47	.14
31 months	(Intercept)	−0.27 [−0.73, 0.19]	0.23	−1.20	.23
	Wordlikeness	0.46 [0.02, 0.87]	0.21	2.20	.03
	Nonword length	−0.81 [−1.25, −0.38]	0.21	−3.80	< .001
	Wordlikeness × Length	0.12 [−0.34, 0.53]	0.21	0.59	.56
25 months: AIC = 1,129.63, BIC = 1,179.58, R^2_m = .23, R^2_c = .39, ICC = .21, RMSE = 0.37					
31 months: AIC = 1,449.32, BIC = 1,505.36, R^2_m = .16, R^2_c = .42, ICC = .31, RMSE = 0.41					

Note. Confidence intervals were calculated using parametric bootstrapping (1,000 iterations). The *p* values are based on asymptotic Wald tests (*z* tests). AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

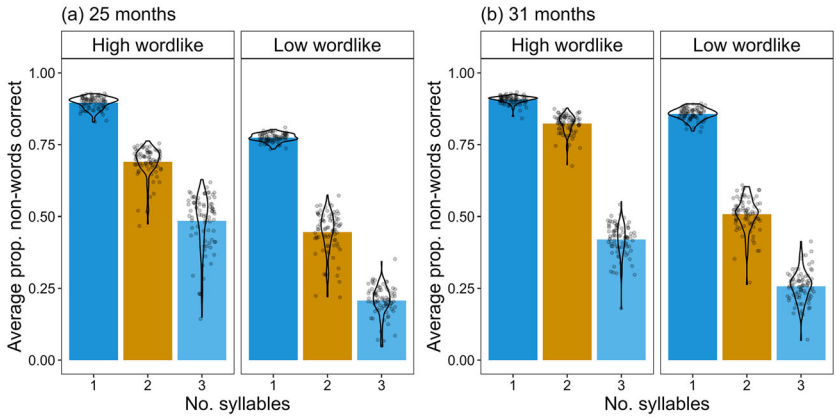

Figure 3 Simulations: Effect of wordlikeness and length on nonword repetition scores.

Nonword Repetition and Vocabulary Development Based on Scores From Communicative Development Inventories

Children

To test for the strength of the relation between NWR performance and concurrent vocabulary, we fitted Pearson's correlations with bootstrapped confidence intervals to NWR scores and vocabulary as measured by the Lincoln CDI at 25 months and the CDI-3 at 31 months. There were significant, medium-sized correlations at both 25 months (Pearson's $r = .51$, bootstrapped 95% CI [.34, .71], $df = 60$, $p < .001$) and 31 months (Pearson's $r = .32$, bootstrapped 95% CI [.10, .55], $df = 66$, $p = .01$).

We then tested whether children's NWR scores predicted vocabulary growth over the subsequent 6 months. Because growth was not linear, we fitted two growth curve mixed-effects models, one using NWR scores from the 25- and one from the 31-month age-point. The outcome measures for each model were, respectively, (1) Lincoln CDI scores between 25 and 30 months and (2) CDI-3 scores between 31 and 37 months. Development was modeled using linear (age^1) and quadratic (age^2) effects of age in months (not centered or standardardized). NWR scores (proportion of responses correct) were entered as a continuous predictor, crossed with both polynomial age predictors. The maximal random-effects structure supported by the data included subject as a random intercept, with age^1 and age^2 as random slopes but with the correlation parameter between random effects removed.

Table 2 Simulations: Results of statistical models estimating the effect of wordlikeness and nonword length (in syllables) on nonword repetition

Age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	0.58 [0.52, 0.65]	0.03	17.83	< .001
	Wordlikeness	0.11 [0.04, 0.17]	0.03	3.31	.01
	Nonword length	−0.20 [−0.26, −0.13]	0.03	−6.15	< .001
	Wordlikeness × Length	0.03 [−0.03, 0.10]	0.03	0.99	.34
31 months	(Intercept)	0.63 [0.52, 0.65]	0.03	24.85	< .001
	Wordlikeness	0.09 [0.04, 0.17]	0.03	3.52	.003
	Nonword length	−0.22 [−0.26, −0.14]	0.03	−8.81	< .001
	Wordlikeness × Length	0.02 [−0.03, 0.10]	0.03	0.92	.38
25 months: AIC = −3,906.00, BIC = −3,863.82, R^2_m = .69, R^2_c = .96, ICC = 0.88, RMSE = 0.05					
31 months: AIC = −4,296.85, BIC = −4,223.04, R^2_m = .81, R^2_c = .97, ICC = 0.84, RMSE = 0.05					

Note. Confidence intervals were calculated using parametric bootstrapping (1,000 iterations). The *p* values are based on Satterthwaite's method (*t* test). AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

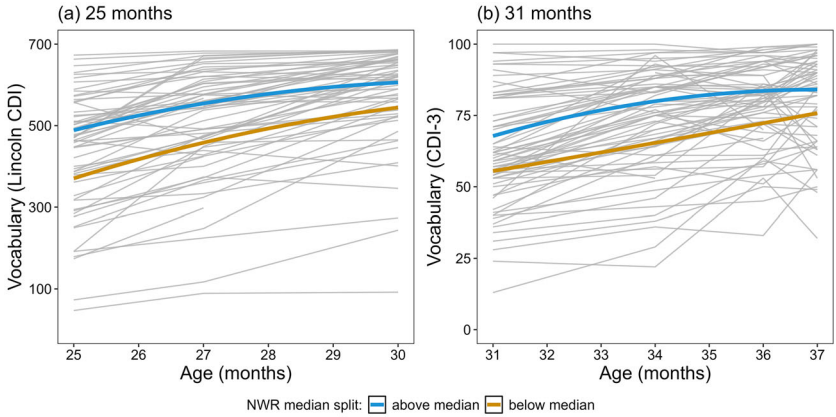

Figure 4 Children: Effect of nonword repetition (NWR) performance at 25 and 31 months on subsequent vocabulary growth as measured by Communicative Development Inventory (CDI) scores. Blue and orange lines illustrate language growth for children with NWR scores below and above the median, respectively. Light gray lines show the developmental trajectories of individual children.

The results are presented in Figure 4 and Table 3. In both models, a main effect of linear age on vocabulary (and quadratic growth at 25 months) indicated that the children's vocabulary grew steadily with age. In both models, a main effect of NWR performance indicated that children with higher NWR scores had larger vocabularies, and that this difference was sustained over the next 6 months. There was a significant interaction between NWR performance at 25 months and linear growth, but this was because children with better NWR performance showed decelerating, rather than accelerating, vocabulary growth over time (see Figure 4A). This is almost certainly because the fastest developing children reached ceiling on the Lincoln CDI before the end of the testing period.

Simulations

To test for the strength of the relation between NWR performance and concurrent vocabulary in the simulations, we fitted Pearson's correlations with bootstrapped confidence intervals to NWR and vocabulary scores as measured by CDIs (Lincoln CDI at 25 months; CDI-3 at 31 months). There were large, significant correlations between NWR performance and vocabulary at both 25 months (Pearson's $r = .93$, bootstrapped 95% CI [.90, .96], $df = 78$, $p < .001$)

Table 3 Children: Results of growth curve mixed-effects models estimating the effect of nonword repetition (NWR) performance on subsequent vocabulary growth as measured by Communicative Development Inventories

NWR age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	508.63 [480.79, 535.53]	13.79	36.89	< .001
	Age^1	762.71 [651.73, 874.70]	57.29	13.32	< .001
	Age^2	−95.90 [−165.93, −22.94]	36.13	−2.66	.01
	NWR	56.62 [28.42, 83.73]	13.72	4.13	< .001
	Age^1 × NWR	−187.12 [−293.06, −66.71]	57.76	−3.24	.002
	Age^2 × NWR	−0.74 [−71.78, 71.39]	36.33	−0.02	.98
31 months	(Intercept)	73.80 [70.24, 77.24]	1.79	41.14	< .001
	Age^1	112.63 [92.94, 133.46]	10.43	10.81	< .001
	Age^2	−13.85 [−31.52, 2.75]	8.55	−1.62	.11
	NWR	5.00 [1.65, 8.85]	1.81	2.77	.01
	Age^1 × NWR	−15.67 [−35.40, 5.59]	10.45	−1.50	.14
	Age^2 × NWR	−4.70 [−20.81, 10.84]	8.54	−0.55	.59
25 months: AIC = 2,103.50, BIC = 2,135.82, R^2_m = .34, R^2_c = .94, ICC = 0.91, RMSE = 23.52					
31 months: AIC = 1,940.56, BIC = 1,986.50, R^2_m = .22, R^2_c = .88, ICC = 0.85, RMSE = 4.64					

Note. Age^1 = linear effects of age in months; age^2 = quadratic effects of age in months. Confidence intervals were computed using parametric bootstrapping (1,000 iterations). The *p* values were obtained via *t* tests with Satterthwaite's method. AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

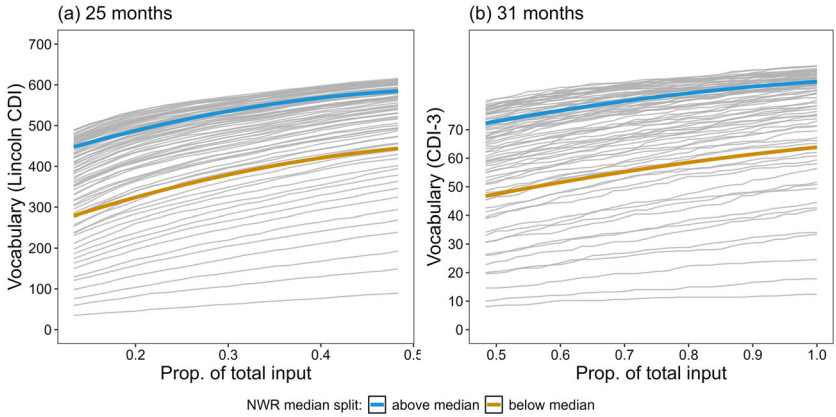

Figure 5 Simulations: Effect of nonword repetition (NWR) performance at 25 months (13% of learning) and 31 months (48% of learning) on subsequent vocabulary growth as measured by Communicative Development Inventory (CDI) scores. Blue and orange lines illustrate vocabulary growth for simulations with NWR scores above and below the median, respectively. Light gray lines show the developmental trajectories of individual simulations.

and 31 months (Pearson's $r = .90$, bootstrapped 95% CI [.85, .96], $df = 78$, $p < .001$).

We assessed the effect of NWR performance on subsequent vocabulary growth by fitting two separate growth curve mixed-effects models, one using NWR scores from the simulations' equivalent of 25 months (1) and one from 31 months (2). The outcome measures were, for Model 1, Lincoln CDI scores after the simulations' equivalent of 25 months up until 31 months, and, for Model 2, CDI-3 scores after 31 months to the end of the learning cycle. Development was modeled using the linear (age^1) and quadratic (age^2) effects of "age" (i.e., learning progress, not centered or standardardized). NWR scores (proportion of responses correct) were entered as a continuous predictor. The maximal random-effects structure supported by the data included subject as a random intercept with a by-subject random slope for age^1 and age^2 and the random-effects correlation parameter.

For results, see Figure 5 and Table 4. The results of the simulations mirrored those of the children. In both models, there were main effects of linear (and quadratic) growth, indicating that the simulations' vocabulary grew over the learning period. There was a significant effect of NWR performance, indicating that simulations with higher NWR scores at both 25 and 31 months

Table 4 Simulations: Results of growth curve mixed-effects models estimating the effect of nonword repetition (NWR) performance on subsequent vocabulary growth as measured by Communicative Development Inventories

NWR age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	457.87 [450.85, 464.66]	3.35	136.52	< .001
	Age^1	315.98 [291.16, 342.64]	14.04	22.51	< .001
	Age^2	−418.66 [−436.59, −399.12]	9.36	−44.72	< .001
	NWR	99.21 [92.92, 105.46]	3.35	29.57	< .001
	Age^1 × NWR	−105.41 [−135.27, −77.23]	14.04	−7.51	< .001
	Age^2 × NWR	−63.11 [−80.59, −46.81]	9.36	−4.97	< .001
31 months	(Intercept)	58.61 [56.58, 60.77]	1.06	55.35	< .001
	Age^1	92.62 [88.04, 97.05]	2.24	41.42	< .001
	Age^2	−13.25 [−14.94, −11.68]	−0.84	−15.88	< .001
	NWR	15.96 [13.92, 18.13]	1.06	15.67	< .001
	Age^1 × NWR	4.93 [0.35, 9.70]	2.24	2.21	.03
	Age^2 × NWR	−2.32 [−4.08, −0.69]	0.83	−2.78	.01
25 months: AIC = 10,148.28, BIC = 10,219.43, R^2_m = .92, R^2_c = 1.00, ICC = 0.99, RMSE = 3.03					
31 months: AIC = 4,248.68, BIC = 4,324.70, R^2_m = .78, R^2_c = 1.00, ICC = 1.00, RMSE = 0.40					

Note. Confidence intervals were computed using parametric bootstrapping (1,000 iterations). The *p* values were obtained via *t* tests with Satterthwaite's method. AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

had larger vocabularies. There were also significant interactions between NWR performance and linear growth (and with quadratic growth), but this was in the predicted direction only at 31 months. At 25 months the direction of the effect indicated that the interaction was because simulations with better NWR performance showed decelerating, rather than accelerating, growth over time. Again this is because the fastest developing simulations reached ceiling on the Lincoln CDIs before the end of the testing period.

Nonword Repetition and Vocabulary Development Based on BPVS-3 Scores

Children

Using CDI data, the analyses above revealed no significant evidence for an effect of NWR performance on the speed of vocabulary growth, but this was confounded by the fact that the fastest developing children and simulations reached ceiling on the CDIs, limiting the capacity for growth. Thus we ran new analyses using vocabulary scores from the BPVS-3 administered at 36 and 42 months of age. We fitted two separate linear mixed-effects models (one for 25-, one for 31-month NWR scores) with total BPVS raw score at 36 and 42 months as the outcome measure. Development was modeled using age in months (centered and standardized), and NWR test performance (proportion of responses correct) was entered as a continuous predictor. The maximal random-effects structure supported by the data included subject as a random intercept and a by-subject random slope for age, but with the correlation parameter between random effects removed.

For results, see Figure 6 and Table 5. At 25 months, there was a main effect of both age and NWR scores, and an interaction between age and NWR scores such that children with higher NWR scores not only maintained their initial advantage in vocabulary but also showed slightly faster rates of growth. However, at 31 months, NWR scores did not significantly predict either vocabulary knowledge or the speed of vocabulary growth from 36 to 42 months.

Simulations

We fitted two separate growth curve mixed-effects models, one for 25- and one for 31-month NWR scores. The outcome measures were total BPVS scores (1) after the model's equivalent of 25 months and (2) after the model's equivalent of 31 months. Development was modeled using the linear (age^1) and quadratic (age^2) effects of "age" (not centered or standardized). NWR scores (proportion of responses correct) were entered as a continuous predictor. For both models, the maximal random-effects structure supported by the data included subject

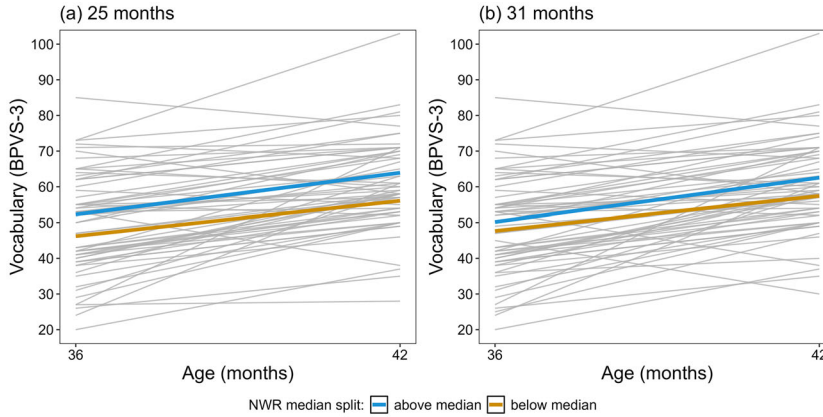

Figure 6 Children: Effect of nonword repetition (NWR) performance at 25 and 31 months on subsequent vocabulary growth as measured by BPVS-3 at 36 and 42 months. Blue and orange lines illustrate language growth for children with NWR scores below and above the median, respectively. Light gray lines show the developmental trajectories of individual children.

as a random intercept with a by-subject random slope for age[1] and age[2], and the correlation parameter between random effects. Note that the simulations' BPVS scores model a slightly earlier developmental period than the children's (median simulation's score at the end of training = 46.8, median child's score at 36 months = 49).

The results are illustrated in Figure 7, and parameter estimates and fit metrics are shown in Table 6. In both models, there were main effects of linear growth (and quadratic growth at 25 months), a significant effect of NWR performance, and significant interactions between NWR performance and linear growth; simulations with higher NWR scores not only maintained their advantage in vocabulary but also showed faster rates of growth.

Simulating Individual Children

The results in the previous section demonstrated that we are able to simulate the relationship between NWR performance and vocabulary growth when we match the median simulation with the median CDI vocabulary score of 2-year-old children. In this section we test the relationship more directly by matching the vocabulary of individual simulations to individual children. As can be seen from Figures 2 and 3, the model's NWR performance was better than that of the children, with less variation in performance across simulations, which

Table 5 Children: Results of linear mixed-effects models estimating the effect of nonword repetition (NWR) performance on subsequent vocabulary growth at 36 and 42 months as measured by the BPVS-3

NWR age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	55.03 [52.11, 57.84]	1.45	38.01	< .001
	Age	5.48 [4.16, 6.84]	0.65	8.40	< .001
	NWR	3.27 [0.35, 6.14]	1.46	2.24	.03
	Age × NWR	1.32 [−0.01, 2.65]	0.65	2.03	.047
31 months	(Intercept)	54.57 [51.91, 57.65]	1.41	38.81	< .001
	Age	5.52 [4.22, 6.81]	0.68	8.18	< .001
	NWR	2.02 [−0.75, 4.71]	1.41	1.43	.16
	Age × NWR	0.19 [−1.06, 1.59]	0.68	0.29	.78
25 months: AIC = 971.79, BIC = 991.70, R^2_m = .24, R^2_c = 1.00, ICC = 1.00, RMSE = 0.02					
31 months: AIC = 1,032.98, BIC = 1,053.26, R^2_m = .20, R^2_c = 1.00, ICC = 1.00, RMSE = 0.01					

Note. Confidence intervals were computed using parametric bootstrapping (1,000 iterations). The *p* values were obtained via *t* tests with Satterthwaite's method. AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

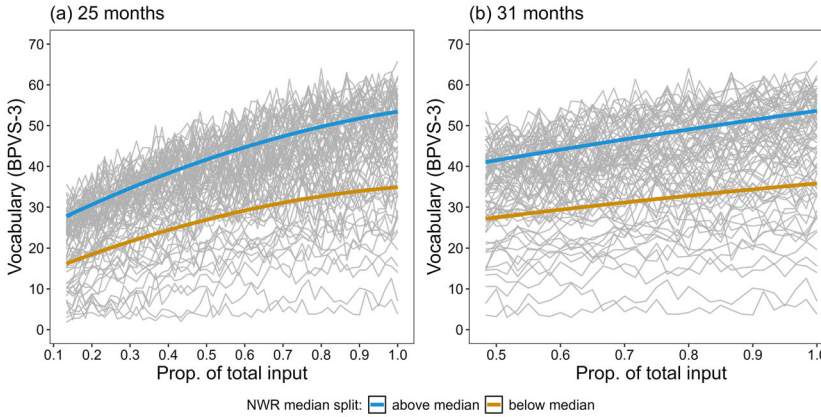

Figure 7 Simulations: Effect of nonword repetition (NWR) performance at 25 months (13% of learning) and 31 months (48% of learning) on subsequent vocabulary growth as measured by BPVS-3. Blue and orange lines illustrate vocabulary growth for simulations with NWR scores above and below the median, respectively. Light gray lines show the developmental trajectories of individual simulations. Note that the slopes for individual simulations are not smooth because, like the children, the model is allowed to "guess" if it does not know the answer.

means that increases in vocabulary may have a limited effect on the simulations' NWR performance when matched to individual children.

For each child, we determined their CDI vocabulary score at 25 and 31 months, and then identified the simulation that was the closest match in terms of vocabulary at the same "age" (i.e., after 13% of learning for the 25-month age-point, and after 48% of learning for the 31-month age-point). This yielded 38 matched child–simulation pairs at 25 months and 53 pairs at 31 months. The other children knew substantially more words than the best performing model so were excluded.

At each age-point, we correlated NWR and vocabulary scores for the subset of the simulations and children matched on vocabulary, and also correlated NWR performance between matched simulation–child pairs. Figures 8 and 9 and Table 7 illustrate the results.

For this subset of simulations, despite substantially less variance across the board, there were still strong and significant correlations between vocabulary and NWR scores at both 25 and 31 months. For this subset of the children, there were significant large (at 25 months) and medium-sized (at 31 months) correlations between vocabulary and NWR scores at 25 months. There were

Table 6 Simulations: Results of linear mixed-effects models estimating the effect of nonword repetition (NWR) performance on subsequent vocabulary growth at 36 and 42 months as measured by the BPVS-3

NWR age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	33.20 [32.32, 33.98]	0.41	80.35	< .001
	Age^1	61.43 [59.52, 63.62]	0.99	61.96	< .001
	Age^2	−9.26 [−10.72, −7.80]	0.72	−12.90	< .001
	NWR	9.51 [8.68, 10.33]	0.41	23.02	< .001
	Age^1 × NWR	12.25 [10.32, 14.22]	0.99	12.36	< .001
	Age^2 × NWR	−0.89 [−2.36, 0.36]	0.72	−1.24	.22
31 months	(Intercept)	34.27 [32.98, 35.48]	0.65	52.81	< .001
	Age^1	51.08 [44.20, 59.99]	3.49	14.65	< .001
	Age^2	−2.85 [−6.60, 0.69]	1.83	−1.56	.12
	NWR	8.81 [7.57, 10.01]	0.65	13.58	< .001
	Age^1 × NWR	11.75 [4.93, 18.76]	3.49	3.37	.001
	Age^2 × NWR	−0.09 [−3.68, 3.69]	1.83	−0.05	.96
25 months: AIC = 21,372.24, BIC = 21,454.82, R^2_m = .86, R^2_c = .96, ICC = 0.66, RMSE = 2.81					
31 months: AIC = 13,169.77, BIC = 13,245.79, R^2_m = .73, R^2_c = .95, ICC = 0.80, RMSE = 2.79					

Note. Confidence intervals were computed using parametric bootstrapping (1,000 iterations). The *p* values were obtained via *t* tests with Satterthwaite's method. AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

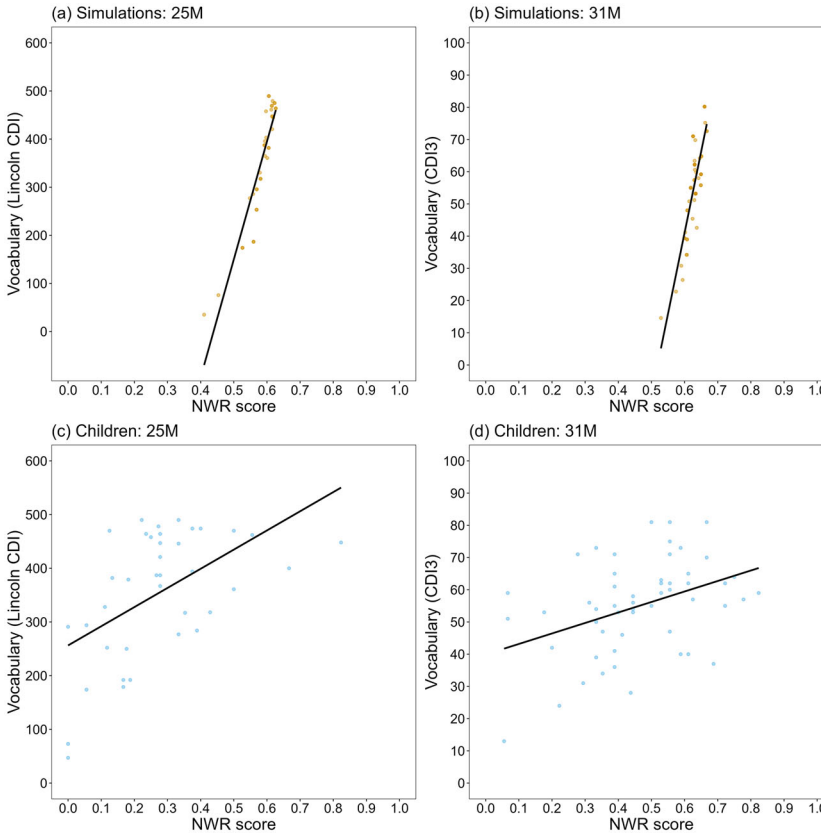

Figure 8 Relation between nonword repetition (NWR) performance and vocabulary for simulations (panels a, b) and children (panels c, d) matched on vocabulary at 25 and 31 months (M). Regression lines illustrate the strength of the linear relationships; points represent individual children or simulations. CDI = Communicative Development Inventory.

also moderate (at 25 months) and small (at 31 months) positive correlations between NWR scores for the children and simulations matched on vocabulary, though these were only significant at 25 months.

We next assessed whether NWR performance predicted longitudinal vocabulary growth as measured by BPVS-3 scores in the subset of simulations matched to the children. We do not include an analysis of growth as measured by CDI scores because the ceiling effects make it hard to draw strong conclusions about growth trajectory (though see supplementary materials on

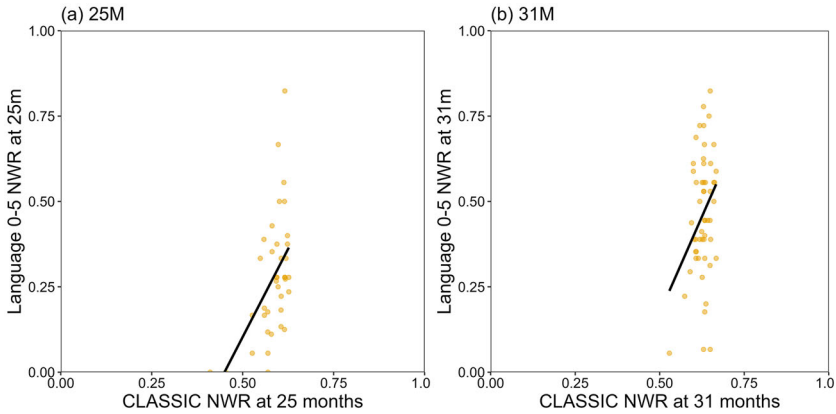

Figure 9 Relation between nonword repetition (NWR) performance in children and their matched simulation at 25 months (panel a) and 31 months (panel b). Regression lines illustrate the strength of the linear relationships; points represent individual child–simulation pairs.

Table 7 Results of bootstrapped correlations (1,000 iterations) for a subset of the children and simulations matched on vocabulary

Correlation pairs	Pearson's *r* [95% CI]	*df*	*p*
Sims: NWR & Lincoln CDI at 25 months	.92 [.88, .95]	36	< .001
Sims: NWR & CDI-3 at 31 months	.86 [.77, .95]	51	< .001
Children: NWR & Lincoln CDI at 25 months	.54 [.35, .73]	36	.01
Children: NWR & CDI-3 at 31 months	.40 [.17, .62]	51	.03
NWR: Children & simulations at 25 months	.52 [.34, .68]	36	.01
NWR: Children & simulations at 31 months	.31 [.03, .63]	51	.13

Note. Sims = simulations; NWR = nonword repetition; CDI = Communicative Development Inventory.

the project OSF site for the results of these analyses). We fitted two separate growth curve mixed-effects models, one for 25- and one for 31-month NWR scores. The outcome measure was total vocabulary as measured by BPVS-3. Development was modeled using the linear (age^1) and quadratic (age^2) effects of "age" (not centered or standardized). NWR scores were entered as a continuous predictor. The maximal random-effects structure supported by the data included subject as a random intercept with random slopes for age^1 and age^2 and the random-effects correlation parameter.

Table 8 and Figure 10 show the results. There were main effects of linear

Table 8 Matched subsample of simulations: Results of growth curve mixed-effects models estimating the linear (age^1) and quadratic (age^2) increase in BPVS-3 scores and the interaction with nonword repetition (NWR) score

NWR age-point	Term	*b* [95% CI]	*SE*	*t*	*p*
25 months	(Intercept)	29.77 [26.14, 33.12]	1.77	16.85	< .001
	Age^1	33.28 [9.48, 55.07]	11.49	2.90	.008
	Age^2	−27.89 [−41.97, −14.00]	6.86	−4.07	< .001
	NWR	8.42 [5.24, 11.67]	1.64	5.14	< .001
	Age^1 × NWR	10.29 [−10.63, 31.56]	10.84	0.95	0.35
	Age^2 × NWR	−4.86 [−17.64, 7.63]	6.50	−0.75	0.46
31 months	(Intercept)	31.98 [30.15, 33.78]	0.91	35.20	< .001
	Age^1	42.66 [33.60, 52.34]	4.82	8.81	< .001
	Age^2	−0.91 [−6.73, 4.75]	2.97	−0.31	0.76
	NWR	6.63 [5.11, 8.23]	0.81	8.19	< .001
	Age^1 × NWR	5.43 [−3.71, 14.08]	4.52	1.20	0.24
	Age^2 × NWR	0.19 [−4.93, 5.29]	2.72	0.07	0.94
25 months: AIC = 4,026.63, BIC = 4,088.10, R^2_m = .81, R^2_c = .94, ICC = 0.66, RMSE = 2.38					
31 months: AIC = 8,576.02, BIC = 8,646.69, R^2_m = .67, R^2_c = .91, ICC = 0.73, RMSE = 2.79					

Note. Confidence intervals were computed using parametric bootstrapping (1,000 iterations). The *p* values were obtained via *t* tests with Satterthwaite's method. AIC = Akaike information criterion; BIC = Bayesian information criterion; R^2_m = marginal R^2; R^2_c = conditional R^2; ICC = intraclass correlation coefficient; RMSE = root-mean-square error.

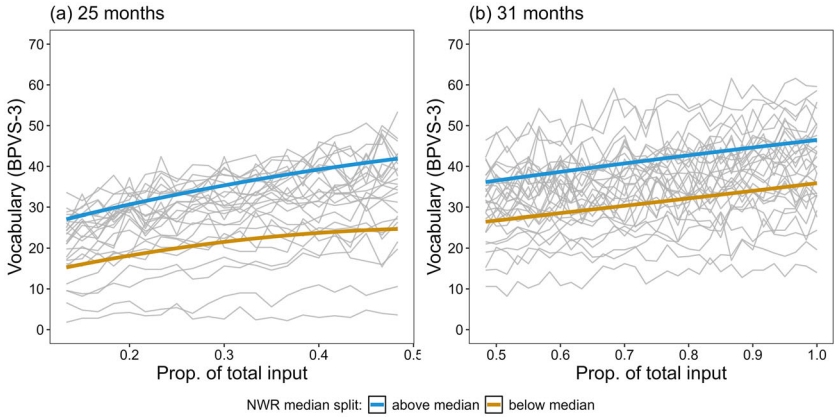

Figure 10 Matched simulations: Effect of nonword repetition (NWR) performance at 25 months (13% of learning) and 31 months (48% of learning) on subsequent vocabulary growth as measured by BPVS-3. Blue and orange lines illustrate vocabulary growth for simulations with NWR scores above and below the median, respectively. Light gray lines show the developmental trajectories of individual simulations.

age and NWR performance (and quadratic age at 25 months). Thus, even in this smaller subsample matched to the children, simulations with better NWR scores had better vocabulary. There was, however, no significant interaction between NWR and linear age at either age.

Discussion

In this study, we showed that a NWR task designed for 2-year-olds reduced dropout rates, was reliable over time, revealed the same length and wordlikeness effects seen in older children, and predicted vocabulary over time. These effects were also successfully simulated in the CLASSIC model. The results from the children suggest that our NWR task is a reliable tool that can complement existing tests in this age range, and the results from the simulations suggest that NWR performance is grounded in the gradual buildup of linguistic knowledge based on increased exposure to language. We expand on these points below.

The new NWR task was extremely successful at eliciting valid responses from 2-year-old children. It yielded a very low dropout rate (at least one valid response from 89% of the 25-month-olds and 95% of the 31-month-olds) and a high number of valid responses per child: *M*/total (*SD*) = 16.43/18 (2.56) at 25 months, 17.23/18 (1.68) at 31 months. We attribute this success to three de-

sign features. First, we embedded the task into an engaging game; the children enjoyed copying Franklin the Frog and building a Fuzzy Felt scene. Second, since 2-year-olds are often reluctant to repeat a recorded voice, the nonwords were produced live by the experimenters, who were trained to pronounce the phonemes in a particular way (to remove accent differences) using a natural prosodic pattern. Third, we avoided, as much as possible, phonemes and phoneme combinations that 2-year-olds find difficult to pronounce, and allowed common articulation errors where this was not possible.

The results were reliable. There was a significant, moderate correlation between individual children's performance at the two age-points, suggesting stability over time. We also replicated the effects of wordlikeness and word length that have been reported in the literature for older children; the children were better at repeating highly wordlike and shorter nonwords. Thus, we conclude that this new task is a reliable tool that researchers can use to assess NWR in 2-year-olds, which complements, in particular, the Preschool Repetition Test (for children aged 2 to 6 years; Chiat & Roy, 2007). The task might also prove useful for identifying language disorder, given that performance in NWR tasks provides a robust behavioral marker of developmental language disorder (Bishop et al., 1996; Conti-Ramsden et al., 2001; Dollaghan & Campbell, 1998). However, please note that this task was designed for English speakers living in the northwest of England, so may need to be modified to accommodate other variants of English. All task instructions and materials are freely available at the project OSF site.

We also determined whether NWR performance in 2-year-olds was associated with language growth. In line with our prediction, NWR performance not only correlated with concurrent vocabulary, but also predicted later vocabulary scores up to 2 years later. Thus, the present study is one of only a handful of studies showing a long-term relation between NWR performance and vocabulary in 2-year-olds, the age at which vocabulary grows most rapidly. That said, with vocabulary assessed using CDIs, there was no evidence that children with better NWR performance also showed faster rates of subsequent vocabulary growth. In fact, at 25 months the highest performers on the NWR task showed decelerating, not accelerating, growth. However, this is almost certainly because the fastest developing children reached ceiling on the Lincoln CDI. When we tested the relationship between NWR scores and language at a later age (36 and 42 months) using a different language scale (BPVS-3), we tentatively concluded that higher NWR scores at 25 (though not 31) months were associated not only with bigger vocabularies later in childhood but also with faster vocabulary growth.

Finally, we simulated the data using a computational model (CLASSIC) to determine whether NWR performance and vocabulary can be explained as emerging from differences in the amount and/or type of knowledge (particularly sublexical knowledge) stored in the mental lexicon. The results from the simulations largely mirrored those of the children. There was a significant correlation between performance at the two age-points and an effect of wordlikeness and word length. The simulations' NWR performance predicted subsequent vocabulary as measured by CDIs. At 31 months, there was some evidence that the simulations with better NWR performance had accelerated vocabulary growth, but at 25 months, the simulations with the highest NWR scores showed decelerating rates of growth, indicating that, like the children, the fastest simulations reached ceiling on the Lincoln CDI before the end of the learning period. Turning to vocabulary as measured by BPVS-3, there was a positive effect of NWR performance on both vocabulary knowledge and on the speed of vocabulary growth at both ages in the full models. In the subset of simulations matched to the children, there was an effect of NWR performance on vocabulary size but not the speed of growth.

Our results replicate those of Jones et al. (2007) and Jones (2016), showing that children's NWR performance and the relationship between NWR performance and language in the preschool years can be explained by simulations that vary in language experience alone. The results extend those studies by showing that this finding holds longitudinally over the next 2 years. Using vocabulary scores to match the simulations to children allowed us to make direct comparisons between child and model performance at critical age-points (25 months old = 13% of learning cycle; 31 months = 48% of learning cycle). These results suggest that the driver of individual and developmental differences in 2-year-olds may not be intrinsic differences in phonological working memory size but differences in the child's current knowledge in terms of lexical and sublexical representations (Jones et al., 2007; Szewczyk et al., 2018). Experience with language leads the simulations and the children to build lexical and sublexical chunks of linguistic information in long-term memory. Simulations and children that receive more input have larger stores of sublexical and lexical chunks, which they can use to solve NWR tasks and to build subsequent vocabulary more quickly (see Jones & Rowland, 2017, for evidence that simulations with a bigger store of sublexical chunks learn new words more quickly).

That said, there were differences in how the simulations and the children performed. Although, like the children, the simulations showed significantly better performance with more wordlike and shorter nonwords, overall

performance was higher than that of the children and showed substantially less variance. The simulations also learned to repeat the nonwords more quickly than the children (e.g., mean proportion correct at 25 months: .59 for simulations vs. .31 for children), and, because of this, there was very little subsequent growth in NWR performance (on average, a .03 increase from 25 to 31 months for the simulations, .16 for children), though the simulations still outperformed children at the later time-point. Even with a small amount of language input, the simulations could accurately repeat many of the nonwords, indicating that some of their constituent phoneme sequences were quickly learned as chunks by virtue of occurring often in the language input. Thus it seems that children may not be as able to capitalize on information shared across nonwords in the same way as the model.

One obvious explanation is that the model simulates a simplified learning process because it is designed to investigate the effects of language experience, not to mimic child performance exactly. For example, the model begins with perfect phonological representations of English phonemes, whereas 2-year-old children are still constructing their phonological representations (Dollaghan et al., 1995; Snowling et al., 1991). In addition, for the model, knowledge of a chunk equates to perfect reproduction of the chunk contents, whereas 2- to 3-year-old children are still developing proficiency in articulation (Roy & Chiat, 2004). To account for this, Jones et al. (2007) added an additional error parameter, but we chose not to include this in order to focus on knowledge gained from language experience alone. A further, substantial difference is that CLASSIC simply learns the phonological form of words without any semantic or syntactic knowledge. Within CLASSIC, as in children, word learning is influenced by the frequency with which a word occurs and the frequency with which the phonological material within the word has occurred (Braginsky et al., 2019), but word learning in children is also influenced by a number of other variables, including semantic density (Borovsky, 2022), concreteness, and valence (Braginsky et al., 2019). None of these predictors are captured in CLASSIC.

Another difference betweeen child and simulation performance was that the effect sizes for children at 31 months were smaller both than those for children at 25 months and for the simulations at both age-points. The correlations between the subset of children and their matched simulations were also smaller at 31 months than at 25 months. It is unlikely that this is due to the properties of the nonwords used at 31 months, because children and simulations received the same set, and because both 25- and 31-month sets were created using the same parameters.

One explanation is that the processing window (as measured by NWR tasks) is less of a constraint on vocabulary learning at 31 months in the children, because by then the store of lexical and sublexical representations available is big enough to enable children to process most of their incoming input. This idea is supported by the fact that the simulations had lower vocabulary scores than the children on average, and none of the simulations matched the vocabulary scores of the most advanced children by the end of the training period. This is almost certainly due to the restricted nature of the simulations' input compared to that of 2-year-old children, who will have heard both substantially more utterances over their 2-year lifespan and a much wider range of word types. This can be solved by increasing the amount and diversity of the input given to simulations.

More generally, our results raise questions about what constructs are measured by NWR tasks, and the relationship between these constructs and language. In CLASSIC, NWR task performance is driven by the interaction of linguistic knowledge and an intrinsic, but fixed and unchanging, constraint on how much of the input can be processed, implemented by a 4.5-chunk processing window. A key question, then, is: What is the equivalent of this processing window in children? It might be a constraint on phonological working memory (Baddeley, 1986b, 2000; Baddeley & Hitch, 1974; Gathercole, 2006; Gathercole & Baddeley, 1989), but could also be a constraint on phonological processing, phonological analysis and/or assembly, retrieval speed of representations from a phonological store, speed of speech motor planning, integration of perceptual and motor wordforms, or a combination of some or all of these capacities (Bowey, 1996, 1997; Davis & Redford, 2023; Dollaghan et al., 1995; Gathercole, 2006; Snowling et al., 1986). In fact, it may even be a general property of the language processing network itself; MacDonald and Christiansen (2002) have argued that "processing capacity emerges from network architecture and experience and is not a primitive that can vary independently" (p. 35). We are agnostic about what the constraint might be, but see this as an area where more work is needed.

Another question concerns the role of linguistic knowledge in NWR task performance. Gathercole (2006) suggested that linguistic knowledge affects the quality of phonological storage in phonological working memory. Hulme and colleagues (Hulme et al., 1991, 1997) suggested that linguistic knowledge contributes to redintegration: Children use stored phonological specifications retrieved from long-term memory to fill in incomplete phonological representations in phonological short-term memory. Schwering and MacDonald (2020) argued, more radically, that verbal working memory is simply the activated

portion of linguistic long-term memory, and is thus an emergent property of linguistic knowledge. The approach we favor, supported by our simulations, is one in which greater exposure to linguistic input leads learners to store phonological knowledge in chunks of information of varying size. This approach explains not only the relationship between NWR performance and language reported here, but also a number of other phenomena in both NWR performance (wordlikeness effects, word-length effects, individual differences, developmental changes; Jones, 2016) and in vocabulary acquisition (e.g., effect of input diversity and quantity, phonotactic probability, neighborhood density; Jones et al., 2021; Jones & Rowland, 2017).

Conclusion

We have demonstrated that a new NWR task designed for 2-year-olds can reliably and robustly capture NWR performance in young children, including wordlikeness effects, word-length effects, and the strong relationship between NWR performance and vocabulary. Our results do not support the view that NWR tasks measure a capacity that intrinsically differs across individual children and increases with age. Instead, we suggest that exposure to linguistic input, filtered through a fixed-capacity processing constraint, leads learners to store phonological knowledge in chunks of information of varying size, and it is this stored knowledge that influences both NWR performance and voabulary growth.

Acknowledgments

Open access funding enabled and organized by Projekt DEAL.

Final revised version accepted 19 June 2024

Notes

1 Differences in processing constraints such as phonological working memory could also potentially be a relevant variable, but, thus far, the model has been able to explain differences using input alone.

2 We added the analyses of BPVS-3 scores after review, because a ceiling on the CDI scores limited the conclusions we could draw.

3 EPAM stands for Elementary Perceiver and Memoriser and CHREST for Chunk Hierarchy and REtrieval STructures (see http://www.chrest.info/).

4 The fact that the model begins with phonemes is at odds with the view that infant speech perception does not (see, e.g., Vihman, 2017). However, this is an implementation decision, and, in fact, individual phoneme representations are quickly subsumed by phoneme sequences during learning. The model could equally

begin with syllables, biphones, or even "indistinct representations from which distinct elements begin to emerge with the child's ability to segment incoming speech," as pointed out by a reviewer.

5 Learning is not allowed to cross word boundaries unless the chunks themselves are words, to simulate the fact that 2-year-old children can already segment words from the speech stream (see, e.g., Johnson & White, 2019). However, this constraint is not essential; see Jessop et al. (2024) for a modified version of CLASSIC that learns both to segment the speech stream and to acquire words through chunking.

6 A limit of 4.5 chunks was chosen to bridge the initial concept of chunking from Miller (1956), who suggested people can hold 7 ± 2 chunks, and more recent work that has suggested this limit may be 4 or fewer chunks (e.g., Cowan, 2001; Gobet & Clarkson, 2004). A limit of 4.5 chunks has been used in all previous work with CLASSIC.

7 We did not code vowels because the transcription of vowels is much less reliable than that of consonants (see, e.g., Davis et al., 2002).

8 A previous version of this paper included an analysis of the effect of NWR performance on syntactic growth. The results can be found in the output file in the supplementary materials on the OSF site.

References

Adams, A.-M., & Gathercole, S. E. (1995). Phonological working memory and speech production in preschool children. *Journal of Speech, Language, and Hearing Research*, *38*(2), 403–414. https://doi.org/10.1044/jshr.3802.403

Adams, A.-M., & Gathercole, S. E. (2000). Limitations in working memory: Implications for language development. *International Journal of Language & Communication Disorders*, *35*(1), 95–116. https://doi.org/10.1080/136828200247278

Baddeley, A. D. (1986a). *Human memory: Theory and practice*. Erlbaum.

Baddeley, A. D. (1986b). *Working memory*. Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bishop, D. V., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of*

Child Psychology and Psychiatry, and Allied Disciplines, *37*(4), 391–403. https://doi.org/10.1111/j.1469-7610.1996.tb01420.x

Borovsky, A. (2022). Lexico-semantic structure in vocabulary and its links to lexical processing in toddlerhood and language outcomes at age three. *Developmental Psychology*, *58*(4), 607–630. https://doi.org/10.1037/dev0001291

Bowey, J. A. (1996). On the association between phonological memory and receptive vocabulary in five-year-olds. *Journal of Experimental Child Psychology*, *63*(1), 44–78. https://doi.org/10.1006/jecp.1996.0042

Bowey, J. A. (1997). What does nonword repetition measure? A reply to Gathercole and Baddeley. *Journal of Experimental Child Psychology*, *67*(2), 295–301. https://doi.org/10.1006/jecp.1997.2408

Bowey, J. A. (2001). Nonword repetition and young children's receptive vocabulary: A longitudinal study. *Applied Psycholinguistics*, *22*(3), 441–469. https://doi.org/10.1017/S0142716401003083

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology*, *32*(5), 867–873. https://doi.org/10.1037/0012-1649.32.5.867

Chiat, S., & Roy, P. (2007). The preschool repetition test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research*, *50*(2), 429–443. https://doi.org/10.1044/1092-4388(2007/030)

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*(6), 635–645. https://doi.org/10.1111/j.1469-7610.2008.01881.x

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468–481. https://doi.org/10.1111/tops.12332

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders*, *43*(1), 1–40. https://doi.org/10.1080/13682820601116485

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(6), 741–748. https://doi.org/10.1111/1469-7610.00770

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science*, *15*(9), 634–640. https://doi.org/10.1111/j.0956-7976.2004.00732.x

Dale, P. (2007). *The MacArthur-Bates Communicative Development Inventory III*. Paul H. Brookes Publishing Company.

Davis, B. L., MacNeilage, P. F., & Matyear, C. L. (2002). Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, *59*(2–3), 75–107. https://doi.org/10.1159/000066065

Davis, M., & Redford, M. A. (2023). Learning and change in a dual lexicon model of speech production. *Frontiers in Human Neuroscience*, *17*, Article 893785. https://www.frontiersin.org/articles/10.3389/fnhum.2023.893785

Dollaghan, C. A., Biber, M., & Campbell, T. F. (1993). Constituent syllable effects in a nonsense-word repetition task. *Journal of Speech, Language, and Hearing Research*, *36*(5), 1051–1054. https://doi.org/10.1044/jshr.3605.1051

Dollaghan, C. A., Biber, M. E., & Campbell, T. F. (1995). Lexical influences on nonword repetition. *Applied Psycholinguistics*, *16*(2), 211–222. https://doi.org/10.1017/S0142716400007098

Dollaghan, C. A., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

Dunn, L. M., Dunn, D. M., & Styles, B. (2009). *The British Picture Vocabulary Scale—Third Edition (BPVS 3)*. GL Assessment.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Paul H Brookes Publishing.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*(1), 83–94. https://doi.org/10.3758/bf03210559

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543. https://doi.org/10.1017/S0142716406060383

Gathercole, S. E., & Adams, A.-M. (1993). Phonological working memory in very young children. *Developmental Psychology*, *29*(4), 770–778. https://doi.org/10.1037/0012-1649.29.4.770

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*(2), 200–213. https://doi.org/10.1016/0749-596X(89)90044-2

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, *81*(4), 439–454. https://doi.org/10.1111/j.2044-8295.1990.tb02371.x

Gathercole, S. E., & Baddeley, A. D. (1996). *Children's test of nonword repetition*. The Psychological Corporation.

Gathercole, S. E., Service, E., Hitch, G. J., Adams, A.-M., & Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology*, *13*(1), 65–77. https://doi.org/10.1002/(SICI)1099-0720(199902)13:1⟨65::AID-ACP548⟩3.0.CO;2-O

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics*, *12*(3), 349–367. https://doi.org/10.1017/S0142716400009267

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology*, *28*(5), 889–898. https://doi.org/10.1037/0012-1649.28.5.887

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four … or is it two? *Memory*, *12*(6), 732–747. https://doi.org/10.1080/09658210344000530

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. https://doi.org/10.1016/s1364-6613(00)01662-4

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Grunwell, P. (1981). The development of phonology: A descriptive profile. *First Language*, *2*(6), 161–191. https://doi.org/10.1177/014272378100200601

Hoff, E., Core, C., & Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-month-olds. *Journal of Child Language*, *35*(4), 903–916. https://doi.org/10.1017/S0305000908008751

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term

memory span. *Journal of Memory and Language*, *30*(6), 685–701. https://doi.org/10.1016/0749-596X(91)90032-F

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, M., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217–1232. https://doi.org/10.1037//0278-7393.23.5.1217

Jessop, A., Pine, J., & Gobet, F. (2024). *Chunk-based incremental processing and learning: An integrated theory of word discovery, vocabulary growth, and speed of lexical processing*. PsyArXiv. https://doi.org/10.31234/osf.io/dukpt

Johnson, E. K., & White, K. S. (2019). Six questions in infant speech and language development. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior*. MIT Press.

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. https://doi.org/10.1016/j.cognition.2016.04.017

Jones, G., Cabiddu, F., Andrews, M., & Rowland, C. F. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, Article 104232. https://doi.org/10.1016/j.jml.2021.104232

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*(2), 298–310. https://doi.org/10.1111/desc.12111

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: A computational model of the learning of new words. *Developmental Science*, *10*(6), 853–873. https://doi.org/10.1111/j.1467-7687.2007.00638.x

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory & Cognition*, *46*(2), 216–229. https://doi.org/10.3758/s13421-017-0759-3

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. https://doi.org/10.1016/j.cogpsych.2017.07.002

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and frequency in specific language impairment: Accuracy and error data from two nonword repetition tests. *Journal of Speech, Language, and Hearing Research : JSLHR*, *53*(6), 1642–1655. https://doi.org/10.1044/1092-4388(2010/09-0222)

Keren-Portnoy, T., Vihman, M. M., DePaolis, R. A., Whitaker, C. J., & Williams, N. M. (2010). The role of vocal practice in constructing phonological working

memory. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1280–1293. https://doi.org/10.1044/1092-4388(2009/09-0003)

Krishnan, S. (2017). Fractionating nonword repetition: The contributions of short-term memory and oromotor praxis are different. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0178356

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54; discussion 55–74. https://doi.org/10.1037/0033-295x.109.1.35

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The Cappuccino model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*, 1619–1624.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*(3), 637–652. https://doi.org/10.1111/tops.12258

Meints, K., Fletcher, K., & Just, J. (2017). *The Lincoln Toddler Communicative Development Inventory—A UK Adaptation of the MacArthur-Bates Communicative Development Inventory: Words and Sentences (Toddler Form)*. https://cpb-eu-w2.wpmucdn.com/blogs.lincoln.ac.uk/dist/b/6736/files/2017/11/Lincoln_toddler_cdiv2-2.pdf

Messer, M. H., Verhagen, J., Boom, J., Mayo, A. Y., & Leseman, P. P. M. (2015). Growth of verbal short-term memory of nonwords varying in phonotactic probability: A longitudinal study with monolingual and bilingual children. *Journal of Memory and Language*, *84*, 24–36. https://doi.org/10.1016/j.jml.2015.05.001

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352. https://doi.org/10.1037/h0043158

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1033–1047. https://doi.org/10.1044/1092-4388(2005/072)

Newbury, J., Klee, T., Stokes, S. F., & Moran, C. (2016). Interrelationships between working memory, processing speed, and language development in the age range 2–4 years. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1146–1158. https://doi.org/10.1044/2016_JSLHR-L-15-0322

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Science*, *11*(3), 520–535. https://doi.org/10.1111/tops.12403

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

R Core Team (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. *Journal of Speech, Language and Hearing Research*, *47*(1), 223–234. https://doi.org/10.1044/1092-4388(2004/019)

Sanchez, A., Meylan, S., Braginsky, M., Macdonald, K., Yurovsky, D., & Frank, M. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*, 1928–1941. https://doi.org/10.3758/s13428-018-1176-7

Schwering, S. C., & MacDonald, M. C. (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in Human Neuroscience*, *14*, 68. https://doi.org/10.3389/fnhum.2020.00068

Snowling, M., Chiat, S., & Hulme, C. (1991). Words, nonwords, and phonological processes: Some comments on Gathercole, Willis, Emslie, and Baddeley. *Applied Psycholinguistics*, *12*(3), 369–373. https://doi.org/10.1017/S0142716400009279

Snowling, M., Goulandris, N., Bowlby, M., & Howell, P. (1986). Segmentation and speech perception in relation to reading skill: A developmental analysis. *Journal of Experimental Child Psychology*, *41*(3), 489–507. https://doi.org/10.1016/0022-0965(86)90006-8

Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *50*(4), 498–505. https://doi.org/10.1111/j.1469-7610.2008.01991.x

Stokes, S. F., Moran, C., & George, A. (2013). Nonword repetition and vocabulary use in toddlers. *Topics in Language Disorders*, *33*(3), 224. https://doi.org/10.1097/TLD.0b013e31829d038c

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition*, *179*, 23–36. https://doi.org/10.1016/j.cognition.2018.06.002

Tamburelli, M., Jones, G., Gobet, F., & Pine, J. M. (2012). Computational modelling of phonological acquisition: Simulating error patterns in Nonword Repetition Tasks. *Language and Cognitive Processes*, *27*(6), 901–46. https://doi.org/10.1080/01690965.2011.583510.

Torrington Eaton, C., Newman, R. S., Ratner, N. B., & Rowe, M. L. (2015). Non-word repetition in 2-year-olds: Replication of an adapted paradigm and a useful methodological extension. *Clinical Linguistics & Phonetics*, *29*(7), 523–535. https://doi.org/10.3109/02699206.2015.1029594

Verhagen, J., Boom, J., Mulder, H., de Bree, E., & Leseman, P. (2019). Reciprocal relationships between nonword repetition and vocabulary during the preschool years. *Developmental Psychology*, *55*(6), 1125–1137. https://doi.org/10.1037/dev0000702

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology*, *108*(1), 1–27. https://doi.org/10.1111/bjop.12207

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Details of Test Creation and Scoring.