Non-word repetition in Yélî Dnye

Alejandrina Cristia[1] & Marisa Casillas[2,3]

[1] Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University

[2] Max Planck Institute for Psycholinguistics

[3] University of Chicago

Author Note

Correspondence concerning this article should be addressed to Alejandrina Cristia, 29, rue d'Ulm, 75005 Paris, France. E-mail: alecristia@gmail.com

Abstract

In nonword repetition (NWR) studies, participants are presented auditorily with an item that is phonologically legal but lexically meaningless in their language, and asked to repeat this item as closely as possible. NWR scores are thought to reflect some aspects of phonological development, saliently a perception-production loop supporting flexible production patterns. In this study, we report on NWR results among children learning Yélî Dnye, an isolate spoken on Rossel Island in Papua New Guinea. Our overarching goal is to reflect on how NWR scores can be compared across participants, studies, languages, and populations, in order to shed light on the factors universally structuring variation in language development. More specifically, this study contributes to three lines of research. First, we contribute to investigations on NWR across diverse languages, by documenting that, in Yélî Dnye, non-word items containing typologically frequent sounds are repeated without changes more often that non-words containing typologically rare sounds, above and beyond any within-language frequency effects. Second, contributing to mounting research suggesting that length effects may be language- or population-specific, we find rather weak effects of item length. Third, we add a datapoint on potential sources of individual variation effects, by establishing that in our sample age has a strong effect on NWR scores, whereas there are weak correlations with gender, maternal education, and birth order. Together, these data provide a unique view of online phonological processing in an understudied language while making preliminary connections between language development and cross-linguistic features.

Keywords: phonology, non-word repetition, development

Word count: 9,000 words

Non-word repetition in Yélî Dnye

TODO Alex

- add missing refs
- correct rest of discussion

TODO Middy

- read over whole thing – does the logic sound ok?
- if you are motivated to, improve figures 1 and 5 (integrating our results to previous lit) – but I also think we can submit the paper with these figures as they are, and propose improvements in a revision

Introduction

Children's perception and production of phonetic and phonological units continues developing well beyond the first year of life, even extending into middle childhood (e.g., Hazan & Barrett, 2000). Much of the evidence for later phonological development comes from nonword repetition (NWR) tasks. In a NWR task, participants hear a short word-like form that is phonologically legal but lexically meaningless in the language(s) they are learning. After hearing this non-word, the participant's task is to try to immediately and precisely repeat it. NWR scores are thought to reflect long-term phonological knowledge (to perceive the item precisely despite not having heard it before) as well as online phonological working memory (to encode the item in the interval between hearing it and saying it back) and flexible production patterns (to produce the item precisely despite not having pronounced it before).

NWR has been used to seek answers to a variety of theoretical questions, including what the links are between phonology, working memory, and the lexicon (Bowey, 2001), and how extensively phonological constraints found in the lexicon affect online production (Gallagher, 2014). NWR is also frequently used in applied contexts, notably as a diagnostic tool for language delays and disorders (Estes, Evans, & Else-Quest, 2007). Since non-words can be generated in any language, it has attracted the attention of researchers working in multilingual and linguistically diverse environments, particularly in Europe (COST Action, 2009; Meir, Walters, & Armon-Lotem, 2016).

In the present study, we use NWR to investigate the phonological development of children learning Yélî Dnye, an isolate language spoken in Papua New Guinea (PNG) that has a large and unusually dense phonological inventory. The study was designed to contribute to four aspects of our understanding of phonological development.

First, we included a subset of non-word items with typologically rare and/or challenging sounds to ask whether these rare sounds are disadvantaged in the perception-production loop involved in NWR. Previous work using NWR has preferred relatively universal and early-acquired phonemes (with the possible exception of Gallagher, 2014), likely as a way to separate phoneme pronunciation from broader syllable structure and word-level prosodic effects (Gallon, Harris, & Van der Lely, 2007). Here, we investigate repetition of non-word items containing cross-linguistically common and cross-linguistically rare phonetic targets.

Second, we varied the length (in syllables) of non-words to contribute to growing research looking at the impact of word length on NWR repetition, and what this may reflect about phonological development. Our reading of previous NWR research is that there are variable effects of length between populations. For instance, Jaber-Awida (2018) reports an average of ~96% correct repetition for items 2 syllables long among children learning an Arabic variety of Israeli at about 5.5 years of age, but ~81% for items 3 syllables long. In contrast, Piazzalunga, Previtali, Pozzoli, Scarponi, and Schindler (2019) observe no decline in performance in similarly-aged Italian learners, with a score of 84% for 2 syllables versus 85% for 3 syllables. It is possible that differences are due to a host of variables, including the modal length of words in the language and/or in child-directed speech in that culture. That said, there could be other causal pathways: Research on adults suggests that illiterate Portuguese speakers repeat monosyllabic non-words just as accurately as literate adults, whereas scores are much lower among illiterate than literate speakers for items 3 or 4 syllables long (de Santos Loureiro et al., 2004). Given that both groups of adults speak the same language (Brazilian Portuguese), then perhaps differences in repetition accuracy reveal differences in how flexible the perception-production loop is. Given the paucity of evidence looking at this question, we do not approach this issue within a hypothesis-testing framework but sought instead to provide one more piece of data on the question, which may be re-used in future meta- or mega-analytic approaches.

Third, there are ongoing discussions as to what the key factors structuring individual variation are. Although the ideal systematic review is missing, a recent paper comes close with a rather extensive review of the literature looking at correlations between NWR scores and a variety of child-level variables (Farabolini, Rinaldi, Caselli, & Cristia, 2021). In a nutshell, most evidence is mixed, suggesting that consistent individual variation effects may be small, and more data is needed to estimate their true size. For this reason, we descriptively report association strength between NWR scores and child age, sex, birth order, and maternal education. Based on previous work, we looked at potential increases with age (Farmani et al., 2018; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014; Vance, Stackhouse, & Wells, 2005). Previous work typically finds no significant differences as a function of maternal education (e.g., Farmani et al., 2018; Balladares, Marshall, & Griffiths, 2016; Kalnak et al., 2014; Meir & Armon-Lotem, 2017) or child gender (Chiat & Roy, 2007). Although past research has not often investigated potential effects of birth order on NWR, there is a sizable literature on these effects in other language tasks (Havron et al., 2019), and therefore we report on these too.

Fourth, these data contribute to the small literature using this task with non-Western, non-urban populations, speaking a language with a moderate to large phonological inventory (see

Maddieson, 2005 for a broad classification of languages based on inventory size). Indeed, NWR has seldom been used outside of Europe and North America (with exceptions including Gallagher, 2014; Cristia, Farabolini, Scaff, Havron, & Stieglitz, 2020), and/or outside urban settings (except for in Cristia et al., 2020), nor with languages having large phonological inventories [e.g., more than 34 consonants and 7 vowel qualities Maddieson (2013b);Maddieson (2013a); no exceptions to our knowledge]. There are no theoretical reasons to presume that the technique will not generalize to these new conditions. That said, Cristia et al. (2020) recently reported relatively lower NWR scores among the Tsimane', a non-Western rural population, interpreting these findings as consistent with the hypothesis that lower levels of infant-directed speech and/or low prevalence of literacy in a population could lead to population-level differences in NWR scores. In view of these results, it is important to bear in mind that NWR is a task developed in countries where literacy is widespread, and it is considered an excellent predictor of reading, for instance better than rhyme awareness (e.g., Gathercole, Willis, & Baddeley, 1991). Therefore, it may not be a general index of phonological development, but more specifically reflect certain non-universal skills. Indeed, Cristia et al. (2020) present the task as being a good index of the development of "short-hand-like" representations specifically, which could thus miss, for example, more holistic phonological and phonetic representations. To our knowledge, there is little discussion of linguistic effects – i.e., of potential differences in NWR as a function of language typology – or cultural effects – i.e., of potential differences in NWR as a function of other differences across human populations, aside from Cristia et al. (2020)'s hypotheses just mentioned. Regarding potential language differences, we note that the very fact that studies compose items by varying syllable structure and word length, while prefering relatively simple and universal phones (notably relying on point vowels, simple plosives, and fricatives that are prevalent across languages, like /s/) may indicate a bias towards Indo-European languages, where syllable structure and word length are indeed important structural dimensions. This bias is, of course, implicit and unintentional, arising as researchers working in other languages attempt to build items that conform to the descriptions of the first people using the method, who tend to work on English. And it does occur that some researchers opt instead to employ dimensions of variation that are more relevant to their language, such as adaptations in Chinese languages that have items varying in tone REF. return to this after reading lit

Before going into the details of our study design we first give an overview of Yélî Dnye phonology as well as a brief ethnographic review of the developmental environment on Rossel Island. As discussed above, NWR has been almost exclusively used in urban, industrialized populations, so we provide this additional ethographic information to contextualize the adaptations we have made in running the task and collecting the data, compared to what is typical in commonly studied sites, which are typically easily accessible. Laying 250 nautical miles off the coast of mainland PNG and surrounded by a barrier reef, transport to and from Rossel Island is both infrequent and irregular. International phone calls and digital exchanges that require significant data transfer are typically not an option. Data collection is therefore typically limited to the duration of the researchers' on-island visits.

Yélî Dnye phonology. Yélî Dnye is an isolate language (presumed Papuan) spoken by approximately 7,000 people residing on Rossel Island, an island found at the far end of the Louisiade Archipelago in Milne Bay Province, Papua New Guinea. The Yélî sound system, much like its baroque grammatical system (Levinson, 2020), is unlike any other in the region. In total, Yélî Dnye uses 90 distinctive segments (not including an additional three rarely used consonants), far outstripping the phonemic inventory size of other documented Papuan languages (Foley, 1986; Levinson, 2020; Maddieson & Levinson, n.d.). Thus, with respect to our first research goal, Yélî Dnye seemed is a good language to attempt an investigation on NWR with sounds varying in cross-linguistic frequency because of its large inventory, which includes some rare sounds.

To provide some qualitative information on this inventory, we add the following observations. With only four primary places of articulation (bilabial, alveolar, post-alveolar, and velar) and no voicing contrasts, the phonological inventory is remarkably packed with acoustically similar segments. The core oral stop system includes both singleton (/p/, /t/, /ṭ/, and /k/) and doubly-articulated (/tp/, /ṭp/, /kp/) segments, with full nasal equivalents (/m/, /n/, /ṇ/, /ŋ/, /nm/, /ṇm/,

/ŋm/), and with a substantial portion of them contrastively pre-nasalized or nasally released (/mp/, /nt/, /ṇṭ/, /ŋk/, /nmtp/, /ṇmṭp/, /ŋmkp/, /ṭṇ/, /kŋ/, /ṭpṇm/, /kpŋm/). A large number of this combinatorial set can further be contrastively labialized, palatalized on release, or both (e.g., /pʲ/, /pʷ/, /pʲʷ/; /tpʲ/; /ṇmḍbʲ/; see Levinson (2020) for details).[1] The consonantal inventory also includes a number of non-nasal continuants (/w/, /j/, /ɣ/, /l/, /βʲ/, /lʲ/, /lβʲ/). Vowels in Yélî Dnye may be oral or nasal, short or long. The 10 oral vowel qualities, which span four levels of vowel height, (/i/, /ɯ/, /u/, /e/, /o/, /ə/, /ɛ/, /ɔ/, /æ/, /ɑ/) can be produced as short and long vowels, with seven of these able to appear as short and long nasal vowels as well /ĩ/, /ũ/, /ə̃/, /ɛ̃/, /ɔ̃/, /æ̃/, /ɑ̃/).

Regarding our second research goal, on the effect of non-word length on NWR, most Yélî Dnye words are bisyllabic (~50%), with monosyllabic words (~40%) appearing most commonly after that, and with tri-and-above syllabic words appearing least frequently (~10%; based on $>$5800 lexemes in the most recent dictionary at the time of writing; Levinson, 2020). The vast majority of syllables use a CV format. A small portion of the lexicon features words with a final CVC syllable, but these are limited to codas of -/m/, -/p/, or -/j/ (e.g., "ndap" /ṇṭæp/ Spondylus shell) and are often resyllabified with an epenthetic /ɯ/ in spontaneous speech (e.g., "ndapî" /'ṇṭæ.pɯ/). There are also a handful of words starting with /æ/ (e.g., "ala" /æ.'læ/ here) and a small collection of single-vowel grammatical morphemes (see Levinson (2020) for details).

Our knowledge of Yélî language development is growing (e.g., Brown, 2011, 2014; Brown & Casillas, n.d.; Casillas, Brown, & Levinson, 2020; Liszkowski, Brown, Callaghan, Takada, & de Vos, 2012), but research into Yélî phonological development has only just begun (e.g., Peute, Fikkert, & Casillas, n.d.). We hope the present study contributes to filling this gap. TODO incorporate brief summary of paper

[1] We use Levinson's (2020) under-dot notation (e.g., /ṭ/) to denote the post-alveolar place of articulation; these stops are, articulatorily, somewhat variable in place, with at least some tokens produced fully sub-apically. In approximating cross-linguistic segment frequency below we use the corresponding retroflex for each stop segment (e.g., /ʈ/, /ʈp/, /ɳ/).

The Yélî community. Some aspects of the community are relevant for interpreting results found when addressing our thir research question, regarding sources of individual variation. Specifically, we investigated potential effects of age, gender, maternal education, and birth order. There is nothing particular to note regarding age and gender, but we have some comments that pertain to the other two factors.

The typical household in our dataset includes seven individuals (typically, a mixed sex couple and children – their own and billeting others, as discussed in the next paragraph) and is situated among a collection of four or more other households, with structures often arranged around an open grassy area. These household clusters are organized by patrilocal relation, such that they typically comprise a set of brothers, their wives and children, and their mother and father, with neighboring hamlets also typically related through the patriline. Land attribution for building one's home is decided collectively based on land availability, and typically does not take into consideration an individual's desire to be close to a school.

Most Yélî parents are swidden horticulturalists, and those who are not may not reside in the island. Within a group of households, it is often the case that most older adolescent and adults spend their day tending to their gardens (which may not be nearby), bringing up water from the river, washing clothes, preparing food, and engaging in other such activities, which leave them little time to spend directly with the children in their household (other than infants). Starting around age two years, children more often spend large swaths of their day playing, swimming, and foraging for fruit, nuts, and shellfish in large (~10 members) independent and mixed-age child play groups (Brown & Casillas, n.d.; Casillas et al., 2020). Formal education is a priority for Yélî families, and many young parents have themselves pursued additional education beyond of what is locally available (Casillas et al., 2020). Local schools are well out of walking distance for many children (i.e., more than 1 hour on foot or by canoe each day), so it is very common for households situated close to a school to billet their school-aged relatives during the weekdays for long segments of the school year. Children start school often at around age six, although the precise age depends on the child's apparent development.

Some general ideas regarding potential maternal education effects on our data may be drawn from the observations above. To begin with, many of our participants above 6 years of age may not be living with their birth mother but with other relatives, which may weaken maternal education effects. Additionally, the importance given to formal education appears relatively stable over the period that Rossel Island has been visited by language researchers (Steven Levinson and Penelope Brown, about 20 years). Together with the fact that land attribution is essentially random with respect to educational hopes, it seems to us that the length of formal education a given individual may have is not necessarily a good index of their socio-economic status or other individual properties, unlike what happens in industrialized sites, and variation may simply due to random factors like living close to a school or having relatives there.

As for birth order, much of the work on birth order effects on cognitive development (including language) has been carried out in the last 70 years and in agrarian or industrialized settings (Barclay, 2015; Grätz, 2018), where nuclear families are more likely to be the prevalent rearing environment (Lancy, 2015). It is possible that birth order effects are stronger in such a setting, because much of the stimulation can only come from the parents, and when there are multiple children, the inter-birth interval is small enough that older siblings may not be of an age that allows them to contribute to their younger siblings' stimulation. This contrasts with this picture just drawn in the Yélî community, where children regardless of their birth order in their nuclear family will typically benefit from a rich and extensive socially stimulating setting, surrounded by siblings, and cousins of several orders.

We add some observations that will help us integrate this study to the broader investigation of NWR across cultures. As mentioned previously, there is one report of lower NWR scores among the Tsimane', which the authors interpret as consistent with long-term effects of low levels of infant-directed speech (Cristia et al., 2020). However, Cristia et al. (2020) also point out that this is based on between-paper comparisons, and thus methods and a myriad other factors have not been controlled for. The Yélî community can help us bring further light into this question because direct speech to children under 3;0 is relatively infrequent in this community too (Casillas et al., 2020). Although infant-directed speech has been measured in different ways among the Tsimane' and the Yélî communities, our most comparable estimates at present suggests that Tsimane' young children are spoken to about 4.2 minutes per hour (Scaff, Stieglitz, Casillas, & Cristia, 2021), and Yélî children about 3.7 minutes per hour (Casillas et al., 2020). Thus, if input quantities in early childhood are a major determinant of NWR scores, we should observe similarly low NWR scores as in Cristia et al. (2020).

NWR design and analysis adaptations. In a basic NWR task, the participant listens to a production of a word-like form, such as /bilik/, and then repeats back what they heard without changing any phonological feature that is contrastive in the language. For instance, in English, a response of [bilig] or [pilik] would be scored as incorrect; a response [bi:lik], where the vowel is lengthened without change of quality would be scored as correct, because English does not have contrastive vowel length. There is some variation in how past NWR studies have designed the presentation procedure and structure of items. For example, while items are often presented orally by the experimenter (Torrington Eaton, Newman, Ratner, & Rowe, 2015), an increasing number of studies have turned instead to playing back pre-recorded stimuli in order to increase control in stimulus presentation (Brandeker & Thordardottir, 2015). Additionally, while some studies have used 10-15 non-words (e.g., Cristia et al., 2020), others have employed up to 46 unique items (Piazzalunga et al., 2019). Authors also often modulate structural complexity, typically measured in terms of item length (measured in number of syllables) and/or syllable structure (open as opposed to closed syllables, Gallon et al., 2007).

Previous work typically steers clear of articulatorily and/or acoustically challenging sounds, but we included some in our experiment to more adequately represent Yélî Dnye's phonology and to contribute data on whether this affects repetition. We ultimately used a relatively large number of items that would enable us to explore both variation in structural complexity and in more vs. less challenging sounds. However, aware that this large item inventory might render the task longer and more tiresome, we split items across children (see below). Naturally, designing the task in this way may make the study of individual variation within the population more difficult because different children are exposed to different items. However, as discussed above, effects of individual differences in NWR are probably relatively small, and thus we reasoned that they would not be detectable with the sample size that we could collect during our short visit. That said, we contribute to the literature by also reporting descriptive analyses of individual variation that could potentially be integrated in meta- or mega-analytic efforts.

Research questions. After some preliminary analyses to set the stage, we perform statistical analyses to inform answers to the following questions:

- Does the cross-linguistic frequency of sounds in the stimuli predict NWR scores? Are rarer sounds more often substituted by commoner sounds?
- How do NWR scores change as a function of item length in number of syllables?
- Is individual variation in NWR scores attributable to child age, sex, birth order, and/or maternal education?

Throughout these analyses and in the Discussion, we will also have in mind our fourth goal, namely integrating NWR results across samples varying in language and culture.

We had considered boosting the interpretational value of this evidence by announcing our analysis plans prior to conducting them. However, we realized that even pre-registering an analysis would be equivocal because we would not have enough power to look at all relationships of interest, in many cases possibly not enough to detect any of the known effects, given the previously discussed variability across studies. Therefore, all analyses in the present study are descriptive and should be considered exploratory.

Methods

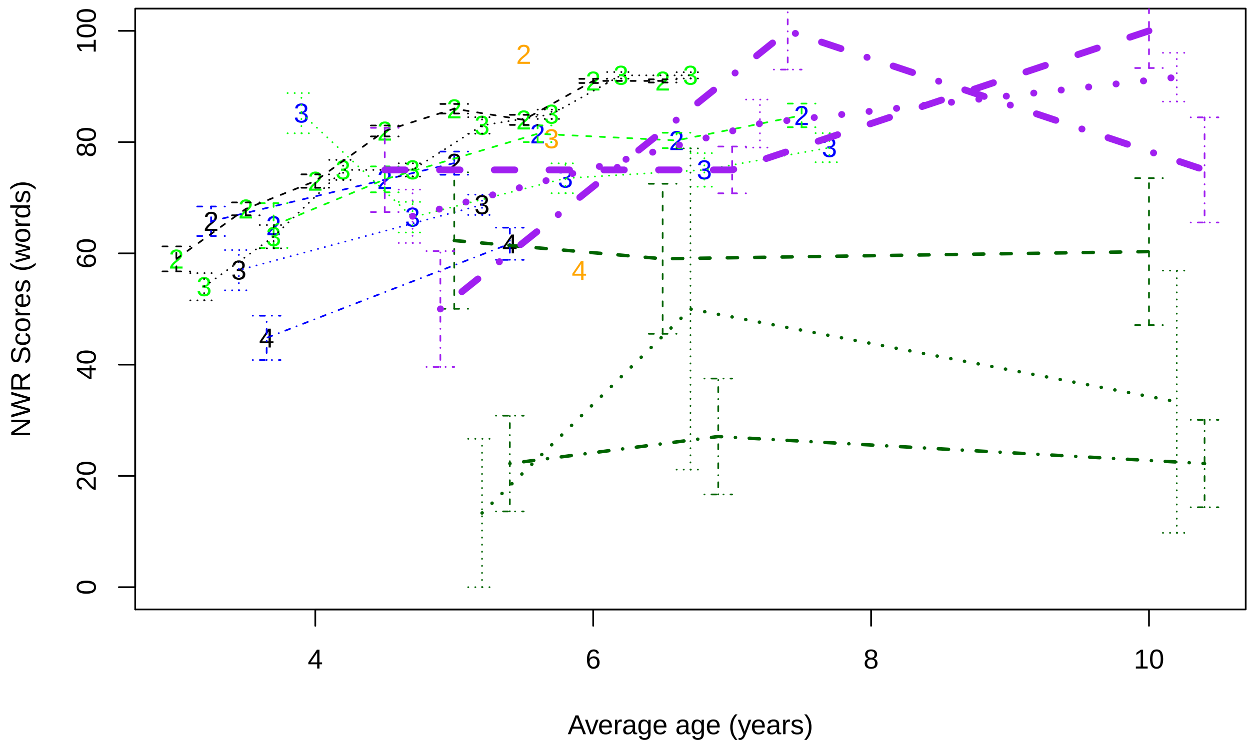

Figure 1. NWR scores as a function of age (in years) and item length for comparable studies (2-4 indicating number of syllables, 2 = dashed, 3 = dotted, 4 = dotted and dashed). Jaber-Awida (2018) reported on 20 Arabic learners (orange); Piazzalunga et al. (2019) reported on groups of 24-60 Italian learners (black); Stokes et al. (2006) on 15 Cantonese learners (blue); Vance et al. (2005) on 17-20 English learners (light green); Cristia et al. (2020) reported on groups of 4-6 Tsimane' learners (dark green); the present study reports on groups of 8-19 Yélî Dnye learners (purple). Central tendency is mean in Arabic and Tsimane', median in Italian and Yélî Dnye; error is one standard error. Age has been slightly jittered for ease of inspection of different lengths at a given age.

Stimuli. Many NWR studies are based on a fixed list of 12-16 items that vary in length between 1 and 4 syllables, often additionally varying syllable complexity and/or cluster presence and complexity, and always meeting the condition that they do not mean anything in the target language (e.g., Balladares et al., 2016; Wilsenach, 2013). We kept the same variation in item length and requirement for not being meaningful in the language, but we did not vary syllable complexity or clusters because these are vanishingly rare in Yélî Dnye. We also increased the number of items an individual child would be tested on, such that a child would get up to 23 items to repeat (other work has also used up to 24-30 items: Jaber-Awida, 2018; Kalnak et al., 2014), with the entire test inventory of 40 final items distributed across children.

A first list of candidate items was generated during a trip to the island in 2018 by selecting simple consonants (/p/, /t/, /ṭ/, /k/, /m/, /n/, /w/, /y/) and vowels (/i/, /o/, /u/, /a/, /e/) and combining them into consonant-vowel syllables, then sampling the space of 2- to 4-syllable sequences. These candidates were automatically removed from consideration if they appeared in Levinson's (2015) dictionary. The second author presented them orally to three local research assistants, all native speakers of Yélî Dnye, who repeated each form as they would in an NWR task and additinally let the experimenter know if the item was in fact a word or phrase in Yélî Dnye. Any item reported to have a meaning or a strong association with another word form or meaning was excluded.

A second list of candidate items was generated in a second trip to the island in 2019, when data were collected, by selecting complex consonants and systematically crossing them with all the vowels in the Yélî Dnye inventory to produce consonant-vowel monosyllabic forms. As before, items were automatically excluded if they appeared in the dictionary. Additionally, since perceiving vowel length in isolated monosyllables is challenging, any item that had a short/long lexical neighbor was excluded. Because there is still much to discover about the phonology and phonetics of Yélî Dnye (Levinson, 2020), it was also possible that we initially generated items with illegal, but currently undocumented constraints. Therefore, we made sure that the precise consonant-vowel sequence occurred in some real word in the dictionary (i.e., that there was a longer word included the monosyllable as a subsequence). These candidates were then presented to one informant, for a final check that they did not mean anything. Together with the 2018 selection, they were recorded, based on their orthographic forms, using a Shure SM10A XLR dynamic headband microphone and an Olympus WS-832 stereo audio recorder (using an XLR to mini-jack adapter) by the same informant, monitored by the second author for clear production of the phonological target. The complete recorded list was finally presented to two more informants, who were able to repeat all the items and who confirmed there were no real words present. Despite these checks, one monosyllable was ultimately frequently identified as a real word in the resulting data (intended "yî" /yɯ/; identified as "yi" /yi/, tree). Additionally, an error was made when preparing files for annotation, resulting in two items being merged ("tpâ" /tpɑ/ and "tp:a" /tpæ̃/). These three problematic items are not described here, and removed from the analyses below.

The final list includes three practice items and 40 test items (across infants): 16 monoysllables containing sounds that are less frequent in the world's languages than singleton plosives; 8 bisyllables; 12 trisyllables; and 4 quadrisyllables (see Table 1).

A Praat script (Boersma & Weenink, 2020) was written to randomize this list 20 times, and split it into two sublists, to generate 40 different elicitation sets. The 40 elicitation sets are available online from osf.io/dtxue/. The split had the following constraints:

- The same three items were selected as practice items and used in all 40 elicitation sets.
- Splits were done within each length group from the 2018 items (i.e., separately for 2-, 3-, and 4-syllable items); and among onset groups for the difficult monosyllables generated in 2019 (i.e., all the monosyllables starting with /tp/ were split into 2 sublists). Since some of these groups had an odd number of items, one of the sublists was slightly longer than the other (20 vs. 23).
- Once the sublist split had been done, items were randomized such that all children heard first the 3 practice items in a fixed order (1, 2, and 4 syllables), a randomized version of their sublist selection of difficult onset items, and randomized versions of their 2-syllable, then

3-syllable, and finally 4-syllable items.

To inform our analyses, we estimated the typological frequency of all phonological segments present in the target items using the PHOIBLE cross-linguistic phonological inventory database (Moran & McCloy, 2019). For each phone in our task, we extracted the number and percentage of languages noted to have that phone in its inventory. While PHOIBLE is an unprecedentedly comprehensive database, with phonological inventory data for over 2000 languages at the time of writing, it is of course still far from complete, which may mean that frequencies are estimates rather than precise descriptors). Note that nearly half of the segment types are only attested in one language (Steven Moran, personal communication). Extrapolating from this observation, we treat the three segments in our stimuli that were unattested in PHOIBLE (/lβʲ/, /ṭp/, and /tp/) as having a frequency of 1 (i.e., appearing in one language), with a (rounded) percentile of 0% (i.e., its cross-linguistic percentile is zero).

Additionally, we estimated frequency of the phones present in the target items in a corpus of child-centered recordings (Casillas et al., 2020) by counting the number of word types in which they occurred, and applied the natural logarithm.[2] Here, unattested sounds were not considered (i.e., they were declared NA so that they do not count for analyses).

Procedure. In adapting the typical NWR procedure for this context, we balanced three desiderata: That children would not be unduly exposed to the items before they themselves had to repeat them (i.e., from other children who had participated); that children would feel comfortable doing this task with us; and that community members would feel comfortable having their children do this task with us.

We tested in four different sites spread across the northeastern region of the island, making a single visit to each, conducting back-to-back testing of all eligible children present at the time of

[2]We also carried out analyses using token (rather than type) phone frequency, but this measure was not correlated with whole-item NWR scores, and therefore the fact that it did not explain away the predictive value of cross-linguistic phone frequency was less informative than the relationship discussed in the Results section.

our visit in order to prevent the items from “spreading” between children through hearsay. Whenever children living in the same household were tested, we tried to test children in age order, from oldest to youngest, to minimize intimidation for younger household members, and always using different elicitation sets. Because space availability was limited in different ways from hamlet to hamlet, the places where elicitation happened varied across testing sites. More information is available from the online supplementary materials.

We fitted the child with a headset microphone (Shure SM10A or WH20 XLR with a dynamic microphone on a headband, most children using the former) that fed into the left channel of a Tascam DR40x digital audio recorder. The headsets were designed for adult use and could not be comfortably seated on many children’s heads without a more involved adjustment period. To minimize adjustment time, which was uncomfortable for some children given the proximity of the experimenter and equipment, we placed the headband on children’s shoulders in these cases, carefully adjusting the microphone’s placement so that it was still close to the child’s mouth. A research assistant who spoke Yélî Dnye natively sat next to the child throughout the task to provide instructions and, if needed, encouragement. The research assistant coached the child throughout the task to make sure that they understood what they were expected to do. An experimenter (the first author) delivered the pre-recorded stimuli to the research assistant and the child over headphones.

The first phase of the experiment involved making sure the child understood the task. We explained the task and then orally presented the first practice item. At this point, many children did not say anything in response, which triggered the following procedure: First, the assistant insisted the child make a response. If the child still did not say anything, the assistant said a real word and then asked the child to repeat it, then another and another. If the child could repeat real words correctly, we provided the first training item over headphones again for children to repeat. Most children successfully started repeating the items at this point, but a few needed further help. In this case, the assistant modeled the behavior (i.e., the child and assistant would hear the item again, and the assistant would repeat it; then we would play the item again and ask the child to repeat it). A small minority of children still failed to repeat the item at this point. If so, we tried again with the second training item, at which point some children demonstrated task understanding and could continue. A fraction of the remaining children, however, failed to repeat this second training item, as well as the third one, in which case we stopped testing altogether (see Participants section for exclusions).

The second phase of the experiment involved going over the list of test items randomly assigned to each child. This was done in the same manner as the practice items: the stimulus was played over the headphones, and then the child repeated it aloud. NWR studies vary in whether children are allowed to hear and/or repeat the item more than one time. We had a fixed procedure for the test items (i.e., the non-practice items) in which the child was allowed to make further attempts if their first attempt was judged erroneous in some way by the assistant. The procedure worked as follows: When the child made an attempt, the assistant indicated to the experimenter whether the child's production was correct or not. If correct, the experimenter would whisper this note of correct repetition into a separate headset that fed into the right channel of the same Tascam recorder and we moved on to the next item. If not, the child was allowed to try again, with up to five attempts allowed before moving on to the next item. Children were not asked to make repetitions if they did not produce a first attempt. In total, test sessions took approximately six minutes, with the first minute attributed to practice and five minutes to the actual test list.

Coding. The first author then annotated the onset and offset of all children's productions from the audio recording using Praat audio annotation software (Boersma & Weenink, 2020), then ran a script to extract these tokens, pairing them with their original auditory target stimulus, and writing these audio pairs out to .wav clips. The assistant then listened through all these paired target-repetition clips randomized across children and repetitions, grouped such that all the clips of the same target were listened to in succession. For each clip, the assistant indicated in a notebook whether the child production was a correct or incorrect repetition and orthographically transcribed the production, noting when the child uttered a recognizable word or phrase and adding the translation equivalent of that word/phrase into English. The assistant was also provided with some general examples of the types of errors children made without making specific reference to Yélî sounds or the items in the elicitation sets.

Analyses. Previous work typically reports two scores: a binary word-level exact repetition score, and a phoneme-level score, defined as the number of phonemes that can be aligned across the target and attempt, divided by the number of phonemes of whichever item was longer (the target or the attempt; as in Cristia et al., 2020). Previous work does not use distance metrics, but we report these rather than the phoneme-level scores because they are more informative. To illustrate these scores, recall our example of an English target being /bilik/ with an imagined response [bilig]. We would score this response as follows: at the whole item level this production would receive a score of zero (because the repetition is not exact); at the phoneme level this production would receive a score of 80% (4 out of 5 phonemes repeated exactly); and the phone-based Levenshtein distance for this production is 20% (because 20% of phonemes were substituted or deleted). Notice that the phone-based Levenshtein distance is the complement of the phoneme-level NWR score. An advantage of using phone-based Levenshtein distance is that it is scored automatically with a script, and it can then easily be split in terms of deletions and substitutions (insertions were not attested in this study).

Participants. This study was approved as part of a larger research effort by the second author. The line of research was evaluated by the Radboud University Faculty of Social Sciences Ethics Committee (Ethiek Commissie van de faculteit der Sociale Wetenschappen; ECSW) in Nijmegen, The Netherlands (original request: ECSW2017-3001-474 Manko-Rowland; amendment: ECSW-2018-041). As discussed in subsection “The Yélî community”, the combination of collective child guardianship practices and common billeting of school-aged children for them to attend school is that adult consent often comes from a combination of aunts, uncles, adult cousins, and grandparents standing in for the child’s biological parents. Child assent is also culturally pertinent, as independence is encouraged and respected from toddlerhood (Brown & Casillas, n.d.).

Participation was voluntary; children were invited to participate following indication of approval from an adult caregiver. Regardless of whether they completed the task, children were given a small snack as compensation. Children who showed initial interest but then decided not to participate were also given the snack.

We tested a total of 55 children from 38 families spread across four hamlets. We excluded test sessions from analysis for the following reasons: refused participation or failure to repeat items presented over headphones even after coaching (N = 8), spoke too softly to allow offline coding (N = 5), or were 13 years old or older (N = 2; we tested these teenagers to put younger children at ease). The remaining 40 children (14 girls) were aged from 3 to 10 years (M = 6.50 years, SD = 1.50 years). In terms of birth order, 6 were first borns, 5 second, 2 third, 7 forth, 5 fifth, and 1 sixth, with birth order missing for 14 children. These children were tested in a remote hamlet, and we unfortunately did not ask about birth order before leaving the site. Maternal years of education averaged 8.22 years (range 6-12 years).[3] We also note that there were 34 only exposed to Yélî Dnye at home, 6 children exposed to Yélî Dnye plus one or more other languages at home.[4]

Results

Preliminary analyses. We first checked whether whole-item NWR scores varied between first and subsequent presentations of an item by averaging word-level scores at the participant level separately for first attempts and subsequent repetitions. We excluded 1 child who did not have data

[3] We asked for mothers' highest completed level of education. We then record the number of years entailed by having completed that level under ideal conditions.

[4] Most speakers of Yélî Dnye grow up speaking it monolingually until they begin attending school around the age of 7 years; school instruction is in English. While monolingual Yélî Dnye upbringing is common, multilingual families are not unusual, particularly in the region around the Catholic Mission—the same region in which the current data were collected—where there is a higher incidence of married-in mothers from other islands (Brown & Casillas, n.d.). Children in these multilingual families grow up speaking Yélî Dnye plus English, Tok Pisin, and/or other language(s) from the region.

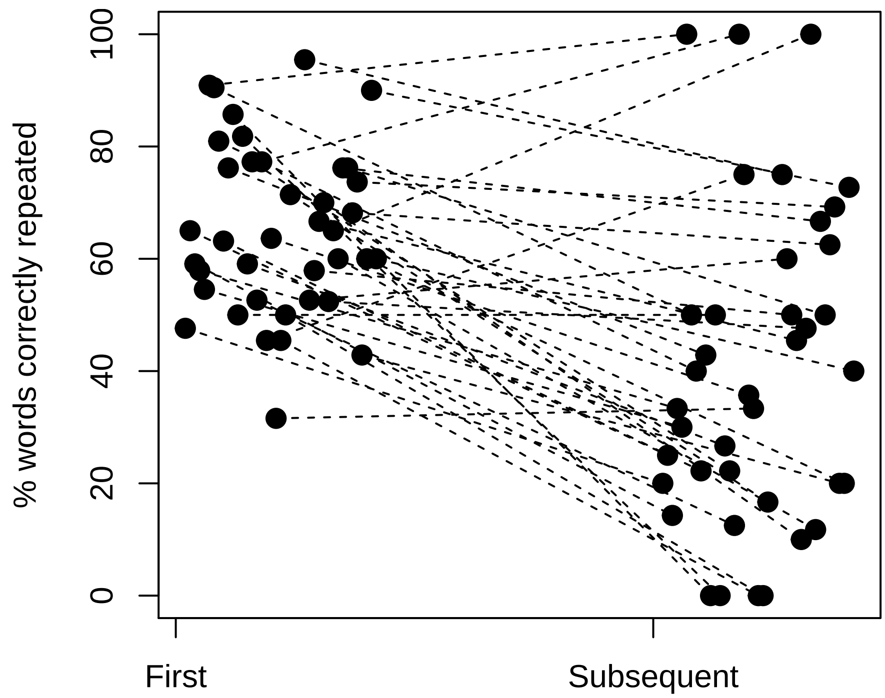

Figure 2. Whole-item NWR scores for individual participants averaging separately their first attempts and all other attempts.

for one of these two types. As shown in Figure 2, participants' mean word-level scores became more heterogeneous in subsequent repetitions. Surprisingly, whole-item NWR scores for subsequent repetitions ($M = 40$, $SD = 28$) were on average lower than first ones ($M = 65$, $SD = 15$), $t(38) = 5.89$, $p < 0.001$; Cohen's $d = 1.13$). Given uncertainty in whether previous work used first or all repetitions, and given that score here declined and became more heterogeneous in subsequent repetitions, we focus the remainder of our analyses only on first repetitions, with the exception of qualitative analyses of substitutions.

Taking into account only the first attempts, we derived overall averages across all items. The overall NWR score was $M = 65\%$ ($SD = 15\%$), Cohen's $d = 4.39$. The phoneme-based normalized Levenshtein distance was $M = 21\%$ ($SD = 9\%$), meaning that about a fifth of phonemes were substituted, inserted, or deleted.

We also looked into the frequency with which mispronunciations resulted in real words. In fact, two thirds of incorrect repetitions were recognizable as real words or phrases in Yélî Dnye or English: 63%. This type of analysis is seldom reported. We could only find one comparison point: Castro-Caldas, Petersson, Reis, Stone-Elander, and Ingvar (1998) found that illiterate European Portuguese adults' NWR mispronunciations resulted in real words in 11.16% of cases, whereas literate participants did so in only 1.71% of cases. The percentage we observe here is much higher than reported in Castro and colleagues' study, but we do not know whether age, language, test structure, or some other factor explains this difference.

NWR as a function of cross-linguistic phone frequency. Turning to our first research question, we analyzed variation in whole-item NWR scores as a function of the average frequency with which sounds composing individual target words are found in languages over the world. To look at this, we fit a mixed logistic regression in which the outcome variable was whether the non-word was correctly repeated or not. The fixed effect of interest was the average cross-linguistic phone frequency; we also included child age as a control fixed effect, and allowed slopes to vary over the random effects child ID and target ID.

We could include 826 observations, from 40 children producing in any given trial one of 40 potential target words. The analysis revealed a main effect of age (ß = 0.35, SE ß = 0.13, $p < 0.01$); and a significant estimate for the scaled average cross-linguistic frequency of phones in the target words (ß = 0.78, SE ß = 0.19, $p < 0.001$): Target words with phones found more frequently across languages had higher correct repetition scores, as shown in Figure 3. Averaging across participants, the Pearson correlation between scaled average cross-linguistic phone frequency and whole-item NWR scores was $r(38) = 0.54$.

We next checked whether the association between whole-item NWR scores and cross-linguistic phone frequence could actually be due to frequency of the sounds within the language: One can suppose that sounds that occur more frequently across languages are also more frequent within a language, and therefore may be easier for children to represent and repeat because of the additional exposure. Phone corpus-based frequencies were correlated with phone

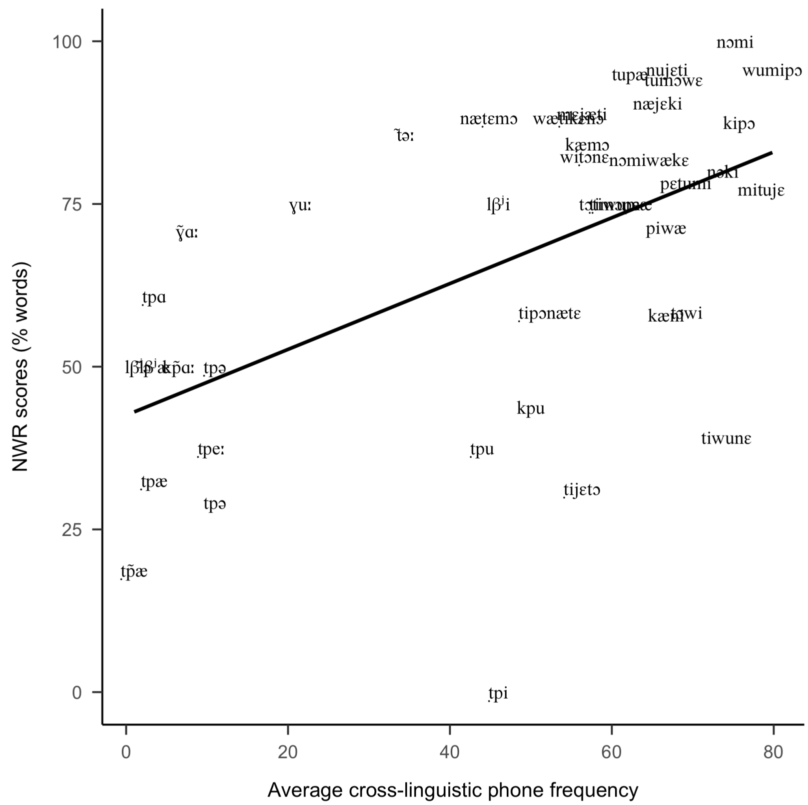

Figure 3. NWR scores for individual target words as a function of the average frequency with which each phone is found across languages.

cross-linguistic frequencies [r(27) = 0.50, p < 0.01]; and item-level average phone corpus-based frequencies were correlated with the corresponding cross-linguistic frequencies [r(38) = 0.73, p < 0.001]. Moreover, averaging across participants, the Pearson correlation between scaled average corpus phone frequency and whole-item NWR scores was r(38) = 0.43, p < 0.01. Therefore, we fit another mixed logistic regression, this time declaring as fixed effects both scaled cross-linguistic and corpus frequencies (averaged across all attested phones within each stimulus item), in addition to age. As before, the model contained random slopes for both child ID and target. In this model, both cross-linguistic phone frequency (ß = 0.78, SE ß = 0.27, p < 0.01) and age (ß = 0.35, SE ß = 0.13, p < 0.01) were significant predictors of whole-item NWR scores, but corpus phone frequency (ß = 0.00, SE ß = 0.25, p = 0.99) was not.

Patterns in NWR mispronunciations. We addressed our first research question in a second way, by investigating patterns of error, looking at all attempts so as to base our generalizations on more data. There were no cases of insertion, and deletions were very rare: there were only 12 instances of deleted vowels (~0.28% of all vowel targets), and 6 instances of deleted consonants (~0.19% of all consonant targets). We therefore focus our qualitative description here on substitutions: There were 820 cases of substitutions, ~16.95 of the 4839 phones found collapsing across all children and target words, so that substitutions constituted the frank majority of incorrect phones (~97.74 of unmatched phones). To inform our understanding of how cross-linguistic patterns may be reflected in NWR scores, we asked: Is it the case that cross-linguistically less common and/or more complex phones are more frequently mispronounced, and more frequently substituted by more common ones than vice versa?

We looked for potential asymmetries in errors for different types of sounds in vowels by looking at the proportion of vowel phones that were correctly repeated or not separately for nasal and oral vowels. The nasal vowels in our stimuli occur in ~1.40% of languages' phonologies (range 0% to 3%); whereas oral vowels in our stimuli occur in ~31.55% of languages' phonologies (range 3% to 92%). As noted above, type frequency within the language is correlated with cross-linguistic frequency, and thus these two types of sounds also differ in the former: Their type frequencies in Yélî Dnye are: nasal vowels ~0.03‰ (range 0.00‰ to 0.05‰) versus oral ~0.23‰ (range 0.02‰ to 0.76‰).

We distinguished errors that included a change of nasality (and may or may not have preserved quality), versus those that preserved nasality (and were therefore a quality error), shown in Table 2. We found that errors involving nasal vowel targets were more common than those involving oral vowels (35.90 versus 11.90). Additionally, errors in which a nasal vowel lost its nasal character were 10 times more common than those in which an oral vowel was produced as a nasal one. Note that this analysis does not tell us whether cross-linguistic or within-language frequency is the best predictor, an issue to which we return below.

For consonants, we inspected complex ([ṭp], [tp], [kp], [km], [kṇ], [mp], [ɣ], and [lβʲ]) versus simpler ones ([m], [n], [l], [w], [j], [w], [ṭ], [g], [p], [t], [k], [f], [h], and [tʃ]), using the same logic: We looked at correct phone repetition, substitution with a change in complexity category, or a change within the same complexity category.[5] The complex consonants in our stimuli occur in ~17.33% of languages' phonologies (range 0% to 78%); whereas simple consonants in our stimuli occur in ~67.62% of languages' phonologies (range 13% to 96%). Again these groups of sounds differ in their frequency within the language. Their type frequencies in Yélî Dnye are: complex consonants ~0.04‰ (range 0.00‰ to 0.10‰) versus simple consonants ~0.32‰ (range 0.06‰ to 0.55‰).

Table 3 showed that errors involving complex consonants targets were more common than those involving oral vowels (50.90 versus 8.20). Additionally, errors in which a complex consonant was mispronounced as a simple consonant were quite common, whereas those in which a simple consonant was produced as a complex one were vanishingly rare.

```
## [1] 30 15
```

To address whether errors were better predicted by cross-linguistic or within-language frequency, we calculated a proportion of productions that were correct for each phone (regardless of the type of error or the substitution pattern). Graphical investigation suggested that in both cases the relationship was monotonic and not linear, so we computed Spearman's rank correlations between the correct repetition score, on the one hand, and the two possible predictors on the other. Although a direct test is missing, the correlation with cross-linguistic frequency [$r() = 0.76$, $p < 0.001$] was greater than that with within-language frequency [$r() = 0.45$, $p = 0.05$].

[5]Note that the substitutions included phones that are not native to Yélî Dnye but do occur in English (e.g., [tʃ]). These data come from careful transcriptions by a native Yélî Dnye speaker who is very fluent in English. This result suggests that several of our participants have mastered production of some English phones, possibly produced within whole English word forms.

NWR scores as a function of item length. We next turned to our second research question by inspecting whether NWR scores varied as a function of word length (Table 4). In this section and all subsequent ones, we only look at first attempts, for the reasons discussed previously. Additionally, we noticed that participants scored much lower on monosyllables than on non-words of other lengths. This is likely due to the fact that the majority of monosyllables were designed to include sounds that are rare in the world's languages, which may be harder to produce or perceive, as suggested by our previous analyses of NWR scores as a function of cross-linguistic phone frequency and error patterns. Therefore, we set monosyllables aside for this analysis.

We observed the typical pattern of lower scores for longer items only for the whole-item scoring, and even there differences were rather small. In a generalized binomial mixed model excluding monosyllables, we included 479 observations, from 40 children producing, in any given trial, one of 24 (non-monosyllabic) potential target words. The analysis revealed a positive effect of age ($ß = 0.56$, SE $ß = 0.14$, $p < 0.001$) and a negative but non-significant estimate for target length in number of syllables ($ß = -0.15$, SE $ß = 0.33$, $p = 0.65$).

Factor structuring individual variation. Our final exploratory analysis assessed whether variation in scores was structured by factors that vary across individuals, as per our third research question. As shown in Figure 4, there was a greater deal of variance across the tested age range, with significantly higher NWR scores for older children (Spearman's rank correlation, given inequality of variance, rho $(5{,}649.08) = 0.47$, $p < 0.01$). In contrast, there was no clear association between NWR scores and sex (Welch $t\ (27.33) = -0.60$, $p = 0.56$), birth order (data missing for 15 children, rho $= (3{,}502.90) = -0.20$, $p = 0.33$), or maternal education (rho $(9{,}628.60) = 0.10$, $p = 0.55$).

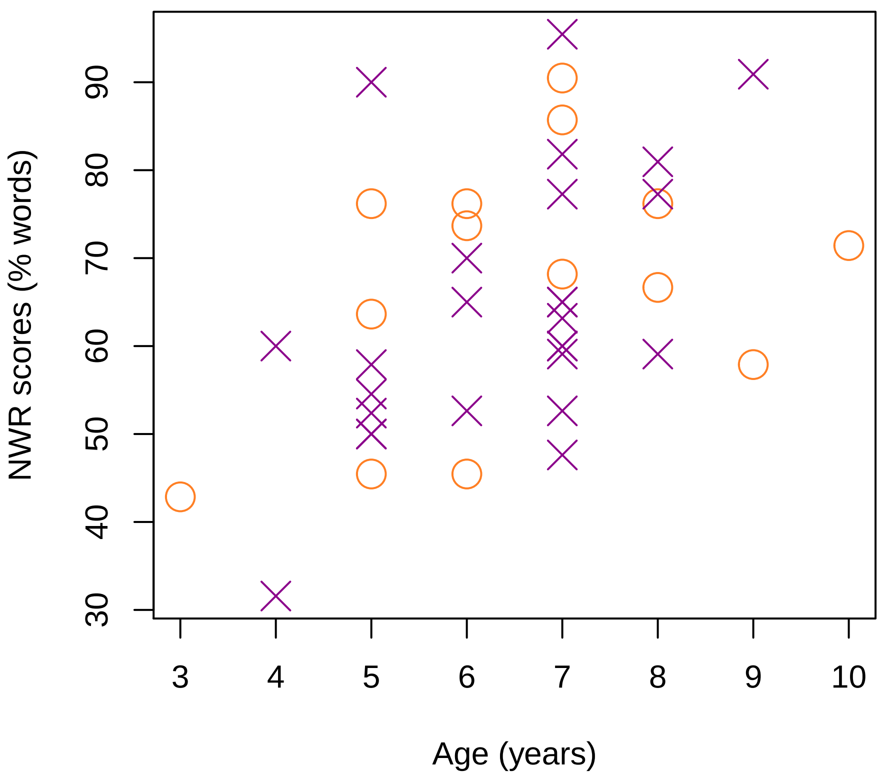

Figure 4. NWR whole-item scores for individual participants as a function of age and sex (purple crosses = boys, orange circles = girls).

Discussion

We used non-word repetition to investigate phonological development in a language with a large phonological inventory (including some typologically rare segments). We aimed to provide additional data on two questions already visitied in NWR work, namely the influence of stimulus length and individual variation, and one that has received less attention, regarding the possible relationship between phone frequency and NWR scores. An additional overarching goal was to revisit this task, which is very commonly used to document phonological development, particularly in children raised in urban settings, and who are learning IndoEuropean languages. We will thus consider our results not only in light of previous work, but also with attention to potential linguistic and population differences.

Associations between NWR and cross-linguistic frequency. FIX LOGIC IN THIS PARA Arguably the most innovative aspect of our data relate to the inclusion of phones that are less commonly found across languages, and rarely used in NWR tasks. Our monosyllabic items included typologically rare segments so that we could test whether lower average segmental frequency is associated with lower NWR scores. Typologically common sounds are associated with higher performance on a handful of other tasks (REFS – M2A: Alex, I added this based on your note, where it sounded like you had some particular studies in mind?) though to our knowledge this has not yet been tested with non-word repetition. Regarding Yélî Dnye in particular, the phonemic inventory is both large and acoustically packed, in addition to containing several typologically infrequent (or unique) contrasts. We therefore expected to see that, while NWR scores would be lower for stimuli with lower average frequency, this effect would be relatively weak because the ambient language puts pressure on Yélî children to distinguish (perceptually and articulatorily) fine-grained phonetic differences in order to successfully communicate with others. Indeed, we found a robust effect of average segmental frequency on NWR performance: Even accounting for age and random effects of item and participant, we see that target words with more frequent segments were repeated correctly more often. This effect is large, with a magnitude more than twice the size of the effect of participant age. This significant effect remains even once also accounting for the frequencies of these segments in Yélî Dnye child-directed speech, which are correlated with their typological frequencies. In sum, typological frequency effects, which have been found in other measurements of phonological processing. appear to strongly affect NWR performance, and do not appear mitigated by language-specific pressure to make finer-grained differences earlier in development.

REWRITE THIS PARA ~With respect to the types of errors in repetition made, we did not see clear patterns to further guide our discussion: base rates of deletion and substitution were fairly low and the relative distribution of errors over, e.g., nasal vs. oral vowels and simple vs. complex consonants, revealed no remarkable bias in error types.~ That said, the lack of a difference could be due to relative imbalance across our stimuli in the use of these phonemic features (e.g., we included many more more oral than nasal targets) and future work should investigate such sources of error bias more systemtically.

Item Length. We investigated the effect of item complexity on NWR scores by varying both the number of syllables in the item. Based on previous work, we had predicted that children would have higher NWR scores for shorter items. That said, previous work has shown both very small (Piazzalunga et al., 2019) and very large (Cristia et al., 2020; Jaber-Awida, 2018) effects of stimulus length and, further, the Yélî Dnye dictionary suggests that mono- and bi-syllabic words are nearly equally frequent in the current language, with trisyllabic and longer words making up a non-trivial 10% of the remaining words. Compare this to, for example, English, which is substantially more skewed toward monosyllabic word forms M2A: Alex I'm going off your note here ("Prediction for Yélî made before seeing the data: The length distribution in Yélî words is more balanced than that in English, and thus the score decline for poly- versus mono-syllables may be less pronounced than that for English.""). I don't have a reference for this, can you please finish the thought or nix this bit?. Setting aside our monosyllabic stimuli, which all contained typologically infrequent segments, we can examine effects of item length among the remaining stimuli, which range between 2 and 4 syllables long. While indeed NWR scores were overall lower for longer items (e.g., see Figure 1), the effect of item length was not significant in a statistical model that additionally accounted for age and random effects of item and participant. In light of mixed prior results of item length, we propose two possible (and non-mutually exclusive) explanations for this minimal impact of item length. First, further extensions of this type of analysis in more populations may reveal that, in general (and cross-linguistically), item length effects are variable between languages, potentially reflecting the distribution of word lengths in the ambient language and other (morpho-)phonological tendencies in the lexicon. Second, above and beyond these language-specific effects, the general impact of item length on NWR score may be relatively small, as shown in Piazzalunga et al.'s (2019) study on Italian and as borne out in the current dataset once controlling for other factors. ADD WHAT LANGUAGES WOULD BE IDEAL TO TEST THIS HYPOTHESIS

Individual differences. A review of previous work (see Introduction) suggested that our anticipated sample size would not be sufficient to detect most individual differences using NWR. We give a brief overview of individual difference patterns of four types in the present data—age, sex, birth order, and maternal education—hoping that these findings can contribute to future meta-analytic efforts aggregating over smaller studies such as ours.

Following prior work, we expected that NWR scores would increase with participant age (Farmani et al., 2018; Kalnak et al., 2014; Vance et al., 2005). Indeed, age was significantly correlated with NWR score and also showed up as a significant predictor of NWR score when included as a control factor in the analyses of both item length and average segmental frequency. In brief, our results underscore the idea that phoonlogical development continues well past the first few years of life, extending into middle childhood and perhaps later (Hazan & Barrett, 2000).

In contrast, previous work shows little evidence for effects of maternal education (e.g., Farmani et al., 2018; Kalnak et al., 2014; Meir & Armon-Lotem, 2017) or participant gender (Chiat & Roy, 2007) on NWR scores. In addition to this prior work, education on Rossel Island, while generally highly valued, is not at all essential to ensuring one's success in society and may not be a reliable index of local socioeconomic variation. There is also limited variation in maternal education across the families in the region of the island where we sampled. We therefore expected little evidence for impact of either participant gender or maternal education in the present study. On the other hand, these predictors have established effects on other language development measures (REFS: M2A: Alex go ahead and pick your faves here). So to the extent that NWR scores share causal links to gender-based differences in development and maternal linguistic input with these other language outcome measures, we might then expect these factors to appear in NWR data. In fact, participant gender and maternal education correlated with NWR score at about r~.1, which is small.

Last but not least, we investigated whether birth order might affect NWR scores, as it does other language tasks, resulting in first-born children showing higher scores on standardized language tests than later-born children (Havron et al., 2019), presumably because later-born children receive a smaller share of maternal input than their older siblings. Given shared caregiving practices and the hamlet organization typical of Rossel communities, children have many sources of adult and older child input that they encounter on a daily basis and first-born children quickly integrate with a much larger pool of both older and younger children with whom they partly share caregivers. Therefore we expected that any effects of birth order on NWR would be attenuated in this context. In line with this prediction, our descriptive analysis showed a non-significant correlation between birth order and NWR score. However, the effect size was larger than that found for the other factors, at r~.2, and thus we believe it may be worth revisiting this question with larger samples in similar child-rearing environments, to further establish whether distributed child care indeed results in more even language outcomes for first- and later-born children.

NWR across languages and cultures. One of the questions in our mind when designing this study was whether NWR was a fair test of phonological development across languages and cultures. Although our data cannot answer this question because we have only sampled one language and culture here, we would like to spend some time discussing the integration of these results to the wider NWR literature. It is important to note at the outset that we cannot obtain a final answer because integration across studies implies not only variation in languages and child-rearing settings, but also in methodological aspects including non-word length, non-word design (e.g., the syllable and phone complexity included in the items), and task administration, among others. Nonetheless, we feel the NWR task is prevalent enough to warrant discussion about this, as it is done for other tasks sometimes used to describe and compare children's language skills across populations, like the MacArthur-Bates Communicative Development Inventory (???).

At first sight, when we had compared our results to those of other studies, we thought the range of performance we observed overlapped with previously observed levels of performance. Paired with our thorough training protocol, we had interpreted the NWR scores among Yélî Dnye learners as indicating that our adaptations to NWR for this context were successful, even given a number of non-standard changes to the training phase and to the design of the stimuli. Additionally, it seemed that Yélî children show edcomparable performance to others tested on a similar task, despite the many linguistic, cultural, and socioeconomic differences between this and previously tested populations, unlike the case that had been reported for the Tsimane'.

To enrich this discussion, we looked for previous studies on monolingual children with normative development learning diverse languages, and entered them when they reported non-word repetition scores based on whole item scoring. We entered data from 14 studies (including ours), presenting data from 12 languages. Specifically, Arabic was represented by Jaber-Awida (2018); Cantonese by (???); English by Vance et al. (2005); Italian by Piazzalunga et al. (2019); Mandarin by (???); Persian by Farmani et al. (2018); Slovak by (???) and (???); Sotho by Wilsenach (2013); Spanish by Balladares et al. (2016); Swedish by Kalnak et al. (2014) and (???); Tsimane' by Cristia et al. (2020); and Yélî Dnye by the present study. Studies varied in the length of non-words that were considered; whenever results were reported separately for different lengths, we calculated overall averages based on lengths of 2 and 3 syllables, for increased comparability. Results separating different age groups are shown in 5.

Several observations can be drawn from this Figure. To begin with, we focus on the comparison between Yélî Dnye and Tsimane'. These two groups have been described as having roughly similar levels of child-directed speech, yet they exhibit very different results, with lower overall NWR scores and (integrating with effect of length in 5) length effects. This may indicate that the conclusion tentatively drawn in Cristia et al. (2020) about lower NWR scores consistent with long-term effects of lower levels of child-directed speech was premature. Naturally, there is an alternative interpretation, namely that input estimation suggesting very slightly higher levels of child-directed speech among the Tsimane' than among Yélî Dnye learners is inaccurate. In fact, careful reading of previous reports highlight important methodological differences in how input quantity has been estimated across papers: Casillas et al. (2020) hand-coded speech with the help of a native research assistant, and then summed all child-directed speech, which effectively

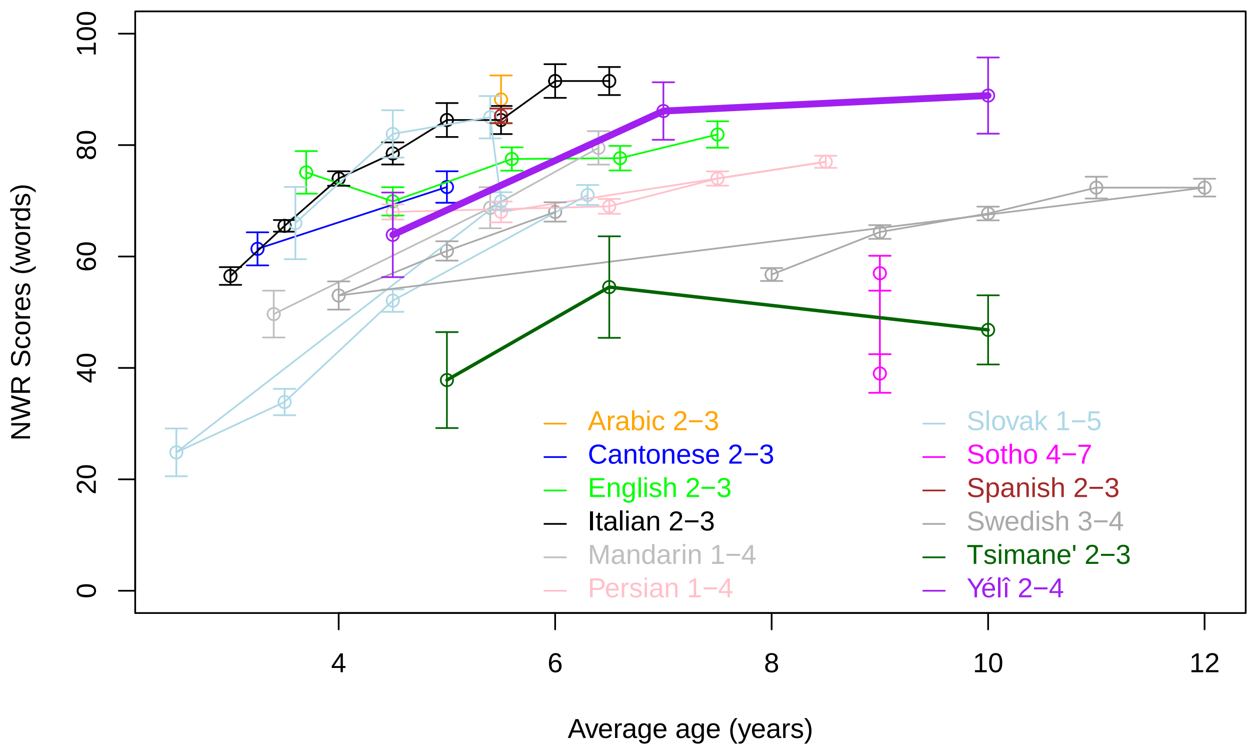

Figure 5. NWR scores as a function of age (in years), averaged across multiple non-word lengths, as a function of children's native languages. The legend indicates language and the length of non-words (in syllables). Central tendency is mean; error is one standard error.

establishes an upper boundary of the speech children could potentially process. (???) estimated quantities from behavioral observations on the frequency of child-directed one-on-one conversation, which is probably closer to a lower boundary. Finally, Scaff et al. (2021) used human annotation for detecting speech but an automated temporal method for assigning speech as child-directed or not, in a way that could lead to over-estimation (because any speech by e.g. a female adult that was not temporally close to speech by others would count as child-directed). A final answer to the question of how much child-directed speech is afforded to Yélî and Tsimane' children must await fully comparable methods.

That said, Cristia et al. (2020) also pointed out another characteristic of the Tsimane' culture, and this was the relatively low prevalence of literacy, and generally the variable access to formal education. This is a very different case from the Yélî population studied here, where all adults have accumulated several years of schooling, and literacy is widespread. If this second hypothesis holds, then this may mean that there are phonetic effects of learning to read in the input afforded to young children, and that this has consequences for young children's encoding and decoding of sounds in the context of NWR tasks. Notice that this is not the same as the oft-recorded effect of learning to read affecting NWR performance, illustrated for instance in the data for Sotho in 5. These two data points have been gathered from two groups of children, all exposed mainly to Sotho, but children with higher NWR had been learning to read in Sotho, whereas those with lower scores were learning to read in English. What is at stake in the second interpretation of the lower scores observed among the Tsimane' is related to literacy in the broader population (rather than in the tested children themselves).

Although exciting, this hypothesis is only one of many. Another plausible explanation is that the Tsimane' results are not comparable to the previous body of literature, and specifically to our study. Cristia et al. (2020) administered the NWR in the form of a group game played outside, with a non-native experimenter providing the target, and each person of the group attempting it in their stead. This immediately means a number of important methodological differences with the standard implementation of NWR, where children are tested individually, they hear items spoken by a native speaker (often over headphones), the experimenter tends to belong to the same community as the children, and testing occurs in quiet conditions (with little background noise). Thus, a priority is for additional data gathered using this more novel testing paradigm in other populations, or from the Tsimane' using the more traditional paradigm.

Broadening our discussion to all of the studies in our literature review, we notice that there is rather wide variation of the range of NWR scores found across these samples, as well as the strength of age effects. We performed some exploratory analyses to see whether features of the languages children were learning could be related to their overall NWR scores. We extracted the number of phonemes in the language from PHOIBLE and coded whether words in the language tended to be longer or shorter based on information in the papers or other sources. Neither of these two predictors explained variance in 5. It is possible that average word length plays a role, but often researchers incorporate this into their design by including longer items when the native language allows this, with e.g. Sotho non-words having 4-7 syllables in length. To be more certain whether language characteristics do account for meaningful variation in NWR scores, it will be necessary to design NWR tasks that are cross-linguistically valid. We believe this will be excedingly difficult (or perhaps impossible), since it would entail defining a 10-20 set of items that are meaningless in all of the languages as well as phonotactically legal. An alternative may be to find ways to regress out some of these effects, and thus compare languages while controlling for choices of phonemes, syllable structure, and overall length of the NWR items.

Additional observations. Some portion of the errors were introduced when the participant produced a real word (in Yélî Dnye or English) in response to the stimulus. Real-word repetitions here made up two thirds of errorful repetitions—this is quite high compared to past work (e.g., Castro-Caldas et al., 1998), but it is unclear what caused this pattern in the current study: Castro and colleagues' (1998) study focused on adults rather than children, the task was administered by a team including a foreign, English-speaking researcher, and the particularities of the Yélî Dnye phonological inventory result in many true-word phonetic neighbors. Follow-up work exploring this type of error in children from other populations in addition to further work on Yélî children may clarify this effect.

Conclusions. While NWR can, in theory, be used to test a variety of questions about phonological development in any language, previous work has been primarily limited to a handful of related languages spoken in urban, industrialized contexts. The present study shows that, not only can NWR be adapted for very different populations than have previously been tested, but that effects of age and typological frequency may strongly influence phonological development across these diverse settings, while effects of item length, participant gender, maternal education, and birth order, may either have little impact on this facet of language development or have an impact that vaies depending on the linguistic, cultural, and sociodemographic properties of the population under study. Because these latter predictors strongly relate to other language outcomes, the present findings raise the issue of why NWR would pattern differently, what that could tell us about the relationship between lexical development, phonological development, and the input environment and, last but not least, what is implied about the joint applicability of these outcome measures as a diagnostic indicator for language delays and disorders. In the meanwhile, we take the present findings as robustly supporting the idea that phonological development continues well past early childhood and as yielding preliminary support for a connection between individual learners and global language patterns when it comes to acoustic and articulatory markedness.

Acknowledgments

We are grateful to the individuals who participated in the study, and the families and communities that made it possible. The collection and annotation of these recordings was made possible by Ndapw:éé Yidika, Taakêmê Ńamono, and Y:aaw:aa Pikuwa; with thanks also to the PNG National Research Institute, and the Administration of Milne Bay Province. We owe big thanks also to Stephen C. Levinson for his invaluable advice and support and Shawn C. Tice for helpful discussion during data collection. AC acknowledges financial and institutional support from Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. MC acknowledges financial support from an NWO Veni Innovational Scheme grant (275-89-033).

References

Balladares, J., Marshall, C., & Griffiths, Y. (2016). Socio-economic status affects sentence repetition, but not non-word repetition, in Chilean preschoolers. First Language, 36(3), 338–351. https://doi.org/10.1177/0142723715626067

Barclay, K. J. (2015). A within-family analysis of birth order and intelligence using population conscription data on swedish men. Intelligence, 49, 134–143.

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer (Version 6.1.35). Retrieved from http://www.praat.org/

Bowey, J. A. (2001). Nonword repetition and young children's receptive vocabulary: A longitudinal study. Applied Psycholinguistics, 22(3), 441–469.

Brandeker, M., & Thordardottir, E. (2015). Language exposure in bilingual toddlers: Performance on nonword repetition and lexical tasks. American Journal of Speech-Language Pathology, 24(2), 126–138.

Brown, P. (2011). The cultural organization of attention. In A. Duranti, E. Ochs, & and Bambi B Schieffelin (Eds.), Handbook of Language Socialization (pp. 29–55). Malden, MA: Wiley-Blackwell.

Brown, P. (2014). The interactional context of language learning in Tzeltal. In I. Arnon, M. Casillas, C. Kurumada, & B. Estigarribia (Eds.), Language in interaction: Studies in honor of Eve V. Clark (pp. 51–82). Amsterdam, NL: John Benjamins.

Brown, P., & Casillas, M. (n.d.). Childrearing through social interaction on Rossel Island, PNG. In A. J. Fentiman & M. Goody (Eds.), Esther Goody revisited: Exploring the legacy of an original inter-disciplinarian (pp. XX–XX). New York, NY: Berghahn.

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Papuan community. Journal of Child Language, XX, XX–XX.

Castro-Caldas, A., Petersson, K. M., Reis, A., Stone-Elander, S., & Ingvar, M. (1998). The illiterate brain. Learning to read and write during childhood influences the functional organization of the adult brain. Brain: A Journal of Neurology, 121(6), 1053–1063. https://doi.org/10.1093/brain/121.6.1053

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.

COST Action. (2009). Language impairment in a multilingual society: Linguistic patterns and the road to assessment. Brussels: COST Office. Available Online at: Http://Www.bi-Sli.org.

Cristia, A., Farabolini, G., Scaff, C., Havron, N., & Stieglitz, J. (2020). Infant-directed input and literacy effects on phonological processing: Non-word repetition scores among the Tsimane'. PLoS ONE, 15(9), e0237702. https://doi.org/https://doi.org/10.1371/journal.pone.0237702

de Santos Loureiro, C., Braga, L. W., Nascimento Souza, L. do, Nunes Filho, G., Queiroz, E., & Dellatolas, G. (2004). Degree of illiteracy and phonological and metaphonological skills in unschooled adults. Brain and Language, 89(3), 499–502. https://doi.org/10.1016/j.bandl.2003.12.008

Estes, K. G., 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.

Farabolini, G., Rinaldi, P., Caselli, C., & Cristia, A. (2021). Non-word repetition in bilingual children: The role of language exposure, vocabulary scores and environmental factors. Speech Language and Hearing.

Farmani, H., Sayyahi, F., Soleymani, Z., Labbaf, F. Z., Talebi, E., & Shourvazi, Z. (2018). Normalization of the non-word repetition test in Farsi-speaking children. Journal of Modern Rehabilitation, 12(4), 217–224.

Foley, W. A. (1986). The Papuan languages of New Guinea. Cambridge, UK: Cambridge University Press.

Gallagher, G. (2014). An identity bias in phonotactics: Evidence from Cochabamba Quechua. Laboratory Phonology, 5(3), 337–378. https://doi.org/10.1515/lp-2014-0012

Gallon, N., Harris, J., & Van der Lely, H. (2007). Non-word repetition: An investigation of phonological complexity in children with Grammatical SLI. Clinical Linguistics & Phonetics, 21(6), 435–455.

Gathercole, S. E., Willis, C., & Baddeley, A. D. (1991). Differentiating phonological memory and awareness of rhyme: Reading and vocabulary development in children. British Journal of Psychology, 82(3), 387–406.

Grätz, M. (2018). Competition in the family: Inequality between siblings and the intergenerational transmission of educational advantage. Sociological Science, 5, 246–269.

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., & Group, E. M.-C. C. S. (2019). The effect of older siblings on language development as a function of age difference and sex. Psychological Science, 30(9), 1333–1343.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. Journal of Phonetics, 28(4), 377–396.

Jaber-Awida, A. (2018). Experiment in non word repetition by monolingual Arabic preschoolers. Athens Journal of Philology, 5(4), 317–334. https://doi.org/10.30958/ajp.5-4-4

Kalnak, N., Peyrard-Janvid, M., Forssberg, H., & Sahlén, B. (2014). Nonword repetition–a clinical marker for specific language impairment in Swedish associated with parents' language-related problems. PloS One, 9(2), e89544.

Lancy, D. F. (2015). The anthropology of childhood. Cambridge, UK: Cambridge University Press.

Levinson, S. C. (2020). A grammar of Yélî Dnye, the Papuan language of Rossel Island. Berlin, Boston: De Gruyter Mouton.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & de Vos, C. (2012). A prelinguistic gestural universal of human communication. Cognitive Science, 36(4), 698–713. https://doi.org/10.1111/j.1551-6709.2011.01228.x

Maddieson, I. (2005). Correlating phonological complexity: Data and validation. UC Berkeley PhonLab Annual Report, 1(1).

Maddieson, I. (2013a). Consonant inventories. The World Atlas of Language Structures Online. Retrieved from https://wals.info/chapter/1

Maddieson, I. (2013b). Vowel quality inventories. The World Atlas of Language Structures Online. Retrieved from https://wals.info/chapter/2

Maddieson, I., & Levinson, S. C. (n.d.). The phonetics of yélî dnye, the language of rossel island.

Meir, N., & Armon-Lotem, S. (2017). Independent and combined effects of socioeconomic status (SES) and bilingualism on children's vocabulary and verbal short-term memory. Frontiers in Psychology, 8, 1442.

Meir, N., Walters, J., & Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. International Journal of Bilingualism, 20(4), 421–452.

Moran, S., & McCloy, D. (Eds.). (2019). PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. Retrieved from https://phoible.org/

Peute, A. A. K., Fikkert, P., & Casillas, M. (n.d.). Early consonant production in Yélî Dnye and Tseltal.

Piazzalunga, S., Previtali, L., Pozzoli, R., Scarponi, L., & Schindler, A. (2019). An articulatory-based disyllabic and trisyllabic Non-Word Repetition test: reliability and validity in Italian 3-to 7-year-old children. Clinical Linguistics & Phonetics, 33(5), 437–456.

Scaff, C., Stieglitz, J., Casillas, M., & Cristia, A. (2021). Daylong audio recordings of young children in a forager-farmer society show low levels of verbal input with minimal age-related changes. Draft.

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.

Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3–7 years. International Journal of Language & Communication Disorders, 40(1), 29–48.

Wilsenach, C. (2013). Phonological skills as predictor of reading success: An investigation of emergent bilingual Northern Sotho/English learners. Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer, 29(2), 17–32. https://doi.org/10.5785/29-2-554

Table 1

NWR stimuli in orthographic (Orth.) and phonological (Phon.) representations.

Practice		Monosyll		Bisyll		Trisyll		Tetrasyll	
Orth.	Phon.	Orth.	Phon.	Orth.	Phon.	Orth.	Phon.	Orth.	Phon.
nopimade	nɔpimæṭɛ	dp:a	ṭp̃æ	kamo	kæmɔ	dimope	ṭimɔpɛ	dipońate	ṭipɔnætɛ
poni	pɔni	dpa	ṭpæ	kańi	kæni	diyeto	ṭijɛtɔ	ńomiwake	nɔmiwækɛ
wî	wɯ	dpâ	ṭpɑ	kipo	kipɔ	meyadi	mɛjæṭi	todiwuma	tɔṭiwumæ
		dpê	ṭpə	ńoki	nɔki	mituye	mitujɛ	wadikeńo	wæṭikɛnɔ
		dpéé	ṭpeː	ńomi	nɔmi	ńademo	næṭɛmɔ		
		dpi	ṭpi	piwa	piwæ	ńayeki	næjɛki		
		dpu	ṭpu	towi	tɔwi	ńuyedi	nujɛṭi		
		gh:ââ	ɣ̃ɑː	tupa	tupæ	pedumi	pɛṭumi		
		ghuu	ɣuː			tiwuńe	tiwunɛ		
		kp:ââ	kp̃ɑː			tumowe	tumɔwɛ		
		kpu	kpu			widońe	wiṭɔnɛ		
		lv:ê	lβʲ̃ə			wumipo	wumipɔ		
		lva	lβʲæ						
		lvi	lβʲi						
		t:êê	t̃əː						
		tpê	tpə						

Table 2

Number (and percent) of vowel targets that were correctly repeated (Corr.), deleted (Del.), or substituted, as a function of vowel type, and whether the error resulted in a nasality change (Nasal Err.) or only a quality change (Qual. Err.)

	Corr.	Del.	Nasal Err.	Qual. Err	% Corr.	% Del.	% Nasal Err.	% Qual Err.
Nasal Target	100	0	39	17	64.1	0.0	25.0	10.9
Oral Target	1992	12	52	205	88.1	0.5	2.3	9.1

Table 3

Number (and percent) of consonant targets that were correctly repeated (Corr.), deleted (Del.), or substituted, as a function of the complexity of the consonant, and whether the error resulted in a change of complexity (Cmpl Err.) or not (Othr Err.)

	Corr.	Del.	Cmpl Err.	Othr Err.	% Corr.	% Del	% Cmpl Err.	% Othr Err.
Complex Target	257	0	218	48	49.1	0.0	41.7	9.2
Simple Target	1425	6	2	120	91.8	0.4	0.1	7.7

Table 4

NWR means (and standard deviations) measured in whole-word scores and normalized Levenshtein Distance (NLD), separately for the four stimuli lengths.

	Word	NLD
1 syll	48 (22)	40 (18)
2 syll	79 (22)	8 (9)
3 syll	78 (19)	7 (7)
4 syll	74 (32)	9 (12)