CrossMark

Nonword repetition stimuli for Vietnamese-speaking children

Giang Pham[1] · Kerry Danahy Ebert[2] · Kristine Thuy Dinh[1] · Quynh Dam[1]

Abstract

Nonword repetition (NWR) has been a widely used measure of language-learning ability in children with and without language disorders. Although NWR tasks have been created for a variety of languages, minimal attention has been given to Asian tonal languages. This study introduces a new set of NWR stimuli for Vietnamese. The stimuli include 20 items ranging in length from one to four syllables. The items consist of dialect-neutral phonemes in consonant–vowel (CV) and CVC sequences that follow the phonotactic constraints of the language. They were rated high on wordlikeness and have comparable position segments and biphone probabilities across stimulus lengths. We validated the stimuli with a sample of 59 typically developing Vietnamese–English bilingual children, ages 5 to 8. The stimuli exhibited the expected age and length effects commonly found in NWR tasks: Older children performed better on the task than younger children, and longer items were more difficult to repeat than shorter items. We also compared different scoring systems in order to examine the individual phoneme types (consonants, vowels, and tones) and composite scores (proportions of phonemes correct, with and without tone). The study demonstrates careful construction and validation of the stimuli, and future directions are discussed.

Keywords Tonal languages · Phonological memory · Language assessment · Phonotactic properties

Nonword repetition (NWR) is a relatively straightforward task in which a child is presented with auditory nonsense words and asked to repeat them aloud. Despite its apparent simplicity, NWR has attracted a great deal of attention in the recent literature due to its abilities to index language development and, potentially, to identify language disorders. Versions of NWR have been constructed and tested for a variety of languages. However, minimal attention has been given to Asian languages, particularly to those that utilize tones. Additional consideration of this group of languages is needed in order to determine whether the properties of NWR tasks, and particularly their ability to identify children with developmental language disorders, are indeed universal across languages.

This study describes the development and validation of an NWR task for Vietnamese, an Asian tonal language. A set of carefully constructed Vietnamese NWR stimuli are presented, along with an assessment of their properties in a group of 59

typically developing school-age children. To frame the study, we first discuss the uses and interpretation of NWR tasks and the expected properties of NWR stimuli, and then we consider evidence across multiple languages.

What does nonword repetition assess?

NWR tasks have attracted attention for their relations with typical language development and their potential to contribute to the identification of developmental language disorders. Children's performance on NWR tasks is most commonly interpreted as a reflection of their phonological memory skills, although performance is also influenced by speech perception, lexical knowledge, and motor skills (Coady & Evans, 2008). There is a clear association between NWR performance and vocabulary size (e.g., Gathercole, Service, Hitch, Adams, & Martin, 1999), such that children with larger vocabularies perform better on NWR tasks. It has been posited that NWR reflects the ability to store unfamiliar phonological forms, a necessary step in vocabulary acquisition (Gathercole et al., 1999). This hypothesis is supported by evidence that NWR performance can predict new word learning (Gathercole, Hitch, & Martin, 1997). Thus, the task may tap into crucial aspects of children's language-learning ability.

✉ Giang Pham
  gpham@sdsu.edu

1   School of Speech, Language, and Hearing Sciences, San Diego State University, 5500 Campanile Drive, SLHS Room 238, San Diego, CA 92182, USA

2   Rush University, Chicago, IL, USA

It is perhaps not surprising, then, that poor NWR performance is associated with the presence of developmental language disorder[1] (DLD). Studies of monolingual English-speaking children consistently find significant differences in NWR performance between children with DLD and their typically developing peers (see Coady & Evans, 2008, for a review). Meta-analysis (Estes, Evans, & Else-Quest, 2007) has indicated that the average effect size in these studies is large (with between-group differences averaging 1.27 standard deviations) and that the relationship holds across a wide age range (4 to 12 years). NWR has been termed a possible clinical marker for DLD (Bishop, North, & Donlan, 1996), meaning that poor performance could be a reliable behavioral indicator of the disorder. This judgment was originally based on studies of NWR in English, but subsequent work on the diagnostic accuracy of NWR has spanned several languages (e.g., Dispaldro, Leonard, & Deevy, 2013; Kapalková, Polišenská, & Vicenová, 2013; Thordardottir & Brandeker, 2013). In reviewing the literature, Coady and Evans reached the conclusion that NWR "does make a good diagnostic tool" for DLD (p. 33). Others have tempered this conclusion by suggesting that the task should be combined with other tools (Archibald & Joanisse, 2009).

Stimuli properties and scoring procedures

NWR stimuli have a number of expected properties. First, they must adhere to the phonotactic constraints of a language. In other words, they can contain only admissible phonemes, phoneme sequences, and suprasegmental patterns for the target language. Within this broad constraint, additional considerations may exist. For example, nonwords can be made more or less wordlike. One method of estimating "wordlikeness" is through adult ratings. Gathercole, Willis, Emslie, and Baddeley (1991) had adults rate nonwords on a 5-point scale (5 = *very much like a real word*). Words that are more like real words are easier to repeat (Coady & Evans, 2008; Gathercole et al., 1991). This effect may be due to the ability to draw upon long-term lexical knowledge (and rely less on phonological memory; Gathercole, 1995). Because of this increased reliance on lexical knowledge, it is possible that more wordlike nonwords are better able to distinguish between children with and without DLD (Archibald & Gathercole, 2006). Thus, potential advantages for using more wordlike NWR stimuli with younger children include increasing the feasibility of the repetition task and creating a task that could be used to identify DLD.

Nonword stimuli have also been evaluated in terms of phonotactic probability, which is generally defined as the frequency with which phonemes occur within a given language. Vitevitch and Luce (2004) described two types of phonotactic probability: positional segment frequency and biphone frequency. *Positional segment frequency* refers to how often a phoneme occurs in a certain position in a word. For example, /s/ in the word-initial position has a higher phonotactic probability than /j/ in the word-initial position in English, because /s/ occurs more often in the initial position than /j/. *Biphone frequency* refers to how often a pair of phonemes co-occur in a word. For example, the biphone sequence /hɪ/ has a higher phonotactic probability in English than does /vʊ/, because it occurs more frequently. Nonwords with high phonotactic probabilities have been rated higher on wordlikeness than have nonwords with low phonotactic probabilities (Vitevitch & Luce, 2004). Because nonwords with high phonotactic probability sound more like real words in a given language, they are generally repeated faster and more accurately than nonwords with low phonotactic probability (Munson, Kurtz, & Windsor, 2005; Vitevitch, Luce, Pisoni, & Auer, 1999; Vitevitch & Luce, 2005). However, Munson et al. found that children with DLD are more affected by phonotactic probability than are typically developing children; the performance gap between children with DLD and those with typical language development was larger on words with low phonotactic probability than on those with high phonotactic probability.

Performance on NWR tasks is also influenced by the length of the stimuli. When children are asked to repeat nonwords with increasing numbers of syllables, poorer accuracy is anticipated on longer stimuli, which presumably tax the phonological memory system. Length effects have indeed been robustly supported in both English (e.g., Ellis Weismer et al., 2000) and other languages (Ebert, Kalanek, Cordero, & Kohnert, 2008; H. J. Lee, Kim, & Yim, 2013; Santos & Bueno, 2003). Word length also appears to affect children with DLD more than their typically developing peers, as the NWR deficits of children with DLD tend to be larger on longer words (Dispaldro et al., 2013; Dollaghan & Campbell, 1998; Girbau & Schwartz, 2007; Munson et al., 2005). However, cross-linguistic differences in phonology can influence the difficulty of any given word length, such that children tend to be able to repeat longer nonwords in languages (such as Spanish) that tend to have longer real words (Ebert et al., 2008; Windsor, Kohnert, Lobitz, & Pham, 2010).

NWR stimuli should demonstrate an age effect as well, such that older children perform better than younger children in repeating the nonwords. This effect is expected, because phonological memory grows during childhood (as do the other constructs that are likely tapped by NWR). Again, the expected age effects have been robustly demonstrated in the literature, both for English (e.g., Gathercole, Willis, Baddeley, & Emslie, 1994) and for other languages (e.g.,

---

[1] A variety of names have been used for this disorder, including specific language impairment (SLI) and primary language impairment (PLI). Following the recommendation of the recent international CATALISE Consortium (Bishop, Snowling, Thompson, & Greenhalgh, 2017), we adopt the term *developmental language disorder*.

Santos & Bueno, 2003). More specifically, though, a well-constructed NWR task will show sensitivity to age effects within the age range for which it is designed. That is, a task designed for preschool children should show an age effect for preschoolers, although older children may perform at ceiling on the task.

Children's performance on NWR tasks have been scored using two primary methods: the percentage of phonemes produced correctly (e.g., Dollaghan & Campbell, 1998) or a holistic score in which the nonwords have been considered either entirely correct or incorrect (e.g., Gathercole et al., 1994). The practical advantage to holistic scoring is the ability to score online. Online scoring can be quite useful, particularly for clinicians who are using NWR tasks as a screening tool for DLD (Archibald & Joanisse, 2009). However, holistic scoring cannot provide information at the phoneme level, which limits the ability to conduct further analyses of the types of errors produced and the contexts in which errors are made.

In addition to overall phonemes correct, recent studies of NWR performance have examined variations in scoring that focus on specific types of segments. In a study of Swedish nonword repetition, Sundström, Samuelsson, and Lyxell (2014) applied various scoring systems from Shriberg, Austin, Lewis, McSweeny, and Wilson (1997), including phonemes correct, the percentage of consonants correct, and the percentage of vowels correct. Researchers found that children made more errors with consonants than with vowels (i.e., vowels correct > consonants correct). Greater accuracy on vowels in NWR tasks has also been reported in other languages, such as Kannada (Shylaja, Abraham, Leela Thomas, & Swapna, 2011) and Spanish (Girbau, & Schwartz, 2007). Relatively few studies of NWR tasks have reported separate scores for consonants versus vowels. Further studies that examine both consonant and vowel accuracy in NWR will indicate whether vowels are easier for children to repeat or whether this is a language-specific effect.

NWR with linguistically diverse populations

Flexibility in stimulus construction, ease of administration, and links to language development and disorders make NWR a promising task to include in assessment. One additional task advantage is a potential reduction in assessment bias in diverse populations. Because NWR inherently uses unfamiliar stimuli, the task emphasizes language processing over experience-dependent linguistic knowledge (Thordardottir & Brandeker, 2013). For children who speak a nonmainstream English dialect, NWR tasks have been shown to be less biased in comparison to traditional language assessments (Ellis Weismer et al., 2000).

It should be noted that although linguistic knowledge plays a reduced role in NWR, such knowledge does still influence performance on the task. For bilingual children, exposure to a specific language has been shown to influence NWR performance (Sorenson Duncan & Paradis, 2016). Some studies have shown that typically developing bilingual children completing an NWR task in their second language perform, on average, below typically developing monolingual speakers of that language (Windsor et al., 2010); however, this finding has not been universal (S. A. S. Lee & Gorman, 2013). The task does appear to be less affected by prior linguistic experience (i.e., limited second language exposure) than other tasks used to assess language disorders (Thordardottir & Brandeker, 2013).

Most studies on NWR with bilingual populations have focused on second language performance (e.g., Sorenson Duncan & Paradis, 2016; Windsor et al., 2010). Few studies have compared bilingual speakers tested in their *first* language to monolingual speakers of that language. One such study revealed no difference between these groups. H. J. Lee, Kim, and Yim (2013) compared typically developing Korean–English preschoolers to monolingual Korean peers using NWR and vocabulary tests. They found that the monolingual group outperformed the bilingual group in Korean vocabulary. However, both groups showed similar performance on Korean NWR. Findings suggest that, unlike lexical knowledge, which is highly dependent on language exposure, phonological memory in bilingual children's first language, as measured by NWR, remains relatively comparable to that in monolingual children.

NWR has received relatively little attention in tonal languages. NWR stimuli have been developed and tested in both Cantonese (Ho, Leung, & Cheung, 2011; Stokes, Wong, Fletcher, & Leonard, 2006; A. M. Y. Wong, Kidd, Ho, & Au, 2010) and Mandarin (Barbosa, Jiang, & Nicoladis, 2017), yet to our knowledge only one of these studies has published the set of stimuli and described in detail how the stimuli were designed and developed (Stokes et al., 2006). Using comparisons between a typically developing preschool group, a preschool group with DLD, and a typically developing toddler group matched to the language-impaired group by receptive grammar, Stokes et al. found an effect for age, in that older children performed better than younger children on the NWR task. However, they did not find differences between age-matched preschool children with and without DLD in terms of NWR performance.
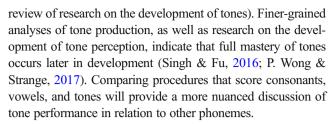
The authors focused on language-specific characteristics of Cantonese in explaining this result, arguing that it may have stemmed from the lack of complex syllable structures and stress patterns in Cantonese, or from phonotactic frequency effects (Stokes et al., 2006). Leonard (2014) later attributed the lack of between-group differences on Cantonese NWR to the tones themselves; that is, he argued that the tones make each syllable exceptionally salient, reducing the effectiveness of NWR. It should be noted that Stokes et al. scored only the reproduction of consonants and vowels; reproduction of the tones was not included in the task score.

Another result did suggest that reducing the salience of tones resulted in a more effective Cantonese NWR task. A. M. Y. Wong et al. (2010) tested school-age children with and without DLD using an NWR subtest of the Hong Kong Test of Specific Learning Difficulties in Reading and Writing for Primary School Students (second edition). This task was reported to present strings of syllables that are phonotactically permissible in Cantonese but that do not carry typical Cantonese tones. Instead, the same tone was used, and the syllables were presented at a steady rate, causing the authors to describe the task as "more akin to serial recall" (A. M. Y. Wong et al., 2010, p. 31). Nonetheless, performance on the NWR task in this study was significantly poorer for children with DLD than for typically developing children.

Beyond the language-specific characteristics of Cantonese (in particular, tone), small sample size limited the statistical power to detect group differences in the Stokes et al. (2006) investigation. We also note that the properties of the Cantonese stimuli themselves may have influenced the results. Stokes and colleagues included syllables in their stimuli that did not follow conventional rules for creating NWR. That is, these items included a combination of syllables that conformed to the phonotactic constraints of Cantonese (i.e., IN syllables) and of syllables that did not (i.e., OUT syllables). The rationale for including OUT syllables was to disassociate NWR performance from the influence of lexical knowledge. However, the inclusion of OUT syllables may have contributed to the poor overall accuracy by all three groups of children in the Cantonese study (Stokes et al., 2006). Although NWR stimuli with low phonotactic probability have been used to reduce the role of lexical knowledge in other NWR studies (e.g., Vitevitch & Luce, 2005), OUT syllables, by definition, would have zero probability of occurring in the language. Furthermore, IN and OUT syllables were combined within single stimulus items, making performance on these two different types of syllables difficult to separate.

Whether NWR can be an effective tool to identify DLD in Asian tonal languages is still an open empirical question. In the present study, we distinguished the NWR stimuli for Vietnamese in several ways. First, the Vietnamese stimuli did not include OUT syllables. Having all items conform to the phonotactic constraints of the language would increase overall accuracy, particularly among young children. Second, the present study included an in-depth analysis of the phonological properties of the stimuli themselves as part of validating the measure. Careful consideration and reporting of stimulus properties will contribute to the methodological discussion on creating NWR tasks across languages. Third, we included a variety of scoring systems in order to further examine the role of tones in comparison to consonants and vowels in NWR. Children tend to be able to produce tones that are perceived to be correct early in phonological development, with minimal errors noted after 2 years of age (see Singh & Fu, 2016, for a

review of research on the development of tones). Finer-grained analyses of tone production, as well as research on the development of tone perception, indicate that full mastery of tones occurs later in development (Singh & Fu, 2016; P. Wong & Strange, 2017). Comparing procedures that score consonants, vowels, and tones will provide a more nuanced discussion of tone performance in relation to other phonemes.

In sum, NWR appears to be a less biased—but not completely unbiased—task that may contribute to the identification of DLD. NWR stimuli have been developed and tested in a host of languages. In many cases, the developers have published complete sets of NWR stimuli, allowing additional research on the language. There are now publicly available NWR stimuli for languages including Brazilian Portuguese (Santos & Bueno, 2003), Slovak (Kapalková et al., 2013), Swedish (Radeborg, Barthelom, Sjöberg, & Sahlén, 2006), Korean (H. J. Lee et al., 2013), and Italian (Dispaldro et al., 2013). Investigation of NWR properties across languages is crucial to understanding the task and what it measures, as well as to developing clinical and research tools.

Research questions

The purpose of this study was to develop and validate a set of NWR stimuli based on the Vietnamese language. This article describes the construction of the nonword stimuli and presents the full set of stimuli. We also present results of a preliminary validation study in which the NWR task was administered to 59 typically developing Vietnamese-speaking school-age children. This preliminary validation aimed to establish the expected properties of the NWR stimuli within this population. We thus considered the following research questions (RQs) for this new set of NWR stimuli:

RQ1: What are the properties of the NWR stimuli in terms of phonotactic probability?
RQ2: Does NWR performance decrease as the length of the nonwords increases?
RQ3: Does NWR performance increase with age within the sample?
RQ4: Are there differences in scoring systems and how they relate to age and length?

Method

NWR stimuli for the Vietnamese language

The nonword stimuli followed the structures of the Vietnamese language. Vietnamese words can consist of one, two, three, or four syllables, and even, in rare instances, six

syllables (Thompson, 1963). Most words in Vietnamese are one or two syllables in length, such as *mẹ* "mother" and *xe đạp* "bicycle" (Nguyen, 1997). We used the syllable as the base unit for stimulus creation because there continues to be debate on what constitutes a word in Vietnamese (Thompson, 1963). The Vietnamese syllable consists of three main components—the onset, rime, and tone—and both closed and open syllables (e.g., both consonant–vowel–consonant [CVC] and CV syllables) are permitted (Tang, 2007). Phonemes in Vietnamese consist of consonants in syllable-initial position, consonants and semivowels in syllable-final position, vowel singletons and diphthongs, and tones (B. Pham & McLeod, 2016; Tang, 2007). The numbers of initial consonants (21–24), final consonants (6), and tones (4–6) vary across regional dialects (B. Pham & McLeod, 2016).

To create the nonword stimuli, we first selected the syllable-initial consonants, syllable-final consonants, vowels, and tones that are shared across dialects, to make the task accessible to Vietnamese-speaking children across Vietnam and the diaspora. Of the 16 syllable-initial consonants shared across dialects (B. Pham & McLeod, 2016), we selected six phonemes that are consistently found in the phonetic inventories of young Vietnamese-speaking children (Tang & Barlow, 2006): /b, d, t, s, k, ɣ/. Vietnamese syllable-final consonants are restricted to six nasals and unreleased stops (/m, n, ŋ, pˀ, tˀ, kˀ/), of which two of the phonemes are produced differently across dialects when preceded by certain vowels: /n, tˀ/ (Pham & McLeod, 2016). Therefore, we selected the four consonants that would not be influenced by vowel contexts: /pˀ, kˀ, m, ŋ/. Finally, we selected nine vowel singletons (/i, ɛ, a, ă, ɤ, ɤ̆, o, ɯ, u/) and three tones (rising, *sắc*; level, *ngang*; and falling, *huyền*) that are found across dialects. Using the select set of phonemes, we created three lists of (a) CV syllables with a falling tone, (b) CV syllables with a level tone, and (c) CVC syllables with a rising tone. We excluded from these lists syllables that violated phonotactic properties of the Vietnamese language. For example, "xă" /să/ and "xâ" /sɤ̆/ cannot be open syllables in Vietnamese and, therefore, were excluded from the CV lists. Two adults fluent in Vietnamese reviewed the lists in order to exclude syllables that were real words. Exclusion of syllables was based on discussions between the two speakers and referrals to online Vietnamese dictionaries.

The finalized lists consisted of syllables that were phonologically possible and classified as not real words. From these lists, we combined syllables to form the stimuli across four stimulus lengths (one, two, three, or four syllables). One-syllable stimuli followed a CVC structure; two-syllable stimuli had the form CVC CV; three-syllable stimuli had the form CVC CVC CV; and four-syllable stimuli had the form CVC CV CVC CV. These syllable structures are found in real Vietnamese words and are similar to the NWR stimuli for Cantonese (Stokes et al., 2006), the only other tonal Asian language for which such a task has been created and stimuli

have been made publicly available. Similar to the Cantonese NWR, we planned a consistent tone pattern within each stimulus length: All one-syllable items had a rising tone (*sắc*), written as superscript /5/; two-syllable items had a rising–level tone pattern (*sắc–ngang*), written as the superscripts /5/ and /1/, respectively; three-syllable items had a rising–rising–falling pattern (*sắc–sắc–huyền*), written as /5/, /5/, and /2/, respectively; and four-syllable items had a rising–falling–rising–level pattern (*sắc–huyền–sắc–ngang*), written as /5/, /2/, /5/, and /1/, respectively. These tones patterns reflected patterns found in real Vietnamese words.

It was our intention to create a set of stimuli that could be repeated by young to old children. To this end, we focused on items that were rated by adults to be highly wordlike. A set of 24 items were piloted with four speakers, six items per stimulus length. The speakers (two female and two male) were adult native speakers of Vietnamese who had immigrated to the United States after the age of 12. All adult speakers were able to repeat the items with 100% accuracy. As in the procedures outlined by Gathercole et al. (1991), the adults listened to each item and were asked to rate its wordlikeness using a 4-point scale (4 = *most like a Vietnamese word*). Two or more speakers rated the following items as a 1 (i.e., *not at all like Vietnamese*): the one-syllable item /fɛm5/, the two-syllable item /bin5 ɣɤ1/, the three-syllable item /fuɯ5 sip5 to2/, and the four-syllable item /kɛŋ5 ɣɤ2 mɤ̆m5 duɯ1/. After rating these items as a 1, the adults confirmed that these nonwords did not sound like Vietnamese words. On the basis of the adult responses, we omitted these items (one item from each stimulus length), for a final set of 20 items. Appendix A displays the task protocol, which includes instructions for administering the task in English and in Vietnamese and the final list of stimuli. The items were audio-recorded by a female native speaker of Vietnamese in an audio booth using an Edirol R-09 digital audio recorder and divided into separate audio files using the Goldwave software (Craig, 2008). The audio recordings of the stimuli are publicly available from the IRIS Digital Repository (www.iris-database.org).

Participants

The participants were typically developing Vietnamese–English bilingual children enrolled in kindergarten, first grade, or second grade in a Southern California school district. The participants were recruited for a larger study in which bilingual children completed a set of language tasks in both Vietnamese and English. School records were used to verify that the children met the criteria for typical development. These criteria included normal hearing, as shown by consistent passes on school-based hearing screenings, and normal language development, as shown by the absence of any current special education services received at school. We verified normal cognition using the Primary Test of Nonverbal

Intelligence (Ehrler & McGhee, 2008) and excluded children who scored below a standard score of 80. Seventy children met all of these inclusionary criteria for typical development.

All of the participants spoke Vietnamese as their home language, per parent report, and school instruction took place exclusively in English. To keep our sample as homogeneous as possible, we excluded seven children who were also exposed to a third language, such as Chinese or Cambodian, one additional child who could not complete a narrative task in Vietnamese, and three additional children who did not complete the NWR task. A total of 59 children thus participated in this study (36 girls/23 boys). The mean age of participants was 7 (years);0 (months) (SD = 10 months, range from 5;8 to 8;6).

Procedures

Individual children participated in the study in a quiet area of their schools during school hours with a trained research assistant fluent in Vietnamese. The NWR task was part of a battery of Vietnamese language measures that was administered within an hour-long session. Participants completed the language measures in generally the same order; at times, adjustments to the order of tasks were made in order to accommodate changes in the children's school schedule.

The NWR stimuli were administered using audio files that were uploaded to an iPad. The research assistant and child each wore headphones connected with a headphone splitter to listen jointly to the stimuli. The research assistant first played the instructions via headphones to the child, and then paused the iPad to practice two items aloud. The practice items were repeated if needed, and the child received feedback on the accuracy of the repetition. The research assistant then played the audio recording in order to administer the task items. Children heard each task item a single time and were instructed to repeat each nonword immediately after hearing it. As in NWR studies in other languages (i.e., Dollaghan & Campbell, 1998; Stokes et al., 2006), the items were presented in order of increasing difficulty (see the score form in Appx. A). The research assistant paused the audio recording between words to allow time for the child to respond. If an item was inadvertently skipped, the research assistant continued to the next word and returned to any missed items at the end. The child responses were audio-recorded for later transcription and scoring.

Appendix A displays the task instructions and scoring form. The scoring guidelines were consistent with Dollaghan and Campbell (1998): Substitutions and omissions were considered errors, whereas additions or slight distortions (e.g., lisps) were not considered errors. Each phoneme (consonant, vowel, or tone) was scored on the basis of accuracy (1 = correct, 0 = incorrect). Some of the final consonant sounds could be co-articulated due to the preceding vowel, such as the one-syllable item "tóng" /tɤŋ^{m5}/. In this case, the consonant-final velar /ŋ/ is produced along with lip closure so as to co-articulate the /m/. (Co-articulations are marked with superscript consonant sounds following the syllable-final consonants on the scoring sheet.) Because it was unclear from the audio recordings alone whether the child had produced lip closure alongside velar closure, a correct score of 1 was entered if the transcriber heard either the consonant-final sound /ŋ/ or /m/ in cases of co-articulation (see the superscript font in App. A).

Accuracy was then calculated using five scoring systems: proportions of phonemes correct (PPC), including consonants, vowels, and tones; proportions of consonants correct (PCC); proportions of vowels correct (PVC); proportions of tones correct (PTC); and proportions of consonants and vowels correct (i.e., no tones: PPCnoT). The different scoring systems highlighted either a composite of phonemes (PPC and PPCnoT) or the accuracy of an individual type of phoneme (PCC, PVC, and PTC). The fifth scoring system (PPCnoT) was calculated in order to allow for comparisons with previous studies. PPC for English and other Indo-European languages is based on solely consonants and vowels (e.g., Dollaghan & Campbell, 1998), since lexical tone is not applicable. Additionally, other studies on NWR in a tonal Asian language (e.g., Stokes et al., 2006) have calculated PPC without tone.

Interrater reliability was calculated with a second independent transcriber using the audio recordings from nine randomly selected participants (15.6% of the sample). Both transcribers spoke Vietnamese as a first language, had an educational background in speech–language pathology, and had completed prior training in phonetic transcription. The total point-by-point agreement across all phoneme types (180 possible) ranged from 91% to 97%, with an average of 94%. Point-by-point agreement for tones (50 possible) was the highest (M = 99%; range = 96% to 100%), followed by consonants (80 possible: M = 94%; range = 93% to 98%), and vowels (50 possible: M = 87%; range = 80% to 98%).

Data analysis

Research Question 1 (RQ1) examined the phonotactic probabilities across stimulus lengths.

We calculated the positional segment frequency and biphone frequency according to the guidelines outlined by Vitevitch and Luce (2004). The positional segment frequencies for initial, medial, and final positions were calculated as the frequency of a target phoneme in a given segment position, divided by the sum of the frequencies of all possible phonemes that could occur in the same segment position. Similarly, the biphone frequency for CV and for VC sequences was calculated as the frequency of the target biphone divided by the sum of the frequencies of all possible two-phoneme combinations that could occur in the same sequence.

Frequency counts were obtained from the Corpora of Vietnamese Texts (CVT; G. Pham, Kohnert, & Carney, 2008), which is a written language corpus that includes over one million words collected from Vietnamese newspaper articles and children's books. The orthographic features of Vietnamese facilitate the search for phonological properties in a text-based corpus. Vietnamese uses a Romanized orthography in which each syllable is separated by a space. The orthography is highly transparent, with a near one-to-one sound–letter correspondence (B. Pham & McLeod, 2016). Consistent with Vitevitch and Luce (2004), frequency counts were converted to log (base 10) values.

Phonotactic probability calculations aligned with the syllable structure of Vietnamese (for a review, see B. Pham & McLeod, 2016), in which the initial segment consists of a consonant sound, the medial segment consists of the main vowel, and final segment consists of a limited set of consonants and semivowels. Tone is a suprasegmental feature that encompasses the entire syllable and is orthographically marked above the main vowel (e.g., đáp /dap^5/ "to respond"). Because the CVT (G. Pham et al., 2008) is a written-language corpus, the frequency for a given vowel was the sum of the frequencies of the target vowel with each of the six tones. To illustrate, the positional segmental frequency for the medial vowel /a/ was calculated as the log (base 10) of sum of the frequencies of /a^1 + a^2 + a^3 + a^4 + a^5 + a^6/ divided by the sum of the log values for all main vowels found in the medial position (and all tone combinations). Similarly, tone was included in the calculations of CV and VC biphone frequencies. To illustrate, the phonotactic probability of the biphone /ba/ was calculated as the log (base 10) of the sum of the frequencies of /ba^1 + ba^2 + ba^3 + ba^4 + ba^5 + ba^6/ divided by the sum of the log values for all possible CV combinations (including tones).

RQ1 asked whether the phonotactic probabilities were comparable across stimulus lengths. Consistent with the phonotactic probability calculator (Vitevitch & Luce, 2004), we focused on two dependent measures to address this question: (a) the sum of all the segment probabilities (initial C, medial V, and final C), and (b) the sum of all the biphone probabilities (both CV and VC). We compared phonotactic probabilities across stimulus lengths using separate analyses of variance (ANOVAs) for both phoneme probabilities and biphone probabilities, with stimulus length (one, two, three, or four syllables) as the between-subjects factor.

RQ2 asked whether NWR performance would decrease as the length of the nonwords increased. To answer this question, we conducted separate repeated measures multivariate ANOVAs (MANOVAs) for each scoring system, with length (one, two, three, or four syllables) as the within-subjects factor. Significant omnibus tests were followed by pairwise comparisons using a least significant difference (LSD) adjustment. RQ3 asked whether NWR performance increased with age. To

answer this question, we conducted bivariate correlations to identify the associations between age and each stimulus length within a given scoring system. RQ4 asked whether there were differences between the scoring systems in how they related to age and length. To answer this question, we conducted a repeated measures MANOVA with total scores as the dependent measures and the type of scoring system as the within-subjects factor. Significant omnibus tests were followed by pairwise comparisons using an LSD adjustment.

Results

RQ1: Phonotactic probability

As we described previously, two Vietnamese adult speakers verified that the syllables used in the stimuli were not real words. Nonetheless, the raw frequency counts revealed that 14 whole syllables had frequency counts greater than zero (see Appx. B). These syllables were parts of borrowed or obscure words found in newspaper articles. The highest frequency count occurred for the one-syllable item "cóp" /kɔp^5/, which occurred 23 times (<0.002% frequency) in the CVT (G. Pham et al., 2008). "Cóp" is a borrowed word from English meaning "to copy [an answer]." The second highest frequency count occurred within the second syllable "mống"/moŋ^{m5}/ of a three-syllable item, which occurred 15 times (<0.002% frequency) in the CVT. The syllable "mống" was part of the word "mầm mống," which relates to germination. The remaining syllables occurred five or fewer times within the one-million-word corpus. Table 1 displays the phoneme and biphone probabilities of the NWR stimuli. The ANOVA results were nonsignificant for both phoneme probabilities [$F(3, 26) = 1.80, p = .17$] and biphone probabilities [$F(3, 26) = 2.37, p = .09$], indicating comparable position segment and biphone probabilities across stimulus lengths.

RQ2: Effects of length

Table 2 displays descriptive statistics for the children's performance on the NWR task and a summary of the pairwise comparisons of the stimulus lengths for each scoring system. For PPC, the omnibus test for length was significant [$F(3, 56) = 24.09, p < .001, \eta_p^2 = .56$]. Pairwise comparisons revealed differences between all stimulus lengths: Performance was more accurate on one-syllable than on two-syllable items ($p = .02$); two-syllable items were more accurate than three-syllable items ($p < .001$); and three-syllable items were more accurate than four-syllable items ($p = .001$).

For PPCnoT, the omnibus test was again significant [$F(3, 56) = 32.73, p < .001, \eta_p^2 = .64$]. Pairwise comparisons revealed significant differences between all lengths: One-

Table 1 Phonotactic probabilities of NWR stimuli by length

	One Syllable		Two Syllables		Three Syllables		Four Syllables	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Initial C	.04370	.00240	.04151	.00229	.04093	.00489	.04007	.00543
Medial V	.01131	.00047	.01026	.00048	.01059	.00057	.01084	.00075
Final C	.08082	.00613	.08643	.00653	.08355	.00486	.08098	.00438
Total phoneme	.13582	.00776	.13737	.00744	.13675	.00537	.13036	.00808
CV	.00271	.00034	.00215	.00040	.00242	.00041	.00237	.00041
VC	.00247	.00051	.00218	.00072	.00262	.00051	.00254	.00047
Total biphone	.00519	.00067	.00430	.00066	.00517	.00071	.00494	.00053

See the text for the calculation of these values. C = consonant sounds; V = vowel sounds

syllable items were more accurate than two-syllable items ($p =$ .001); two-syllable items were more accurate than three-syllable items ($p = .005$); and three-syllable items were more accurate than four-syllable items ($p < .001$).

For PCC, the omnibus test for length was significant [$F(3, 56) = 33.81$, $p < .001$, $\eta_p^2 = .64$]. Pairwise comparisons revealed differences between two of the three pairs: Performance on one-syllable and two-syllable items did not differ ($p = .07$); one and two-syllable items were each more accurate than three-syllable items ($p < .001$ and $p = .04$, respectively); and three-syllable items were more accurate than four-syllable items ($p < .001$).

For PVC, the omnibus test was significant [$F(3, 56) = 11.10$, $p < .001$, $\eta_p^2 = .37$]. Pairwise comparisons revealed differences between two of the three pairs: One-syllable items were more accurate than two-syllable items ($p = .03$); two-syllable items were more accurate than three-syllable items ($p = .01$); and three- and four-syllable items did not differ ($p = .47$).

The omnibus test for PTC approached statistical significance [$F(3, 56) = 2.72$, $p = .05$, $\eta_p^2 = .13$]. Pairwise comparisons revealed a difference for one of the three pairs: One-

syllable items did not differ from two-syllable items ($p = .09$); two-syllable items were more accurate than three-syllable items ($p = .01$); and three-syllable items did not differ from four-syllable items ($p = .05$).

RQ3: Effects of age

Table 3 displays the bivariate correlations between chronological age and scores. Age was positively related to four of the five total scores: PPC ($r = .34$, $p = .008$), PPCnoT ($r = .33$, $p = .010$), PCC ($r = .26$, $p = .047$), and PVC ($r = .33$, $p = .011$). Similarly, age was positively related to four-syllable items for four of the five scoring systems: PPC ($r = .36$, $p = .005$), PPCnoT ($r = .37$, $p = .004$), PCC ($r = .31$, $p = .018$), and PVC ($r = .35$, $p = .007$). Age was not related to the PTC for either total score ($r = .18$) or four-syllable items ($r = .15$).

RQ4: Comparing scoring systems

An omnibus repeated measures MANOVA revealed a significant difference between the five scoring systems [$F(4, 55) =$

Table 2 Descriptive statistics by stimulus length and scoring system

Length	One Syllable			Two Syllables			Three Syllables			Four Syllables			Total			Compare Lengths
	M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range	
PPC	.92	.06	.75–1.00	.90	.06	.69–1.00	.86	.08	.64–1.00	.82	.09	.57–.97	.86	.06	.71–.97	1 > 2 > 3 > 4
PPCnoT	.91	.07	.73–1.00	.87	.08	.56–1.00	.83	.09	.55–1.00	.77	.11	.42–.96	.82	.07	.62–.95	1 > 2 > 3 > 4
PCC	.92	.10	.60–1.00	.89	.09	.60–1.00	.85	.10	.52–1.00	.75	.12	.50–.97	.83	.08	.60–.96	1 = 2; 2 > 3 > 4
PVC	.89	.14	.60–1.00	.84	.13	.50–1.00	.79	.13	.40–1.00	.77	.14	.30–1.00	.80	.10	.50–.94	1 > 2 > 3; 3 = 4
PTC	.96	.11	.60–1.00	.98	.04	.80–1.00	.95	.07	.67–1.00	.97	.05	.75–1.00	.97	.04	.82–1.00	1 = 2; 2 > 3; 3 = 4

Pairwise comparisons were conducted using repeated measures MANOVAs. PPC = proportions of phonemes correct (consonants, vowels, and tones); PPCnoT = proportions of consonants and vowels correct (no tones); PCC = proportions of consonants correct; PVC = proportions of vowels correct; PTC = proportions of tones correct.

Table 3 Bivariate correlations between chronological age and scores

Length	PPC	PPCnoT	PCC	PVC	PTC
One syllable	.12	.18	.21	− .01	− .10
Two syllables	.07	.07	− .02	.12	.06
Three syllables	.23	.22	.12	.26	.18
Four syllables	.36**	.37**	.31*	.35**	.15
Total scores	.34**	.33*	.26*	.33*	.18

PPC = proportion of phonemes correct (consonants, vowels, and tones); PPCnoT = proportion of consonants and vowels correct (no tones); PCC = proportion of consonants correct; PVC = proportion of vowels correct; PTC = proportion of tones correct. $^* p < .05$, $^{**} p < .01$.

74.19, $p < .001$, $\eta_p^2 = .84$]. Pairwise comparisons revealed that each scoring system was different from the remaining four, with p values ranging from <.001 to .014. The following summarizes the average performance from highest to lowest total scores: PTC > PPC > PCC > PPCnoT > PVC. PTC had the highest total score, with a mean of .97 ($SD = .04$), indicating that the average performance on tones reached ceiling levels (see Table 2 for the means and averages of the total scores). Differences in these scoring systems indicate that children performed highest on tones, followed by consonants, and then vowels.

Discussion

The purpose of the present study was to develop, validate, and disseminate a set of Vietnamese NWR stimuli. The stimuli were constructed with careful attention to the phonology of Vietnamese and to the desirable qualities of nonwords. These stimuli were then tested within a group of 59 Vietnamese–English bilingual children.

Properties of the stimuli

The procedures and analyses conducted here revealed several strengths of the stimuli. First, the stimuli were constructed to include only dialect-neutral phonemes of Vietnamese, making them suitable for all Vietnamese-speaking children. Phonological complexity across items was demonstrated by similar phonotactic probabilities across stimulus lengths. Additionally, the stimuli were rated as being highly wordlike by adult native speakers, which increases their utility for young children. The youngest children in the present study were 5 years of age, and no child scored below 50% total accuracy (on PVC, the lowest of the scoring systems); thus, even young children may be

able to achieve some success in repeating these Vietnamese nonwords. Finally, the possibility of real words embedded within the nonwords was investigated. Although a few nonwords contained one-syllable real words, all were of low frequency and unfamiliar to the adult Vietnamese speakers who reviewed the words. Moreover, NWR stimuli in other languages have also sometimes included real-word syllables, given language-specific constraints on consonants, vowels, and syllable structure (Ebert et al., 2008).

In addition to the careful construction and inspection of the phonological properties of the stimuli, the analyses supported the expected effects of age and length within a sample of typically developing children. Four of five scoring systems related positively to age, meaning that older children tended to have higher scores. Significant length effects were seen for four of the five scoring systems. Moreover, because the phonological complexity of the syllables within the nonwords was comparable across stimulus lengths, the length effect was not the result of inadvertently including more complex syllables within longer words. In other words, the decreased performance on longer words can be attributed to difficulty in retaining more phonological information. These results suggest that the stimuli did indeed tax the capacities of the children tested here and can provide an appropriate assessment of phonological memory within the tested age range.

One limitation of scoring certain items was related to co-articulation (e.g., the simultaneous production of /ŋ/ and /m/ in the item /tɔŋᵐ⁵/). Because scoring was conducted using audio recordings, it was difficult to determine whether the children produced both consonant-final sounds. Therefore, we opted for a lenient scoring procedure, awarding the child a correct phoneme when either consonant-final sound was produced. To increase the precision in scoring these segments (i.e., to award a correct phoneme only if the child co-articulated both consonant-final sounds), future studies could incorporate video recording or combine live scoring of coarticulated segments with audio recordings.

Another potential limitation to these nonwords is the consistency of tone and syllable patterns. All words of a given length utilized the same syllable shape and tone pattern. This approach follows that of Dollaghan and Campbell (1998) and may allow for a clearer demonstration of length effects (because the complexity is consistent within a given length). Stokes et al. (2006), too, argued for consistency of tonal patterns within a given word length. Because this was the first study to examine NWR in Vietnamese, we opted to manipulate one parameter at a time—that is, increasing length while maintaining phonological complexity. An alternative approach

would be to vary the syllable shapes within a word length (e.g., Gathercole et al., 1994). Future studies of NWR in Vietnamese that manipulate tone while maintaining stimulus length could further examine the role of tone in repetition tasks (see Sundström et al., 2014, for a similar study of prosody in Swedish).

Alternate scoring systems

Analyses comparing the five scoring systems in this study (i.e., PPC, PPCnoT, PCC, PVC, and PTC) revealed significant differences among the scores. The two composite scoring systems (PPC and PPCnoT) had the advantage of showing more consistent effects of length, as performance differed significantly for each stimulus length. In contrast, the PVC and PCC scoring systems were less consistent in length effects, with significant differences between only three of the four pairs. Both PPC and PPCnoT also showed the expected age effects. Combining the age effects with length effects suggests that the two composite scoring systems (PPC and PPCnoT) may be the most optimal. One potential advantage of PPCnoT is the ability to make comparisons to nontonal languages such as English (Dollaghan & Campbell, 1998), Spanish (Ebert et al., 2008), and Italian (Dispaldro et al., 2013), as well as to previous studies of tonal languages (e.g., Stokes et al., 2006) that utilized a PPC scoring system without tones. At the same time, PPC encompasses all phonemes (consonants, vowels, and tones), which could provide a more comprehensive picture of children's phonological memory.

Participants were most accurate in repeating tones and least accurate in repeating vowels, with consonants falling in between. Indeed, children performed near ceiling levels in tone repetition. This finding is consistent with tone acquisition studies in similar languages, such as Cantonese, in which children acquire tones early in development (e.g., So & Dodd, 1995; see Singh & Fu, 2016). However, we also note that fine-grained analyses (i.e., using band pass filtering) of tone production in Mandarin-speaking children (P. Wong & Strange, 2017) have indicated that tone production in 6-year-olds is not yet adult-like; it is possible that a more fine-grained analysis would reveal subtle differences between our participants' repetition of tones and the adult models. Near-ceiling-level performance in tones may have also been related to the restricted set of tones used in the stimuli—namely, the three tones that are shared across regional dialects of Vietnamese (B. Pham & McLeod, 2016). Including only dialect-neutral tones makes the task accessible to all Vietnamese-speaking children. However, it may have also limited the difficulty level of the tones in this task. It should also be noted that these

NWR stimuli used the same tone pattern for each word of a given length; it is possible that varying the tones within each word length would increase the difficulty of tone repetition.

Participants were also significantly more accurate at repeating consonants than vowels. This finding stands in contrast to previous studies of NWR tasks in other languages, such as Swedish (Sundström et al., 2014), Spanish (Girbau & Schwartz, 2007), and Kannada (Shylaja et al., 2011). There are at least two possible explanations for this finding. First, methodological considerations regarding phoneme selection and transcription may have influenced the results. The nonword stimuli consisted of a select set of phonemes (consonants, vowels, and tones) that were shared across regional dialects of Vietnamese. Additionally, consonants were limited to ones that are considered to be early acquired (Tang & Barlow, 2006) in order for the stimuli to be accurately repeated by young children, resulting in a total of six syllable-initial consonants and four syllable-final consonants in the task. The higher accuracy on consonants might have been related to the restricted set of consonants included in the stimuli. Additionally, whereas interrater reliability for the NWR transcription was high across phoneme types, interrater agreement on the vowels was relatively lower on average and had a wider range of values than did agreement on the consonants and tones (see the Procedures section above). Greater variability in transcription might have contributed to the lower accuracy on vowels than on other phoneme types.

The second explanation is related to the important role of vowels in the Vietnamese language. A Vietnamese syllable requires at minimum a main vowel at the segmental level and a tone at the suprasegmental level; consonants are optional (B. Pham & McLeod, 2016). Vietnamese has a large vowel inventory that consists of 14 to 16 singletons and diphthongs, depending on the dialect (B. Pham & McLeod, 2016). Unlike in English, the vowels in Vietnamese are relatively constant across dialects, whereas the consonants and tones vary by region. Variation in consonants and tones across dialects may make them more flexible or mutable in the word recognition process. Using a word reconstruction paradigm in a similar tonal language, Mandarin Chinese, Wiener and Turnbull (2016) found that adult native speakers were more likely to change tones or consonants in order to create a word from a nonword; when forced to change vowels, participants were less accurate and slower to create words from nonwords. Wiener and Turnbull interpreted this finding in terms of information load—namely that vowels contribute more to word recognition than do tones and consonants in

Mandarin. Relatedly, consonant repetition in the present study (i.e., PCC scores) was not significantly related to age, whereas vowel repetition (PVC) did relate positively to age for both total scores and four-syllable items; this result suggests that younger children are less accurate on vowels than older children. These results underscore the importance of transcribing and scoring vowels, as they were the most taxing (and thus perhaps the most sensitive) segment within the Vietnamese nonwords.

Future directions

This study was intended as a first step in validating the Vietnamese NWR stimuli, and it provides the foundation for future work. As we discussed previously, NWR tasks are expected to index language development and to discriminate between children with and without language disorders. It will be important to relate children's performance on these NWR stimuli to their performance on other indices of Vietnamese language development, particularly to receptive vocabulary, due to its close links with NWR (Gathercole et al., 1999).

This new set of NWR stimuli can contribute to work with bilingual children who speak Vietnamese as a first or second language. There are large Vietnamese-speaking communities living outside of Vietnam, in countries such as the United States, Australia, and Canada. In the U.S., there are over 1.5 million Vietnamese people, of whom nearly 400,000 are children (U.S. Census Bureau, 2010) who may speak Vietnamese as a first language. Within Vietnam there are 54 ethnicities, among which most minority groups speak Vietnamese as a second language (B. Pham & McLeod, 2016). Gathering tools to assess language-learning ability will be vital for addressing the educational needs of these populations. Comparing NWR performance across the two languages of bilingual children will contribute to a broader understanding of how phonological memory relates to language exposure (e.g., Thordardottir & Brandeker, 2013).

Finally, comparing the performance of children with DLD to typically developing children is an important next step. The few NWR studies on tonal languages have had mixed findings. A. M. Y. Wong et al. (2010) found that children with DLD performed more poorly than did their typically developing peers on a nonword task. However, this study's nonword stimuli are not publicly available for independent replication of these results. Stokes et al. (2006) found that NWR was not a sensitive measure to identify DLD in Cantonese. The inability to distinguish between typical and impaired groups in this study may have been in part related to the target age range, the small sample size, and the type of nonwords used. Although the phonemes used in the NWR stimuli were found in the Cantonese language, many of the biphone sequences violated the phonological rules of the language. The relatively low accuracy in the study suggests that this task was too difficult. In the present study, we introduced NWR stimuli that followed the phonological rules of Vietnamese and were rated high in wordlikeness. Repetition accuracy was substantially higher than in Stokes et al. and was consistent with previous studies of NWR tasks with school-age children (e.g., Dollaghan & Campbell, 1998). In sum, it is still an open empirical question whether NWR tasks can be used to identify DLD in tonal languages. Further studies will be needed with a variety of tonal languages and age ranges. Careful construction and validation of the stimuli will provide a crucial foundation for this work.

Appendix A Vietnamese nonword repetition task (Lặp lại từ không có nghĩa)

Instructions: Turn on the audio recording of the NWR task. Use two headphones with a splitter so that child and tester are both listening. Pause in between words to allow the child time to repeat each word. Listen to each word only once. If needed, continue to the next word and then come back to missed words at the end. Say aloud the two practice items and let the child know if she or he repeated them correctly. You can repeat the practice items if needed. When the child is ready, begin the task. Task items are played from the audio recording and can only be heard one time each. Based on the audio recording, score the child's responses using the following form.

Say to the child "*Bây giờ con sẽ nghe một số từ không có thật. Con lặp lại y như con nghe nhé. Con sẽ làm thử hai từ.* [Practice items here] *Bây giờ con sẽ bắt đầu nhé. Con lắng nghe và lặp lại y như con nghe nhé.*" [Now you will hear some words that are not real. Repeat them exactly how you hear them. Here are two words for practice. . . . Now you will begin. Listen carefully and repeat exactly what you hear.]

NWR Scoring Form

Instructions: Score each phoneme as correct (1) or incorrect (0). Score consonants and vowels on the lines and tones in the boxes provided.

Child ID: _____   Test date: _____

1 Syllable	2 Syllables	3 Syllables	4 Syllables
báp /bap^5/ _ _ _ ☐	đéng mư /dɛŋ^5 mɯ^1/ _ _ _ ☐ _ _ ☐	gốp cớm xè /ɣop^5 kɤm^5 sɛ^2/ _ _ _ ☐ _ _ _ ☐ _ _ ☐	ghím bừ phóp ke /ɣim^5 bɯ^2 fɔp^5 kɛ^1/ _ _ _ ☐ _ _ ☐ _ _ _ ☐ _ _ ☐
đúm /dum^5/ _ _ _ ☐	gứng te /ɣɯŋ^5 tɛ^1/ _ _ _ ☐ _ _ ☐	đép mống gò /dɛp^5 moŋ^{m5} ɣɤ^2/ _ _ _ ☐ _ _ _ ☐ _ _ ☐	đấp xè gống mư /dɤp^5 sɛ^2 ɣoŋ^{m5} mɯ^1/ _ _ _ ☐ _ _ ☐ _ _ _ ☐ _ _ ☐
cóp /kɔp^5/ _ _ _ ☐	xéng đư /sɛŋ^5 dɯ^1/ _ _ _ ☐ _ _ ☐	bẳm phóc xò /bɐm^5 fɔk^{p5} so^2/ _ _ _ ☐ _ _ _ ☐ _ _ ☐	búm xò phíp te /bum^5 so^2 fip^5 tɛ^1/ _ _ _ ☐ _ _ ☐ _ _ _ ☐ _ _ ☐
xớp /sɤp^5/ _ _ _ ☐	phóm ke /fɤm^5 kɛ^1/ _ _ _ ☐ _ _ ☐	xím cót bừ /sim^5 kɔt^5 bɯ^2/ _ _ _ ☐ _ _ _ ☐ _ _ ☐	cóm xừ bếm gơ /kɔm^5 sɯ^2 bem^5 ɣɤ^1/ _ _ _ ☐ _ _ ☐ _ _ _ ☐ _ _ ☐
tóng /tɔŋ^{m5}/ _ _ _ ☐	kép gư /kɛp^5 ɣɯ^1/ _ _ _ ☐ _ _ ☐	múng tóm ghè /muŋ^{m5} tɤm^5 ɣɛ^2/ _ _ _ ☐ _ _ _ ☐ _ _ ☐	míp tồ ghém bư /mip^5 to^2 ɣem^5 bɯ^1/ _ _ _ ☐ _ _ ☐ _ _ _ ☐ _ _ ☐
PPC ____ / 20	____ / 35	____ / 55	____ / 70 = Total PPC ____ / 180
PPCnoT ____ / 15	____ / 30	____ / 50	____ / 65 = Total PPCnoT ____ / 160
PCC ____ / 10	____ / 15	____ / 25	____ / 30 = Total PCC ____ / 80
PVC ____ / 5	____ / 10	____ / 15	____ / 20 = Total PVC ____ / 50
PTC ____ / 5	____ / 10	____ / 15	____ / 20 = Total PTC ____ / 50

Appendix B Segmental and biphone probabilities by stimulus length for all NWR items

Item	IPA	Spelling	Initial C	Medial V	Final C	CV	VC
One Syllable							
1	bap[5]	báp	.04134	.01187	.07696	.00309	.00272
2	dum[5]	đúm	.04375	.01142	.08219	.00245	.00208
3	kɔp[5]	cóp/kóp	.04621	.01133	.07696	.00307	.00226
4	sɤp[5]	xớp/sớp	.04124	.01058	.07696	.00239	.00205
5	tɔŋ[m5]	tóng	.04596	.01133	.09100	.00257	.00326
Two Syllables							
1	dɛŋ[5]	đéng	.04375	.00981	.09100	.00230	.00149
	mɯ[1]	mư	.04106	.01075		.00230	
2	ɣɯŋ[5]	gứng	.03957	.01075	.09100	.00194	.00321
	tɛ[1]	te	.04596	.00981		.00181	
3	sɛŋ[5]	xéng/ séng	.04124	.00981	.09100	.00273	.00149
	dɯ[1]	đư	.04375	.01075		.00258	
4	fɤm[5]	phớm	.03858	.01058	.08219	.00135	.00223
	kɛ[1]	ke	.04081	.00981		.00229	
5	kɛp[5]	kép	.04081	.00981	.07696	.00229	.00246
	ɣɯ[1]	gư	.03957	.01075		.00194	
Three Syllables							
1	ɣop[5]	gốp	.03957	.01084	.07696	.00238	.00196
	kɤm[5]	cớm/kớm	.04621	.01058	.08219	.00241	.00223
	sɛ[2]	xè/sè	.04124	.00981		.00273	
2	dɛp[5]	đép	.04375	.00981	.07696	.00230	.00246
	moŋ[m5]	mống	.04106	.01084	.09100	.00296	.00349
	ɣɤ[2]	gờ	.03957	.01058		.00173	
3	băm[5]	bắm	.04134	.00975	.08219	.00259	.00294
	fɔkp[5]	phóc	.03858	.01133	.08599	.00243	.00283

	so^2	xồ/sồ	.04124	.01084		.00243	
4	sim^5	xím/ sím	.04124	.01130	.08219	.00323	.00262
	kɤt^5	cớt/ kớt	.04621	.01058	.08484	.00241	.00212
	bɯ^2	bừ	.04134	.01075		.00219	
5	muŋ^{m5}	múng	.04106	.01142	.09100	.00255	.00328
	tɤm^5	tớm	.04596	.01058	.08219	.00234	.00223
	ɣɛ^2	ghè	.02556	.00981		.00159	
Four Syllables							
1	ɣim^5	ghím	.02556	.01130	.08219	.00184	.00262
	bɯ^2	bừ	.04134	.01075		.00219	
	fɔp^5	phóp	.03858	.01133	.07696	.00243	.00226
	kɛ^1	ke	.04081	.00981		.00229	
2	dɤ̌p5	đấp	.04375	.01057	.07696	.00285	.00288
	sɛ2	xè/sè	.04124	.00981		.00273	
	ɣoŋ^{m5}	gống	.03957	.01084	.09100	.00210	.00349
	mɯ^1	mư	.04106	.01075		.00230	
3	bum^5	búm	.04134	.01142	.08219	.00246	.00208
	so^2	xồ/sồ	.04124	.01084		.00266	
	fip^5	phíp	.03858	.01130	.07696	.00247	.00207
	tɛ^1	te	.04596	.00981		.00181	
4	kɔm^5	cóm/kóm	.04621	.01133	.08219	.00307	.00236
	sɯ^2	xừ/ sừ	.04124	.01075		.00275	
	bem^5	bếm	.04134	.01301	.08219	.00241	.00292
	ɣɤ^1	gơ	.03957	.01058		.00173	
5	mip^5	míp	.04106	.01130	.07696	.00273	.00207
	to^2	tồ	.04596	.01084		.00285	
	ɣɛm^5	ghém	.02556	.00981	.08219	.00159	.00269
	bɯ^1	bư	.04134	.01075		.00219	

Items are displayed in International Phonetic Alphabet (IPA) and in Vietnamese orthography. Graphemes that represented the same phoneme (e.g., /c, k/ or /x, s/) were searched separately and then combined to calculate frequency counts. Frequency counts from the Corpora of Vietnamese Texts (Pham et al., 2008). See the text for calculations.

References

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

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, *52*, 899–914.

Barbosa, P. G., Jiang, Z., & Nicoladis, E. (2017). The role of working and short-term memory in predicting receptive vocabulary in

monolingual and sequential bilingual children. *International Journal of Bilingual Education and Bilingualism*. Advance online publication. https://doi.org/10.1080/13670050.2017.1314445

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*, *37*, 391–403.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T., & the CATALISE-2 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, *58*, 1068–1080. https://doi.org/10.1111/jcpp.12721

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 and Communication Disorders*, *43*, 1–40.

Craig, C. S. (2008). Goldwave (Version 5.23) [Computer software]. St. John's, NF, Canada: Goldwave.

Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Real-word and non-word repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, *56*, 323–336.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*, 1136–1146.

Ebert, K. D., Kalanek, J., Cordero, K. N., & Kohnert, K. (2008). Spanish nonword repetition stimuli development and preliminary results. *Communication Disorders Quarterly*, *29*, 67–74.

Ehrler, D. J., & McGhee, R. L. (2008). PTONI: Primary test of nonverbal intelligence. Austin, TX: Pro-Ed.

Ellis Weismer, S., Tomblin, J. B., Zhang, X., Buchwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *43*, 865–878.

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*, 177–195.

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*, 83–94.

Gathercole, S. E., Hitch, G. J., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, *33*, 966–979.

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*, 65–77.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103–127.

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*, 349–367.

Girbau, D., & Schwartz, R. G. (2007). Non-word repetition in Spanish-speaking children with Specific Language Impairment (SLI). *International Journal of Language and Communication Disorders*, *42*, 59–75.

Ho, C. S. H., Leung, M. T., & Cheung, H. (2011). Early difficulties of Chinese preschoolers at familial risk for dyslexia: deficits in oral language, phonological processing skills, and print-related skills. *Dyslexia*, *17*, 143–64.

Kapalková, S., Polišenská, K., & Vicenová, Z. (2013). Non-word repetition performance in Slovak-speaking children with and without SLI:

Novel scoring methods. *International Journal of Language and Communication Disorders*, *48*, 78–89.

Lee, H. J., Kim, Y. T., & Yim, D. (2013). Non-word repetition performance in Korean-English bilingual children. *International Journal of Speech-Language Pathology*, *15*, 375–382.

Lee, S. A. S., & Gorman, B. K. (2013). Nonword repetition performance and related factors in children representing four linguistic groups. *International Journal of Bilingualism*, *17*, 479–495.

Leonard, L. B. (2014). Specific language impairment across languages. *Child Development Perspectives*, *8*, 1–5.

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*, 1033–1047.

Nguyen, D. H. (1997). Vietnamese. Amsterdam, The Netherlands: Benjamins.

Phạm, B., & McLeod, S. (2016). Consonants, vowels and tones across Vietnamese dialects. *International Journal of Speech-Language Pathology*, *18*, 122–134.

Pham, G., Kohnert, K., & Carney, E. (2008). Corpora of Vietnamese texts: Lexical effects of intended audience and publication place. *Behavior Research Methods*, *40*, 154–163. https://doi.org/10.3758/BRM.40.1.154. Available through IRIS digital repository: www.iris-database.org

Radeborg, K., Barthelom, E., Sjöberg, *M*, & Sahlén, B. (2006). A Swedish non-word repetition test for preschool children. *Scandinavian Journal of Psychology*, *47*, 187–192.

Santos, F. H., & Bueno, O. F. A. (2003). Validation of the Brazilian Children's Test of Pseudoword Repetition in Portuguese speakers aged 4 to 10 years. *Brazilian Journal of Medical and Biological Research*, *36*, 1533–1547.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Percentage of Consonants Correct (PCC) metric extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, *40*, 708–722.

Shylaja, K., Abraham, A., Leela Thomas, G., & Swapna, N. (2011). Nonword repetition in simultaneous and sequential bilinguals. *Journal of the All India Institute of Speech and Hearing*, 30, 176–184.

Singh, L., & Fu, C. S. (2016). A new view of language development: The acquisition of lexical tone. *Child Development*, *87*, 834–854.

So, L. K., & Dodd, B. J. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language*, *22*, 473–495.

Sorenson Duncan, T., & Paradis, J. (2016). English language learners' nonword repetition performance: The influence of age, L2 vocabulary size, length of L2 exposure, and L1 phonology. *Journal of Speech, Language, and Hearing Research*, *59*, 39–48.

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research*, *49*, 219–236.

Sundström, S., Samuelsson, C., & Lyxell, B. (2014). Repetition of words and non-words in typically developing children: The role of prosody. *First Language*, *34*, 428–449.

Tang, G. (2007). Cross-linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States. *Journal of Southeast Asian American Education and Advancement*, *2*, 1–33.

Tang, G., & Barlow, J. (2006). Characteristics of sound systems of monolingual Vietnamese-speaking children with phonological impairment. *Clinical Linguistics and Phonetics*, *20*, 423–445.

Thompson, L. C. (1963). The problem of the word in Vietnamese. *Word*, *19*, 39–52.

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders*, *46*, 1–16.

U.S. Census Bureau. (2010). The Vietnamese population in the United States: 2010. Available online.

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*, 481–487. https://doi.org/10.3758/BF03195594

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, *52*, 193–204. https://doi.org/10.1016/j.jml.2004.10.003

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*, 306–311.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*, 59–82.

Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology*, *19*, 298–310.

Wong, A. M. Y., Kidd, J. C., Ho, C. S. H., & Au, T. K. F. (2010). Characterizing the overlap between SLI and dyslexia in Chinese: The role of phonology and beyond. *Scientific Studies of Reading*, *14*, 30–57.

Wong, P., & Strange, W. (2017). Phonetic complexity affects children's Mandarin tone production accuracy in disyllabic words: A perceptual study. *PLoS ONE*, *12*, e0182337. https://doi.org/10.1371/journal.pone.0182337