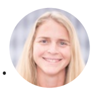
Non-word repetition in 2-year-olds: Replication of an adapted paradigm and a useful methodological extension

Article *in* Clinical Linguistics & Phonetics · April 2015

DOI: 10.3109/02699206.2015.1029594 · Source: PubMed

Rochelle Suzanne Newman
University of Maryland, College P…



Meredith L Rowe
Harvard University

Clinical Linguistics & Phonetics, 2015; Early Online: 1–13

ISSN: 0269-9206 print / 1464-5076 online
DOI: 10.3109/02699206.2015.1029594

informa
healthcare

Non-word repetition in 2-year-olds: Replication of an adapted paradigm and a useful methodological extension

CATHERINE TORRINGTON EATON[1], ROCHELLE S. NEWMAN[2], NAN BERNSTEIN RATNER[2], & MEREDITH L. ROWE[3]

[1]*Department of Communication Sciences and Disorders, Rockhurst University,* [2]*Department of Hearing and Speech Sciences, University of Maryland, and* [3]*Harvard Graduate School of Education, Harvard University, Kansas City, MO, USA*

(*Received 24 August 2014; revised 10 March 2015; accepted 11 March 2015*)

Abstract
Accurate non-word repetition (NWR) has been largely attributed to phonological memory, although the task involves other processes including speech production, which may confound results in toddlers with developing speech production abilities. This study is based on Hoff, Core and Bridges' adapted NWR task, which includes a real-word repetition (RWR) condition. We tested 86 typically developing 2-year-olds and found relationships between NWR and both receptive and expressive vocabulary using a novel measure that controls for speech production by comparing contextually matched targets in RWR. *Post hoc* analyses demonstrated the influence of lexical and sublexical factors in repetition tasks. Overall, results illustrate the importance of controlling for speech production differences in young children and support a useful methodological approach for testing NWR.

Keywords: Articulation, analysis methods, phonological working memory, typically developing, young children

Introduction

The ability to repeat back novel sound patterns, generally referred to as non-word repetition (NWR), has an extensive history in speech-language pathology and the psycholinguistics literature more broadly (Archibald, 2008; Graf Estes, Evans, & Else-Quest, 2007). One consistent finding is that NWR accuracy is strongly related to the size of a child's lexicon, particularly before the age of 4 (Coady & Evans, 2008; Gathercole & Adams, 1993; Jones, Gobet, & Pine, 2007 for review). Gathercole (2006) suggested that this positive correlation was largely based on the shared process of phonological memory, or the ability to temporarily hold verbal information in mind for further processing. In order to learn a new word, the learner needs to hold the sound pattern of the word in memory long enough to form a new representation. As such, phonological memory is important for building these representations; it is used to hold perceptuo-motor programmes in mind while

Correspondence: Catherine Torrington Eaton, Rockhurst University, 407b Van Ackeren Hall, 1100 Rockhurst Road, Kansas City, MO 64151, USA. Tel: 1-816-501-4518. E-mail: catherine.torrington-eaton@rockhurst.edu

creating new representations that are not already stored in long-term memory (Baddeley, Gathercole, & Papagno, 1998; Gathercole, 2006).

NWR has proven to be a very good measure of a constellation of language- and memory-related skills that seem to underlie language acquisition. Poor NWR performance has been implicated in a number of communication disorders including specific language impairment, stuttering, speech sound disorders, dyslexia and hearing impairment (Coady & Evans, 2008; Hakim & Bernstein Ratner, 2004; Ibertsson, Willstedt-Svensson, Radeborg, & Sahlen, 2008; Rispens & Baker, 2012; Shriberg et al., 2009). Although phonological memory may be implicated in these diverse disorders, it is likely that other cognitive–linguistic processes differentially contribute to impaired NWR performance.

A considerable literature in fact has been devoted to evaluating exactly what is measured in NWR. Research has suggested that NWR includes not only phonological memory capacity, but serial memory as well (Majerus, Ponecelet, Greffe, & Van der Linden, 2006). In addition, NWR accuracy is likely affected by more general cognitive factors such as the efficiency of encoding and retrieval processes and the simultaneous processing of both segmental and prosodic information (Marton, 2006; Montgomery & Evans, 2006). It has also been shown that both lexical and sublexical factors including neighbourhood density and phonotactic probability significantly influence NWR performance (Ellis Weismer & Edwards, 2006; Majerus, Van der Linden, Mulder, Meulemans, & Peters, 2004; Metsala & Chisholm, 2010; Vitevitch, 2006; Vitevitch, Luce, Pisoni, & Auer, 1999). In summary, the use of NWR tasks as a proxy of phonological memory should be interpreted with caution because many cognitive and linguistic variables influence NWR performance (Smith, 2006).

Speech production may be a particularly important factor when testing younger participants (i.e. children under 3). For this age group, studies typically include significant numbers of participants who fail to complete the task (Chiat & Roy, 2007; Stokes & Klee, 2009), which could be at least partially accounted for by speech production abilities. It is possible that some children may not attempt the task because their speech production abilities are not well developed. But even in young children who complete the task, differences in speech production abilities pose a potential confound. In a basic NWR task, individuals are given a sound pattern and asked to repeat it back; errors are generally assumed to indicate a failure to successfully encode and remember the sound pattern. But if an individual is unable to produce certain sounds correctly, they could err on this task even if they correctly stored and recalled the item.

To address this potential confound, NWR studies with young children as well as studies with children with speech sound disorders have used a variety of methods to control for speech production differences such as adjusting scoring criteria to accommodate common error patterns (Roy & Chiat, 2004) or designing non-word items with earlier- rather than later-developing sounds (Dollaghan & Campbell, 1998; Shriberg et al., 2009). These approaches affect all children tested, regardless of their actual speech production ability. That is, treating common speech sound error patterns as acceptable responses is something that is done for all children, regardless of whether any particular child makes that error.

Another method of controlling for speech production effects is to use a more individualised approach. Some researchers have analysed children's phonemic inventories using a standardised articulation assessment and then counted as correct those items in the NWR task that were produced with a consistent error (Chiat & Roy, 2007; Stokes & Klee, 2009). An advantage of this methodology is that it assesses an individual child's speech sound production skills and then adjusts scoring criteria specifically for that child. However, collecting and analysing phonemic inventories in order to accurately score NWR items requires an additional, quite time-consuming step in the process, which may be why many studies do not address the potential confound (e.g. Gathercole & Adams, 1993; Rispens & Baker, 2012).

Hoff, Core, and Bridges (2008) designed and tested an adaptation of an NWR task (Gathercole & Adams, 1993) in children aged 21–24 months who also used this individualised method of controlling for speech production ability. Hoff et al.'s task incorporated real-word repetition (RWR) as a control variable for speech production ability. The notion was that repeating real words would depend less on phonological memory than would repeating non-words, but would require the same speech production skills as would NWR. Thus, if children were able to repeat the real words correctly, but failed to repeat a non-word with the same phonemes, this would be indicative of a phonological memory error; if the child repeated both the real word and the non-word incorrectly, this could be indicative of a speech production difficulty. Non-word stimuli were matched phonotactically to real words such that all phonemes occurred in the same word position across both conditions. In their analysis, Hoff et al. factored out speech production abilities by running a partial correlation between NWR and vocabulary percentile with the variance accounted for by RWR removed.

Hoff et al. ran two experiments using their adapted NWR task and analytic approach. Stimuli were elicited in a naturalistic, child-friendly style (i.e. repeating the names of plush toys). Despite minor differences in lists of stimuli, experimental setting and numbers of participants ($n = 15$ and 21, ages 21–24 months), results from both experiments showed correlations with large effect sizes among NWR, RWR and vocabulary. More importantly, when the shared variance from RWR was removed, the partial correlations between NWR and vocabulary were approximately 0.40, which is consistent with results demonstrating this relationship (e.g. Gathercole & Adams, 1993; Stokes, Moran, & George, 2013). However, the correlations only reached significance in one of the two analyses, likely because of the studies' small sample sizes.

The paradigm created by Hoff et al. is useful for a number of reasons. First, it targets a younger population, whereas the majority of NWR protocols are designed for preschool or school-aged children (see Graf Estes et al., 2007 for review). Assessing NWR at an earlier age has the potential to better inform the relationship between vocabulary acquisition and phonological memory throughout development (Chiat, 2006; Gathercole, Tiffany, Briscoe, Thorn, & the ALSPAC team, 2005; Majerus et al. 2006). Second, NWR has considerable potential as an early identification tool for distinguishing late talkers from children with speech or language disorders (Stokes & Klee, 2009). Third, Hoff et al.'s paradigm appears to be a valid measure of phonological memory (and related processes) because the results are consistent with well-established findings showing a strong correlation between NWR accuracy and vocabulary size in children under 4 (Chiat & Roy, 2007). Importantly, this correlation remained even after speech production differences were statistically controlled. Finally, the paradigm is conveniently a ''one-stop-shopping'' assessment; an individual child's speech production can be factored out without additional standardised assessment and analysis.

One important concern raised by Hoff et al. was that statistically removing RWR ability through partial correlation likely eliminated significantly more variance than simply speech production ability. This method may have removed variance from some of the very skills the NWR task is designed to explore, such as children's underlying representations of speech sounds and the contribution of stored lexical items. Although Hoff et al. acknowledged this concern, they were limited by the size of their studies and did not consider alternate analytic methods.

Hoff et al. (2008) designed their stimuli in such a way that it would be possible to create a more fine-grained analysis of the NWR results within a single task. Rather than partialling out shared variation, it would be possible to take a more individualised approach using only those phonemes that children were able to articulate correctly in the real-word stimuli. This might provide a more sensitive measure of NWR because it would in essence control for speech production ability without necessarily removing additional effects of lexical and phonological knowledge. In addition, replication using a larger sample size would allow for a comparison of methodological approaches to refine the NWR measure.

This study is thus an extension of Hoff et al.'s study. One purpose of this study was to replicate the original findings by Hoff et al. with a larger sample size. The more significant goal of this study was a methodological extension exploring a novel approach to removing effects of speech production from the analysis using the single task for eventual clinical utility. We also explored the relationships between NWR and a variety of language measures including receptive and expressive vocabulary and spontaneous language usage. Finally, we examined the influence of lexical and sublexical factors on RWR and NWR in these stimuli.

Participants

The 86 participants (49 female) in this study were part of a larger cohort of monolingual children and their mothers participating in a longitudinal study of language development. Families were recruited through the Infant Studies database at the University of Maryland and were from middle and upper middle-class homes in which all but eight mothers had an associate degree or higher. The participants ranged in age from 23 to 25 months (mean age = 24.06, SD = 0.44).

Thirty-nine (15 female) out of the original group of 125 participants were not included in the study. The dropped cohort included nine children whose scores were below the 10th percentile on the *The Macarthur-Bates Communicative Development Inventory: Words and Sentences* (MCDI) (Fenson et al., 2006), nine children who failed to repeat at least 66% (16 out of 24) of the items and 21 participants who were completely non-compliant with the task. Further discussion regarding excluded participants will be addressed in following sections.

Procedure

The experimental task was part of a single 3-h testing session composed of an extensive battery of both standardised and non-standardised assessments. The NWR task was generally administered in the second half of the session. Similar to methods described in Hoff et al. (2008), children were handed individual toys and asked to repeat their labels (e.g. ''This is a dog. Can you say 'dog'''?). For the non-words, the experimenter handed the child a series of brightly coloured stuffed animals that resembled penguins, asked the child to repeat the ''funny'' name, and then encouraged the child to put the toy in a box either for a party or for bedtime (depending on which context more consistently elicited responses from the child). The list of real words was always administered before the non-words. Each real word or non-word was produced a maximum of two times by the experimenter before moving to the next item. Any item that the child failed to repeat after two presentations was counted as inaccurate in the analysis.

The *Peabody Picture Vocabulary Test, fourth edition* (PPVT; Dunn & Dunn, 2007), a measure of receptive vocabulary, and the *Expressive One-Word Picture Vocabulary Test* (EOWPVT; Brownell, 2000) were administered before the NWR task, but in the same session. For the final activity of the study, mothers played with their children using a pre-designated array of toys for the purpose of eliciting a 10-min conversational speech sample. All sessions were both audio- and video-recorded. The MCDI (Fenson et al. 2006) was most often completed by the family prior to the session, although a few mothers filled out paperwork during the session. Responses for the MCDI, EOWPVT and PPVT were coded independently by two laboratory members. A third coder resolved any scoring discrepancies (<5%). Mother–child play sessions were transcribed in the freeware programme Codes for the Human Analysis of Transcripts (CHAT; MacWhinney, 2000). Measures of expressive language and spontaneous vocabulary usage spoken by the child were derived using the analysis programme Computerized Language Analysis (CLAN; MacWhinney, 2000).

Stimuli

The stimuli listed in Appendix 1 were obtained from Hoff, Core, and Bridges (personal communication, 12/1/2009) and were an updated version of the Hoff et al.'s (2008) studies. Stimuli consisted of four one-syllable, four two-syllable and three three-syllable real words and their matched non-words. Real words and non-words were matched so as to have the exact same consonants and vowel–consonant pairs, only in different combinations (e.g. real words: ''cookie'' and ''puppy''; non-words: ''pookie'' and ''kuppy''). Consequently, real words and non-words were matched for positional phoneme probability (although not necessarily biphone probability), lexical stress and phoneme difficulty. Later-developing sounds, such as consonant clusters and the phoneme /r/, were not included in the updated stimuli because these sounds are commonly produced in error by 2-year-olds.

Analysis

For ease of scoring individual items, real words and non-words were transcribed and then linked to their respective audio files using CHAT (MacWhinney, 2000). All standardised assessment tools were scored independently by 2 of 4 graduate students familiar with phonetic transcription. Only consonants were coded as correct or incorrect and entered into a spreadsheet. In accordance with the majority of literature on phoneme accuracy (e.g. Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997), vowels were not included in Hoff et al.'s study or this study because of variability in scorers' perceptual judgements. All discrepancies between coders were juried and resolved by consensus similar to procedures recommended in Shriberg, Kwiatkowski, and Hoffman (1984). Items that the child did not attempt to repeat after the scripted elicitation, regardless of the reason, were counted as inaccurate in the analysis (6.50% of the data). Items that were too difficult to code due to noise were excluded (0.36% of the data).

In order to explore different methods of analysing the data, two calculations were made for each participant. The first method, which was used in Hoff et al. (2008), was to calculate percentage consonants correct (PCC; Shriberg & Kwiatkowski, 1982). Two PCC values were derived for each participant by dividing the number of accurate consonants produced by the total number of consonant targets for each of the real-word and non-word conditions. As is typical in the NWR literature with young children, consistent sound substitutions found in both real-word and matched non-word targets were not counted as errors (Chiat & Roy, 2007; Stokes & Klee, 2009).

For the second method, an item analysis was conducted such that each of the 33 non-word phoneme targets was compared to its corresponding real-word target (i.e. same word position). Two columns were created for this calculation. In the first column, non-word targets were assigned one point if they were produced correctly and the matched real-word target was also correct. Non-word phoneme substitutions that were the same in both non-word and real-word contexts were also counted as correct (e.g. /p/ for /k/ in word-initial position). If either the non-word target or the corresponding real-word target were incorrect (or not produced), zero points were awarded for that phoneme. In the second column, real-word targets were assigned one point if they were produced correctly or if they were incorrect but matched the non-word target (thus indicating a consistent phoneme substitution). After each column was summed, column one was divided by column two for each participant. Thus, the dependent measure was the proportion of accurate real-word consonants that were produced accurately in matched non-word contexts. This measure will be referred as adjusted non-word accuracy (NW accuracy).

Two measures from spontaneous language samples, mean length of utterance (MLU) and vocabulary diversity (VOCD), were derived using CLAN (MacWhinney, 2000). VOCD is an algorithm that is less influenced by quantity (e.g. verbose vs. taciturn children) than is type–token ratio, but its accuracy is unreliable on samples of fewer than 50 utterances (MacWhinney, 2000). Therefore, only 70 transcripts were included in the VOCD analysis.

To explore the influences of lexical and sublexical factors on RWR and NWR in Hoff et al.'s stimuli, we first scored each item as either completely accurate (e.g. all phonemes were present and produced correctly, or were produced with a consistent articulation error in both the real-word and non-word items) or inaccurate (some portion of the item was in error in a manner inconsistent across productions). Items were then subdivided into those that were either ''high'' or ''low'' in terms of phonotactic probability and neighbourhood density.

For phonotactic probability, we measured the likelihood of each phoneme occurring in that position in a word and summed these values (see, e.g. Jusczyk, Luce, & Charles-Luce, 1994; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). This was done using an online calculator (Vitevitch & Luce, 2004). This act of summing leads to a clear length-dependence; the sum of phoneme probabilities is inevitably greater when the word is longer. To address this, we rank-ordered the items within sequences of a given length; thus, among the three-phoneme items (e.g. cat and dog), the real words and non-words with the highest phonotactic probability were assigned to the ''high'' group, and those with lower phonotactic probability were assigned to the ''low'' group. This was done separately for real words and non-words at each sequence length. Finally, we calculated the accuracy for each set of items for each participant.

Similar methods were employed based on lexical neighbourhood density (defined as the number of words in the lexicon that differed from the target by a single phoneme addition, deletion or substitution, in any location in the word). Here, too, length was a potential confound, in that shorter words have many more neighbours. Moreover, there was an additional concern regarding length and neighbourhood density; the longer items in each real-word or non-word set had the same number of neighbours, zero – as such, they could not have ''high'' density at all. We therefore selected only a subset of items for this analysis, including only 10 of the 24 items.

Results

Descriptive statistics

Scores for all assessment tools and vocabulary measures are included in Table 1. The mean raw score for the MCDI was 358.45 (SD 147.24) and the mean percentile was 54.95 (SD 26.02). Although Hoff et al. used MCDI percentiles in their analyses, it seemed more appropriate to use the true size of the lexicon, or raw score, to address the research questions rather than the norm-referenced score. For consistency, all vocabulary measures from standardised assessments were analysed using raw scores.

Table 1. Descriptive statistics for test participants: MCDI, PPVT and EOWPVT raw scores and spontaneous language measures MLU and VOC-D.

	Mean	SD	Range
MCDI	358.45	147.25	62–664
PPVT	32.90	12.26	12–60
EOWPVT	18.53	6.61	0–32
MLU	1.84	0.47	1.2–3.0
VOC-D	31.96	17.33	6.9–88.2
PCC RW	60.06	19.09	18.2–90.9
PCC NW	49.58	17.83	12.1–97.0
Adjusted NW accuracy	0.65	0.16	0.3–1.0

VOC-D values are for 70 of 86 participants. RW, real word; NW, non-word; PCC, percentage consonant correct.

One potential question that we had was whether toddlers who are non-compliant with NWR tasks are qualitatively different from those who complete the task. Table 2 shows test performance for the 39 participants who did not complete the task (or at least 66% of the task). Of this group, nine children had raw scores of <50 on the MCDI and/or <5 on the EOWPVT suggesting that these children were late talkers who were at risk for language impairment (Stokes & Klee, 2009). As can be seen from the data, overall children in the non-completion group demonstrated lower vocabulary raw scores from the completion group (MCDI: $t(122)=4.55$, $p<0.001$; $d=0.87$; PPVT: $t(122)=4.93$, $p<0.001$; $d=0.97$; EOWPVT: $t(121)=5.79$, $p<0.001$, $d=0.97$). There were no significant between-group differences, however, in spontaneous language measures (MLU morphemes ($t(120)=0.33$, $p=0.74$); VOC-D ($t(98)=1.460$, $p=0.147$)). There were also no statistically significant differences between demographic variables including the number of males versus females or the level of maternal education (gender: $t(123)=1.741$, $p=0.08$; maternal education: $t(123)=1.84$, $p>0.07$).

Table 2. Descriptive statistics for non-completion participants.

	Mean	SD	Range
MCDI	219.15	159.06	0–589
PPVT	21.82	10.15	8–64
EOWPVT	10.78	7.30	0–27
MLU	1.79	0.54	1.11–3.2
VOC-D	32.05	18.66	9.9–87.4

VOC-D values are for 32 of 39 participants.

Replication of Hoff et al. with a larger sample

We next asked whether the original results could be replicated with a much larger data set. Similar to Hoff et al. (2008), children showed a significantly greater percentage accuracy when repeating real words ($t(86)=7.094$, $p<0.001$; $d=0.57$). Sixty-seven of the 86 children showed a greater accuracy when repeating real words than non-words. As seen in Table 3, there were significant correlations between all three vocabulary raw scores and PCC in real words (effect sizes: MCDI: $r^2=0.431$; PPVT: $r^2=0.110$; EOWPVT: $r^2=0.319$) and PCC in non-words (effect sizes: MCDI: $r^2=0.327$; PPVT: $r^2=0.130$; EOWPVT: $r^2=0.266$). There was also a significant correlation between non-word and real-word PCC ($r^2=0.527$).

The second analysis that was conducted by Hoff et al. (2008) was a partial correlation between NWR PCC and vocabulary size with RWR PCC used as a control variable. Again, the purpose of running this analysis, as argued by Hoff et al., was to remove the variance caused by articulatory ability to obtain a more representative measure of phonological memory. In contrast to Hoff et al., we found a much smaller correlation between NWR PCC and vocabulary size, as measured using the MCDI raw score, when RWR PCC was a controlled variable ($r=0.188$, $p=0.086$, $r^2=0.035$). The effect size indicates that only 3.5% of the variance in NWR PCC was explained by vocabulary scores when RWR PCC was controlled. The other two vocabulary measures showed a similar trend in which partial correlations were not statistically significant (EOWPVT: $r=0.189$, $p=0.085$, $r^2=0.036$; PPVT: $r=0.187$, $p=0.089$; $r^2=0.035$).

Alternate analysis to remove articulatory effects

The second part of this study examined an alternate measure using Hoff et al.'s protocol, which factored out speech production errors for each individual child. Results (in Table 3) show that the

Table 3. Correlation matrix for all NWR, vocabulary and language measures.

	1	2	3	4	5	6	7	8
1 PCC NW	–	0.726**	0.784**	0.572**	0.361*	0.516**	0.454**	0.343*
2 PCC RW	0.726**	–	0.315*	0.657**	0.332*	0.565**	0.494**	0.419**
3 adj-NW	0.784**	0.315*	–	0.257*	0.294*	0.314*	0.177	0.128
4 MCDI	0.572**	0.657**	0.257*	–	0.397**	0.488**	0.521**	0.401**
5 PPVT-4	0.361*	0.332*	0.257*	0.397**	–	0.627**	0.216†	0.161
6 EOWPVT	0.516**	0.565**	0.314*	0.488**	0.627**	–	0.354	0.278*
7 MLU	0.454**	0.494**	0.177	0.521**	0.216†	0.354**	–	0.504**
8 VOC-D	0.343*	0.419**	0.128	0.401**	0.161	0.278*	0.504**	–

Significant at †$p<0.05$; *$p<0.01$; **$p<0.001$.

adjusted NW accuracy measure was significantly correlated with all three vocabulary raw scores (effect sizes: MCDI: $r^2=0.066$; PPVT: $r^2=0.086$; EOWPVT: $r^2=0.099$). It was also strongly correlated with non-word PCC, as would be expected given that both measures are a reflection of non-word production accuracy (effect size: $r^2=0.614$). These results suggest that the individualised method of controlling for speech production provides a better measure of NWR performance than using a partial correlation. The small-to-moderate effect sizes between accuracy and vocabulary scores will be considered in the discussion.

As seen in Table 3, there were statistically significant relationships between non-word PCC and the two spontaneous language measures (effect sizes: VOCD: $r^2=0.188$; MLU: $r^2=0.201$). These relationships diminished considerably when speech production was controlled using the adjusted NW accuracy measure (effect sizes: VOCD: $r^2=0.016$; MLU: $r^2=0.031$). The implications of these results will be addressed in the ''discussion'' section.

Sublexical and lexical effects

In light of the non-significant partial correlation, it seemed particularly appropriate to consider sublexical and lexical factors that might additionally affect RWR and NWR performance. To examine the influence of sublexical factors, we ran a 2 (phonotactic probability: high vs. low) × 2 (item type: word vs. non-word) repeated-measures ANOVA. Overall, there was a significant effect of phonotactics, $F(1,85)=22.02$, $p<0.0001$, partial eta-squared = 0.21; children were more accurate at naming items with more common phonotactic patterns. There was also a significant effect of item, $F(1,85)=10.83$, $p<0.002$, partial eta-squared = 0.11, with children more accurate at naming real words than non-words. Critically, there was no interaction, $F(1,85)<1$, $p>0.50$, partial eta-squared = 0.004, suggesting that the effect of phonotatic probability influenced naming equally in both real words and non-words. Despite the non-significant interaction, we looked at the real words and non-words separately as planned comparisons – there was a significant effect of phonotactic probability among the real-word items ($t(85)=2.90$, $p<0.005$, $r=0.30$), but only a marginal effect among the non-words ($t(85)=1.78$, $p<0.08$, $r=0.190$). Thus, it is possible that sublexical effects may be somewhat more variable among non-words than real words; still, the general pattern suggests that even when repeating non-words, children are accessing sublexical knowledge of the language.

A similar analysis was performed based on lexical neighbourhood density. The overall results suggested that children were significantly more accurate at naming items with *few* lexical neighbours ($F(1,85)=21.54$, $p<0.0001$, partial eta-squared = 0.20). They were again more accurate with real words than non-words ($F(1,85)=7.13$, $p<0.01$, partial eta-squared = 0.08), and there was no interaction ($F < 1$, $p > 0.40$, partial eta-squared $= 0.007$). Planned comparisons showed a strong effect of neighbourhood density among the real words, [$t(85) = 2.58$, $p < 0.02$, $r = 0.27$], but only a weak trend among the non-words ($t(85) = 1.11$, $p = 0.272$, $r = 0.12$).

Discussion

This study replicated effects found by Hoff et al. with a considerably larger sample of 2-year-olds. Results showed a strong relationship between NWR and vocabulary size, which is consistent with what has been found in the literature on phonological memory in young children (e.g. Chiat & Roy, 2007). Unlike Hoff et al., however, this relationship was weakened when the variance from RWR was partialled out. An alternate measure, which calculated accuracy of non-word phonemes based solely on the phonemes that were produced correctly in the same word position in real words, provided a more fine-grained measure of task performance. This novel measure of NWR was correlated with parental-reported vocabulary size and expressive and receptive vocabulary raw scores. There was little relationship between measures of lexical diversity or grammatical complexity in spontaneous language and the ability to repeat non-words when speech production was controlled using the adjusted NW accuracy measure. Finally, there were significant lexical and sublexical effects found across RWR and NWR; items with higher phonotactic probability and fewer phonological neighbours were repeated more accurately.

Other NWR paradigms for this population include an RWR condition (Chiat & Roy, 2007); however, to our knowledge, this condition has not previously been used for the purposes of controlling for speech production, but rather to demonstrate that real words are produced with greater accuracy than non-words. While results of this study are consistent with the lexicality effect reported in the previous literature comparing RWR to NWR accuracy (e.g. Conlin & Gathercole, 2006; Majerus & Boukebza, 2013; Vitevitch et al., 1999), Hoff et al.'s stimuli are advantageously designed to allow phonemes in non-words to be compared directly to phonemes in real words within the same context (word-initial, -medial or -final). More specifically, phoneme location effects and lexical stress have been shown to significantly influence phonological memory (Burgess & Hitch, 2006; Majerus et al., 2006; Marton, 2006; Roy & Chiat, 2004).

Other cognitive and linguistic factors might also help explain the non-significant partial correlation between NWR and MCDI raw scores found with this larger data set as compared to Hoff et al.'s results. As noted previously, NWR involves not only phonological memory, but also facilitation from lexical items stored in long-term memory, influences from phonotactic frequency, encoding and retrieval processes, phonological processing, motor planning and articulation (Montgomery & Evans, 2006; Munson, Swenson, & Manthei, 2005; Thorn & Frankish, 2005; Vitevitch et al., 1999). Controlling for speech production by partialling out RWR does not result in a true measure of phonological memory because it likely removes too much of these underlying processes. That is, ''phonological memory'' depends, at least in part, on one's existing knowledge of the language, which includes both lexical and sublexical processes (Vitevitch, 2006).

Results from the *post hoc* analyses examining sublexical influences were in accord with the prior literature; Vitevitch and colleagues reported that phonotactic effects are facilitative (Vitevitch et al., 1997). That is, it is easier to produce items with common sound patterns. In terms of lexical influences, results from this study indicated increased accuracy in both real words and non-words from sparser neighbourhoods. Although contrary to evidence in adults, which shows faster response times when naming items from dense lexical neighbourhoods (Vitevitch, 2002), the current results are consistent with some studies of word learning in children indicating an advantage of sparse neighbourhoods (Storkel, Maekawa, & Hoover, 2010, but see Storkel, 2004).

It is important to point out that the items in this study were based on those from Hoff et al. and were not selected to include variation in lexical or sublexical properties *per se*. As such, these results must be considered merely exploratory. An additional caveat concerns the fact that the neighbourhood density analysis excluded more than half of all items due to length effects; Metsala and Chisholm (2010) found a significant difference between density effects on repetition of two- versus three- and four-syllable items. The potential for spurious effects based on excluded items was considerable, and so results from this analysis should be approached with caution. Nonetheless, the pattern of findings suggests that children rely on previously stored knowledge when asked to repeat items, and that neither RWR nor NWR is likely to be divorced from the child's lexical (and sublexical) knowledge. As researchers work to define the complex relationship between phonological memory and vocabulary acquisition throughout language development, this study provides a more refined alternative to removing only the variance from speech production abilities in NWR in toddlers (Chiat, 2006; Gathercole et al., 2005).

The comparison between the adjusted NW accuracy measure and non-word PCC was also demonstrated in relation to spontaneous language measures. Although NWR and lexical diversity and morphology usage appeared to be related, these relationships were considerably diminished once speech production was more effectively factored out. These results again highlight the importance of using appropriate methods to control for speech production skills in young children in order to avoid misinterpreting perceived relationships.

It should be noted that the effect sizes between the adjusted NW accuracy measure and the vocabulary measures though significant, are not large especially in comparison to the moderate effect sizes found by others (Chiat & Roy, 2007; Stokes et al., 2013). These studies with young children also reportedly controlled for articulatory differences by taking participants' consistent patterns of substitution into account. One possible explanation for the discrepancy is that this study counted all omissions as errors rather than exclusions. Children with smaller vocabularies may have declined to produce the items as a result of poor long-term storage rather than poor working memory, although this would have artificially enhanced rather than diminished the relationship. Although not reported for the sake of brevity, we did re-run the analysis using the alternative method of excluding items that children failed to repeat, and results did not significantly differ. Yet another explanation for the smaller effect sizes may be the nature of the stimuli; it is possible that neighbourhood densities and phonotactic frequencies interacted differently in this list of stimuli compared to lists used in other studies.

The efficacy of this paradigm can be contrasted with Gathercole and Adams' (1993) attempt to administer the standard NWR task to 3-year-olds, which resulted in a >50% rate of non-completion. The 30% non-completion rate in this study might also appear excessive, but other NWR studies report non-compliance rates from 11% to 25% in children younger than 30 months (Chiat & Roy, 2007; Hoff et al., 2008; Polisenska & Kapalkova, 2014; Roy & Chiat, 2004; Stokes & Klee, 2009). In these studies, NWR is one of only a few experimental tasks, in contrast to this experimental task that was part of a 3-h battery of assessments, and thus very demanding for toddlers. Studies also vary considerably on their criteria for excluding participants by requiring completion of all items or by not using standardised assessments to exclude children with delayed language (Roy & Chiat, 2004; Stokes & Klee, 2009).

There is disagreement among researchers on whether non-completion indicates a child's reluctance to participate because the task is beyond his/her linguistic or sensorimotor capability or is better explained by reticence on the part of the child (Polisenska & Kapalkova, 2014). Another possibility is that NWR compliance is affected by the size of the child's phonetic inventories; children may avoid non-words with phones that are not within their phonetic inventories (Schwartz, Leonard, Frome Loeb, & Swanson, 1987). The participants in this study who completed the task significantly out-performed the non-compliant group on vocabulary measures, but the non-compliant group also consisted of at least nine children who were at risk for language disorder. It seems likely that there are subgroups within the larger group of non-compliant participants. Future research should devote more attention to this issue if this tool is to be clinically useful for distinguishing typical from delayed development in early childhood (Stokes & Klee, 2009).

In conclusion, results of this study suggest that the NWR paradigm introduced by Hoff et al. combined with the improved analysis method employed here offers an important methodological contribution by efficiently controlling for speech production in young language learners. The results further demonstrate the consequences of not effectively controlling for speech production. When speech production ability is accurately controlled for, children's ability to repeat non-words seems to indeed capture linguistic skills that underlie language acquisition, as shown by significant correlations between NWR and a variety of vocabulary measures. Results also illustrate the risk of over-controlling for articulation by removing effects of lexical and sublexical knowledge that are important components in children's ability to repeat speech sequences.

Acknowledgments

The authors particularly thank Lisa Tuit, who helped standardise the testing procedures, Sean Hendricks and Amelie Bail, who conducted much of the testing and coding of the NWR data and clinical assessments, and Devon Brunson, who performed most of the scheduling. The authors also thank the following students for assistance in recruiting and scheduling families: Katrina Ablorh, Candace Ali, Saher Ali, Alison Arnold, Megan Askew, Catherine Bender, Taryn Bipat, Alyssa Cook, Jennifer Coon, Justine Dombroski, Sara Dougherty, Sara Edelberg, Daniel Eisenberg, Lauren Evans, Andrea Farina, Lauren Fischer, Arielle Gandee, Richard Garcia, Laura Horowitz, Megan Janssen, Mina Javid, Amanda Jensen, Esther Kim, Stephanie Lee, Perri Lieberman, Rachel Lieberman, Eileen McLaughlin, Kelly McPherson, Giovanna Morini, Vidda Moussavi, Molly Nasuta, Courtenay O'Connor, Sabrina Panza, Elise Perkins, Amanda Pasquarella, Lauren Polovoy, Rachel Ports, Rachel Rhodes, Allie Rodriguez, Maria Rodriguez, Judith Segal, Katie Shniderman, Veronica Son, Mara Steinberg, Justine Taweel, Allison Temple, Dena Tran, Eugene Vassilas, Krista Voelmle, Chelsea Vogel. Amanda Wildman, Cavena Williams, Kimmie Wilson, Catherine Wu, Donna Zack-Williams and Michelle Zobel.

Declaration of interest

This work was supported by NSF grant BCS 0745412 to the University of Maryland.

The authors report no conflicts of interest.

References

Archibald, L. (2008). The promise of nonword repetition as a clinical tool. *Canadian Journal of Speech-Language Pathology and Audiology*, *32*, 21–28.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*, 158–173.

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary* (3rd ed.). Novato, CA: Academic Therapy Publications.

Burgess, N., & Hitch, G. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language*, *55*, 627–652.

Chiat, S. (2006). The developmental trajectory of nonword repetition. *Applied Psycholinguistics*, *27*, 552–556.

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*, 429–443.

Coady, J., & Evans, J. (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.

Conlin, J., & Gathercole, S. (2006). Lexicality and interference in working memory in children and in adults. *Journal of Memory and Language*, *55*, 363–380.

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

Dunn, L., & Dunn, D. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Circle Pines, MN: American Guidance Service Publishing/Pearson Assessments.

Ellis Weismer, S., & Edwards, S. (2006). The role of phonological storage deficits in specific language impairment: A reconsideration. *Applied Psycholinguistics*, *27*, 556–562.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, S., & Bates, E. (2006). *The MacArthur Communicative Development Inventories* (2nd ed.). Baltimore, MD: Brookes.

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

Gathercole, S., & Adams, A. (1993). Phonological working memory in very young children. *Developmental Psychology*, *29*, 770–778.

Gathercole, S., Tiffany, C., Briscoe, J., Thorn, A., & the ALSPAC team (2005). Developmental consequences of poor phonological short-term memory function in childhood: A longitudinal study. *Journal of Child Psychology and Psychiatry*, *46*, 598–611.

Graf Estes, K., Evans, J., & Else-Quest, N. (2007). Differences in nonword repetitions performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*, 177–195.

Hakim, H., & Bernstein Ratner, N. (2004). Nonword repetition abilities of children who stutter: An exploratory study. *Journal of Fluency Disorders*, *29*, 179–199.

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

Ibertsson, T., Willstedt-Svensson, U., Radeborg, K., & Sahlen, B. (2008). A methodological contribution to the assessment of nonword repetition – a comparison between children with specific language impairment and hearing-impaired children with hearing aids or cochlear implants. *Logopedics Phoniatrics Vocology*, *33*, 168–178.

Jones, G., Gobet, F., & Pine, J. (2007). Linking working memory and long-term memory: A computational model of the learning of new words. *Developmental Science*, *10*, 853–873.

Jusczyk, P., Luce, W., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Majerus, S., & Boukebza, C. (2013). Short-term memory for serial order supports vocabulary development: New evidence from a novel word learning paradigm. *Journal of Experimental Child Psychology*, *116*, 811–828.

Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, *51*, 297–306.

Majerus, S., Poncelet, M., Greffe, C., & Van der Linden, M. (2006). Relations between vocabulary development and verbal short-term memory: The relative importance of short-term memory for serial order and item information. *Journal of Experimental Child Psychology*, *93*, 95–119.

Marton, K. (2006). Do repetition errors in children with specific language impairment reflect a weakness in an unidentified skill specific to nonword repetition or a deficit in simultaneous processing? *Applied Psycholinguistics*, *27*, 569–573

Metsala, J., & Chisholm, G. (2010). The influence of lexical status and neighborhood density on children's nonword repetition. *Applied Psycholinguistics*, *31*, 489–506.

Montgomery, J., & Evans, J. (2006). Commentary on keynote. *Applied Psycholinguistics*, *27*, 573–577.

Munson, B., Swenson, C., & Manthei, S. (2005). Lexical and phonological organization in children: Evidence from repetition tasks. *Journal of Speech, Language, and Hearing Research*, *48*, 108–124.

Polisenska, K., & Kapalkova, S. (2014). Improving child compliance on a computer-administered nonword repetition task. *Journal of Speech, Language, and Hearing Research*, *57*, 1060–1068.

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research*, *55*, 683–694.

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

Schwartz, R., Leonard, L., Frome Loeb, D., & Swanson, L. (1987). Attempted sounds are sometimes not: An expended view of phonological selection and avoidance. *Journal of Child Language*, *14*, 411–418.

Shriberg, L., Austin, D., Lewis, B., McSweeny, J., & Wilson, D. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language and Hearing Research*, *40*, 708–722.

Shriberg, L., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, *47*, 256–270.

Shriberg, L., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, *27*, 456–465.

Shriberg, L., Lohmeier, H., Campbell, T., Dollaghan, C., Green, J., & Moore, C. (2009). A nonword repetition task for speakers with misarticulations: The syllable repetition task (SRT). *Journal of Speech, Language and Hearing Research*, *52*, 1189–1212.

Smith, B. (2006). Precautions regarding nonword repetition tasks. *Applied Psycholinguistics*, *27*, 584–587.

Storkel, H. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*, 201–221.

Storkel, H., Maekawa, J., & Hoover, J. (2010). Differentiating the effects of phonotactic probability and neighborhood density on vocabulary comprehension and production: A comparison of preschool children with versus without phonological delays. *Journal of Speech, Language and Hearing Research*, *53*, 933–949.

Stokes, S., & Klee, T. (2009). The diagnostic accuracy of a new test of early nonword repetition for differentiating late talking and typically developing children. *Journal of Speech, Language and Hearing Research*, *52*, 872–882.

Stokes, S., Moran, C., & George, A. (2013). Nonword repetition and vocabulary use intoddlers. *Topics in Language Disorders*, *33*, 224–237.

Thorn, A., & Frankish, C. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 729–735.

Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 735–747.

Vitevitch, M. (2006). Manipulating the characteristics of words and nonwords to better understand word learning. *Applied Psycholinguistics*, *27*, 594–598.

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, and Computers*, *36*, 481–487.

Vitevitch, M., Luce, P., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*, 47–62.

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

Appendix 1

dog – kog
juice – buice
cat – jat
book – dook
balloon – challoon
cookie – pookie
puppy – kuppy
chicken – bicken
banana – baJApop
telephone – teLIna
lollipop – LOlamas
pajamas – paNAphone