The Diagnostic Accuracy of a New Test of Early Nonword Repetition for Differentiating Late Talking and Typically Developing Children

Stephanie F. Stokes
Curtin University, Perth, Australia

Thomas Klee
Newcastle University, United Kingdom

Purpose: To assess the diagnostic accuracy of a new Test of Early Nonword Repetition (TENR) for 2-year-old children.
Method: 232 British-English-speaking children aged 27 (±3) months were assessed on 3 standardized tests (receptive and expressive vocabulary and visual processing) and a novel nonword repetition (NWR) test. Parents completed a British adaptation of the MacArthur-Bates Communicative Development Inventory: Words and Sentences (CDI:WS–UK; Klee & Harrison, 2001). The diagnostic accuracy of two versions (1–3 syllables and 1–4 syllables) of a new NWR test was examined. Standard diagnostic accuracy measures of sensitivity, specificity, positive and negative likelihood ratios, and diagnostic odds ratios were generated.
Results: 177 children (80%) completed the 1–3 syllable task, and 96 children (73%) completed the 1–4 syllable task. The 1–3 syllable version produced a positive likelihood ratio (LR+) of 7.8 (confidence interval [CI] = 4.5–13.6) and a negative likelihood ratio (LR–) of .28 (CI = .12–.65). The 1–4 syllable version of the NWR test produced a LR+ of 14.88 (CI = 6.1–36.2) and a LR– of .13 (CI = .02–.83).
Conclusion: The TENR could be useful for identifying 2-year-old children at risk of language impairment.

KEY WORDS: nonword repetition, late talkers, language impairment, diagnostic accuracy, sensitivity and specificity

In a nonword repetition (NWR) task, the participant imitates a series of nonwords that usually range from one to four syllables in length and that meet the criterion of being as nonwordlike as possible (e.g., *mot, woogelamik*). Beyond those criteria, tests differ widely in the number of syllables included, the use of singleton versus clustered consonants, stress patterns, and tense and lax vowels (Archibald & Gathercole, 2006). Stimuli are presented to children for repetition in the form of live voice (e.g., Adams & Gathercole, 1995; Bishop, North, & Donlan, 1996; Roy & Chiat, 2004) or recorded samples (e.g., Campbell, Dollaghan, Needleman, & Janosky 1997). Scoring is one of two methods, either 1 point for each correct whole nonword or 1 point for each correct segment (consonant or vowel) repeated. NWR is believed to require psycholinguistic processing that does not recruit information from a stored mental lexicon and is therefore considered to be an unbiased assessment of language-related ability (for detailed explanations, see Dollaghan & Campbell, 1998; Gathercole, 2006).

Regardless of differences in test construction and administration, over the last 2 decades various studies have reported the significant

deficit that children with language impairment show in comparison with their age-matched peers or younger language-matched typically developing (TD) children on NWR (for a review, see Coady & Evans, 2008; Graf-Estes, Evans, & Else-Quest, 2007). Graf-Estes et al. explored the size of the NWR deficit in children with specific language impairment (SLI) by screening 60 published and unpublished studies for possible inclusion in a meta-analysis of NWR performance in children with and without SLI. Their meta-analysis of the 23 studies that met their inclusion criteria revealed that children with SLI, on average, performed at 1.27 standard deviations below the mean score of children without SLI (TD children). However, they also reported that four different versions of NWR tests[1] yielded different effect sizes in group comparisons and were thus not interchangeable. Nonetheless, there was no relationship between effect size and the age of the children with SLI, and children with SLI were significantly worse than both their age-matched peers and younger language-matched children at repeating even one-syllable nonwords, not just longer strings of syllables.

These findings of statistically significant differences in mean scores for groups identified as SLI and TD have led researchers to suggest that NWR could be used as a clinical marker of SLI. The value of this is that NWR is considered to be equally familiar (or unfamiliar) to all children and therefore free of cultural bias in identifying impairment, unlike standardized language tests (Montgomery, 2002). In recent years, authors have also begun to suggest that NWR could be used as an indicator of early language delay. Chiat and Roy (2007), Roy and Chiat (2004), and Stokes and Klee (2009) reported the use of repetition tests in children as young as 2 years, with the former two combining scores on word and nonword repetition and the latter using only NWR. In addition, Chiat and Roy (2008) reported the use of a repetition test as a predictor of later language outcome, following up children from 2;0–2;6 (years;months) to 3;6–4;0. All three of these studies of NWR in very young children reported results for regression analyses that demonstrate the high proportion of variance accounted for in language scores by NWR (about 38%).

Thus, at least 2 decades of research has culminated in the suggestion that NWR could be used as a clinical marker of language impairment in preschool and school-aged children or as a predictor of early language delay. However, in the search for a clinical marker of SLI, it is necessary to go beyond studies of mean group differences (i.e., pre-accuracy studies) of NWR. Several studies have reported the sensitivity and specificity of NWR in classifying children previously identified as SLI or TD. Sensitivity is the probability of a fail score on an index test (e.g., NWR) among those with the condition, showing the value of the test in detecting impairment, whereas specificity is the probability of a pass score on NWR among participants without the condition, showing the value of the test in identifying children without language impairment as TD (see Dollaghan, 2007, for a summary of the model and terms used). There is some variation in what is acceptable in sensitivity and specificity values, from 70% (Glascoe & Squires, 2007) to 80% (Meisels, 1988) and 90% (Plante & Vance, 1994), but 80% is generally accepted as the minimum level.

In recent years, sensitivity and specificity values have been supplemented by the use of likelihood ratios (LRs; for detailed explanations, see Dollaghan, 2007; Klee, 2008). Dollaghan (2007) has described why LRs are more robust in evaluating the diagnostic accuracy of a test than just sensitivity and specificity measures, and the reasons will not be repeated here. A positive likelihood ratio (LR+) is calculated from [sensitivity/(1–specificity)] and indicates the likelihood that a fail score on a test such as NWR came from a child with SLI rather than a TD child. A negative likelihood ratio (LR–) is calculated from [(1–sensitivity)/specificity)] and indicates the likelihood that a pass score on the test came from a child with SLI. In using LRs to diagnose individuals, a LR+ of $\geq$ 10.0 and a LR– of $\leq$ .10 are considered desirable (McAlister, Straus, & Sackett 1999). If a test demonstrated these levels of diagnostic accuracy in tandem with reasonable confidence intervals (CIs), we could conclude that a child with SLI was at least 10 times more likely to achieve a fail score than a TD child and only .10 times as likely to achieve a pass score than a TD child.

To facilitate a comparison among tests in terms of their diagnostic accuracy, the diagnostic odds ratio (DOR) may be used (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003). A DOR is "the ratio of the odds of positivity in disease relative to the odds of positivity in the nondiseased" (Glas et al., 2003, p. 1130) and may be computed as LR+/LR–. DOR values range from zero to infinity, with a value of 1 indicating no value of the diagnostic test in discriminating between affected and unaffected individuals. As LR+ increases and LR– decreases, DOR increases. The higher the DOR, the better the diagnostic test is in identifying the true status of affected and unaffected individuals. Note, however, that one still needs to know how many over- and underreferrals a diagnostic test will yield. Thus, there is a role for sensitivity, specificity, LRs, and DORs in either comparing different diagnostic tests or in comparing results of one diagnostic test across studies. Higher values of DOR indicate

[1]The tests investigated were the Children's Test of Nonword Repetition (CNRep; Gathercole & Baddeley, 1996), the Nonword Repetition Test (NRT; Dollaghan & Campbell, 1998), the Montgomery test (Montgomery, 1995), and various 3–4 syllable item tests (Coady, Evans, & Kluender, in press; Edwards & Lahey, 1998; Kamhi & Catts, 1986; Kamhi, Catts, Mauer, Apel, & Gentry, 1988; Munson, Kurtz, & Windsor, 2005).

increasing classification accuracy. Therefore, across a group of studies (as in a meta-analysis), tests yielding the highest DORs could be described as the most useful, clinically.

Given that children with language impairment over the age of 4;0 score significantly lower than age-matched and language-matched peers on tests of NWR and that NWR has been identified as a potential clinical marker for language impairment, our aim was to determine the diagnostic accuracy of a new NWR test for toddlers at risk of early language impairment. Before doing so, we first documented the diagnostic accuracy of studies using NWR with children older than 4;0. In the field of childhood language disorders, comparisons across studies are of limited value because differences in clinical features of groups of children with SLI across studies contribute to spectrum bias, such that studies are not easily comparable and thus generalizability of results is limited (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003). At this point, we can only comment on how well an NWR test fares, in terms of diagnostic accuracy, within a given study, rather than comparing the relative value of tests. Here, we report DORs for published studies.

To fully assess the diagnostic accuracy of a test, published studies must provide sufficient data for calculation of LR and CIs. First, where sensitivity and specificity figures are reported in studies that compare children with and without language impairment on a NWR task, these can be entered into the Stats Calculator on the Web site of the Centre for Evidence-Based Medicine at the University of Toronto (2008) and LRs derived. Where sensitivity and specificity values or LRs were not reported, if an author reported the percentage of children in the TD group who scored as "false positive" and the percentage of children in the language impairment group who scored as "true positive," as well as the sample size for both groups, a crosstabulation was done to yield the remaining required figures ("true negative" and "false negative"). Sometimes, authors do not report pass/fail cut-points. For example, in two studies (Dollaghan & Campbell, 1998, and Ellis et al., 2000), for a child to be classified as language-impaired on the NWR test, he or she would have to score less than 70% accuracy on the test, and for a child to be classified as TD on the NWR test, he or she would have to score at or above 81% accuracy on the NWR test. This means that no single cut-point is used to identify good and poor performance. To overcome this, arbitrary cut-points can be imposed. For example, children scoring below 75% accuracy on a NWR test can be coded as "test positive" (fail) and those who score equal to or above 75% can be coded as "test negative" (pass).

A literature review revealed six reports that provided sufficient data for calculation of DORs and CIs. One study that examined the diagnostic accuracy of the NRT (Oetting, Cleveland, & Cope, 2008) did not provide sufficient data to calculate CIs and is not considered further here. Table 1 shows the resulting sensitivity and specificity values, LRs, DORs, and CIs for the studies that compared children with language impairment with their TD peers on either the NRT or the CNRep. The sensitivity values for these studies ranged from 52% to 94%, with three reports meeting the 80% accuracy level for sensitivity (Archibald & Alloway, 2007; Dollaghan & Campbell, 1998; Gray, 2003). The specificity values ranged from 88% to 99%, with Archibald and Alloway (2007), Dollaghan and Campbell (1998), and Gray (2003) achieving 96%, 90%, and 98%, respectively. Positive LRs varied from 4.48 to 43. Three studies achieved a LR+ figure of at least 10.0 (Archibald & Alloway, 2007; Conti-Ramsden, 2001; Gray 2003). All studies had large CIs. The negative LRs were similarly diverse across studies, with only Gray (2003) reporting a value < 0.10. Large CIs for LR– were evident for most studies, with Gray (2003) and Dollaghan and Campbell (1998) achieving the narrowest intervals. DORs ranged from 641.79 (Gray, 2003) to 8.84 (Ellis Weismer et al., 2000).

Although all of these studies provide valuable information about the presence of deficits in NWR in children with SLI, because of either small sample size or overlap in group scores (and, therefore, wide CIs), they have not yet been able to meet the requirements for demonstrating excellent diagnostic accuracy. Continuing work in this field will need to include larger sample sizes. If we change our focus to ask which of several NWR tests has the best diagnostic accuracy, then studies will also have to become uniform in ages of participants, methods for classification of language impairment, and severity of impairment. At present, this has not been achieved in the field of child language impairment. This is illustrated in Table 1, in which the range of participants, ages, and classification tests are shown for the studies described here.

Nonetheless, it is encouraging to find that tests of NWR may have a LR+ as high as 43 and a LR– as low as 0.07, with DORs as high as 641. All of the research that allows calculation of DORs has been conducted with preschool and school-aged children with language impairment, and few studies have examined NWR in children under the age of 3. Recent calls to redouble efforts to improve the early identification of children at risk for language impairment (Bercow, 2008) and the encouraging DORs for children over 4 reported here focused our attention on the possible benefits of developing a new test of NWR for 2-year-old children. Our aim was to explore the diagnostic accuracy of a test of NWR in 2-year-old children by comparing the performance of children identified as late talkers with their TD peers. The term *late talker* is conventionally used to describe those children with expressive lexicons below the 10th percentile or having fewer than 50 words or no word combinations by age 2 (for a discussion of this term, see Bates, Dale, &

Table 1. Sensitivity, specificity, positive and negative likelihood ratios (LR+, LR–), each with their respective confidence intervals, and diagnostic odds ratios (DORs) of nonword repetition tasks in children with and without language impairment (LI) in English, including participant and classification information for included studies.

	CNRep				NRT	
Value	Gray, 2003[a]	Conti-Ramsden, 2001[b]	Conti-Ramsden & Hesketh, 2003[b]	Archibald & Alloway, 2007	Dollaghan & Campbell, 1998[c]	Ellis Weismer et al., 2000[c]
Sensitivity	.94 (.76–.99)	.52 (.33–.71)	.59 (.42–.75)	.88 (.60–.97)	.84 (.71–.92)	.56 (.41–.69)
Specificity	.98 (.82–.99)	.99 (.87–.99)	.89 (.71–.96)	.96 (.70–.99)	.90 (.78–.96)	.88 (.84–.91)
LR+	43 (2.77–668.4)	34.5 (2.14–555.89)	5.15 (1.71–15.55)	21 (1.38–319.40)	8.62 (3.37–22.06)	4.48 (3.08–6.53)
LR–	.07 (.01–.31)	.49 (.31–.75)	.46 (.29–.71)	.13 (.03–.59)	.18 (.09–.35)	.51 (.37–.71)
DOR	641.79	71.13	11.21	161	48.97	8.84
Control group (*N*)	AM (22)	AM (32)	LM (32)	AM (11)	AM (20)	AM (359)
Age	5;0 (4;0–5;11)	4;9 (4;4–5;8)	2;10 (2;4–3;7)	9;3 (7;0–11;1)	7;10 (6;0–9;9)	8;9 (7;1–8;11)
LI group (*N*)	SLI (22)	SLI (32)	SLI (32)	SLI (11)	LI (20)	SLI (80)
Age	5;0 (4;0–5;11)	5;1 (4;4–5;10)	5;0 (4;4–5;10)	8;10 (6;9–10;10)	7;10 (6;0–9;9)	8;9 (7;1–8;11)
Classification tests	KABC SPELT–II BBTOP PPVT–III	CELF–P WPPSI–R	CELF–P WPPSI–R RDLS–III	Raven's BPVS TROG CELF–UK3 GFA	PPVT–R TOLD–I:2 TOLD–P:2 TONI	PPVT–III CREVT CELF–3 WISC–III

Note. All indices were computed from raw data reported in the original study using the Stats Calculator from the Centre for Evidence-Based Medicine at the University of Toronto (2008). Sensitivity and specificity calculations may differ from those reported in the original study for cases in which the frequency of one or more cells in the 2 × 2 table was zero. LRs differ from the original study for cases in which different calculations were used. CNRep = Children's Test of Nonword Repetition (Gathercole & Baddeley, 1996). NRT = Nonword Repetition Test (Dollaghan & Campbell et al., 1998). AM = age-matched. LM = language-matched. SLI = specific language impairment. KABC = Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983). SPELT–II = Structured Photographic Expressive Language Test—II (Werner & Kreschek, 1983). BBTOP = Bankson-Bernthal Test of Phonology (Bankson & Bernthal, 1990). PPVT–III = Peabody Picture Vocabulary Test—III (Dunn & Dunn, 1997). CELF–P = Clinical Evaluation of Language Fundamentals—Preschool (Wiig, Secord, & Semel, 1992). WPPSI–R = Wechsler Preschool and Primary Scale of Intelligence—Revised (Wechsler, 1992). RDLS–III = Reynell Developmental Language Scales—III (Edwards et al., 1997). Raven's = Raven's Coloured Matrices (Raven, Court, & Raven, 1986). BPVS = British Picture Vocabulary Scales (Dunn, Dunn, Whetton, & Burley, 1997). TROG = Test for Reception of Grammar (Bishop, 1982). CELF–UK3 = Recalling Sentences subtest of Clinical Evaluation of Language Fundamentals—UK3 (Semel, Wiig, & Secord, 1995b). GFTA–2 = Goldman Fristoe Test of Articulation—Second Edition (Goldman & Fristoe, 2000). PPVT–R = Peabody Picture Vocabulary Test—Revised (Dunn & Dunn, 1981). TOLD–I:2 = Test of Language Development—Intermediate, Second Edition (Hammill & Newcomer, 1988). TOLD–P:2 = Test of Language Development—Primary, Second Edition (Newcomer & Hammill, 1988). TONI = Test of Nonverbal Intelligence (Brown, Sherbenou, & Johnson, 1990). CREVT = Comprehensive Receptive and Expressive Test (Wallace & Hammill, 1994). CELF–3 = Clinical Evaluation of Language Fundamentals, Third Edition (Semel, Wiig, & Secord, 1995a). WISC–III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991).

[a]The first time point was used from Gray (2003). [b]Based on the 16th percentile. [c]Based on < 75% percentage phonemes correct.

Thal, 1995). Because a proportion of children with slow onset of language at age 2 do develop to within normal limits by the age of 4 (Bates et al., 1995), we do not refer to these children as being language impaired but rather as being at risk for language impairment. We sought to establish whether or not difficulties with an NWR task would accurately and easily identify late talkers from a community sample of 2-year-old children.

Method

Participants

A total of 232 British-English–speaking children aged 24–30 months (M = 26.83, SD = 1.48) with no severe medical history or reported hearing loss participated in the study. Of these, 134 children were from Southern England (58%), and 98 were from the North East of England (42%). There were 121 girls (52%) and 111 boys (48%). The children were recruited from either local nurseries and parent–toddler groups (NE England) or a university research database of volunteers (Southern England). Two trained research assistants (RAs) were employed to complete all data collection. Both RAs held undergraduate psychology degrees, had considerable postgraduate experience working with children, and were trained in all the procedures used in this study.

Procedures

Parents received a mailed pack containing an information sheet explaining the study, a form to elicit

informed consent, and two questionnaires, in addition to other materials not relevant to the present study. The first questionnaire addressed family/child demographics, including the child's age, gender, birth date, birth order, medical history, family history of speech/language delay, parents' education level, parental concern about language development, and language(s) spoken in the home. The second questionnaire was a British-English version of the MacArthur-Bates Communicative Development Inventory: Word and Sentences (CDI:WS–UK; Klee & Harrison, 2001).

The parent returned the questionnaires and signed consent form to the university lab in a prepaid, addressed envelope. On receipt of the questionnaires, parents were sent a £5.00 ($8) store voucher. Child details were checked for meeting the selection criteria of no major medical history, monolingual English-speaking, and age 24–30 months, and parents were invited to attend the lab for administration of three standardized tests, the Visual-Reception subscale of the Mullen Scales of Early Learning (MSEL–VR; Mullen, 1995) as a measure of nonverbal cognition, the Receptive One-Word Picture Vocabulary Test (ROWPVT; Brownell, 2000b), and the Expressive One-Word Picture Vocabulary Test (EOWPVT; Brownell, 2000a); an experimental test of NWR; and collection of a language sample, in that order. Parents were paid £20 ($33) for participating in the lab phase of the study. The CDI was not scored before the child attended the lab session so that the RA would be blind to the child's measured language status. Although the same RA administered the standardized tests and the NWR test, the tests were not scored until all tasks had been completed, aiming to maintain blinding for the duration of data collection.

Index Test and Reference Standard

Index test. Given that our focus was on 2-year-old children, we considered it appropriate to devise a new Test of Early Nonword Repetition (TENR) that differed from others in both construction and administration to ensure that the nonwords were within the phonetic inventories of 2-year-olds, as far as possible, and that the task would be sufficiently engaging of their attention. Some of the words in the Roy and Chiat (2004) test contained word-like syllables (e.g., *peas* in *lepeese, jam* in *jamic, sign* in *sinodaur*), which we aimed to avoid. The TENR contained 12 one-, two-, and three-syllable nonsense words (4 of each type) consisting of early developing consonants and tense vowels (see the Appendix), and trials were conducted with children aged 18–25 months. The pilot study (N = 24, mean age = 22 months) revealed that children under 24 months were unlikely to complete the task (only 6 children completed the task). One single-syllable word was subsequently excluded, as it resembled a real word. The words were presented in a set order from one to three syllables, with a live voice to engage the attention of these 2-year-olds, as has been done in other studies with young children (e.g., Roy & Chiat, 2004). The child was asked to imitate the nonwords said by the experimenter and then was allowed to roll a ball down a chute as a reward. The child was instructed to "Say what I say and then push the ball down." The task was introduced by asking the child to "Say teddy," after which the task began. Each correct consonant or vowel was awarded a point, and the total percentage correct was calculated. Child errors that reflected consistent substitution errors in the child's spontaneous speech and in the picture vocabulary test were counted as correct (e.g., consistent [t] for /k/). If a child failed to complete the entire TENR, his or her scores were excluded from analysis. The data from the first 98 children had a median score of 81% accuracy. Consequently, 4 four-syllable nonwords were added to the test, and this test was administered to the remaining 131 children. Four of the words in the final set were modifications of items in the CNRep (Gathercole & Baddeley, 1996; see the Appendix), as the original items had low wordlikeness and low articulatory complexity (Archibald & Gathercole, 2006).

Reference standard. The CDI was selected as the reference instrument because it is commonly used to identify late talkers or children at the low end of the vocabulary distribution (e.g., D'Odorico, Assaneli, Franco, & Jacob, 2007; Thal, Tobias, & Morrison, 1991). Some studies have set the 10th percentile on the expressive vocabulary scale as the cut-point for late talker status (e.g., D'Odorico et al., 2007; Thal et al., 1991) and SLI status (see Leonard, 1998; McCauley, 2001), whereas other studies have set cut-points at the 25th and 16th percentiles (e.g., Conti-Ramsden, 2001). To achieve sufficient numbers for analysis, the 16th percentile (i.e., 1 standard deviation below the mean) was selected here.

Results

The Dataset

A total of 222 parents brought the completed CDI form to the assessment session within 1 month of completion of the CDI. Of this sample, 177 children (80%) completed the TENR (55 children either did not start or did not complete the task). Of these, 172 children had both CDI and TENR scores, with missing CDI data for 5 children. To examine whether children in the noncompliant group (NC-NWR) were different from those who complied with the NWR test (C-NWR), one-way analyses of variance (ANOVAs) were run with age, CDI score, MSEL–VR, ROWPVT, and EOWPVT as dependent variables. The NC-NWR group scored significantly lower than the C-NWR group on all measurements except age

(see Table 2), indicating that children who did not complete the TENR had lower developmental indices than children who completed the task. This may indicate that children who are not compliant on this task require further assessment, but such speculation needs to be further explored. There was no difference in the rates of boys and girls who completed the task, $\chi^2(1, N = 222) = 1.01$, $p = .34$, and there was no difference in TENR scores for children across the two data collection sites, $F(1, 176) = 1.94$, $p = .166$.

Classification Accuracy of the TENR 1–3 Syllable Test

The scores for all children who completed one-, two-, and three-syllable nonwords were used for this analysis ($N = 172$). The 16th percentile points for the CDI (hereafter CDI16) and the TENR (hereafter TENR16) were calculated for each month of age. Table 3 shows the breakdown. Children who scored below the 16th percentile on the CDI total vocabulary score or had no word combinations reported on the CDI were coded as "late talkers" (LT); the remainder were coded as "typically developing" (TD). Children who scored below the 16th percentile on the TENR were coded as "test positive," and the remainder were coded as "test negative." Crosstabulation of TENR16 percentile × CDI16 percentile was conducted using SPSS Version 15.0 to generate figures for entry into the calculation of sensitivity, specificity, and likelihood ratios. Crosstabulation results are shown in Table 4.

Table 3. 16th percentile cut-points for the CDI and TENR by age group.

	CDI		TENR	
Age (months)	Raw score	*N*	Percentage correct	*N*
24	167	11	50	7
25	148	32	59	18
26	159	58	64	47
27	226	40	68	36
28	290	43	73	35
29	335	34	75	26
30		4		3

Note. Blank cells indicate that no cut-point was applied because of small sample size.

These figures were entered into the Stats Calculator (Centre for Evidence-Based Medicine, 2008).

The sensitivity value shows the number of LTs who were correctly classified as LT by the TENR (75%; 95% CI = 51%–90%). The specificity value shows the number of TDs who were correctly classified as TD by the TENR (90%; 95% CI = 85%–94%). The LR+ shows the likelihood that a score below the 16th percentile on the TENR came from a child who was classified as LT (7.8; 95% CI = 4.5–13.6), and the LR– shows the likelihood that a score above the 16th percentile on the TENR came from a child coded as LT (.28; 95% CI = .12–.65). That is, children who are LT are about 8 times more likely to have a positive

Table 2. Significant differences in CDI, MSEL–VR, ROWPVT, and EOWPVT scores between children who were and were not compliant for the Test of Early Nonword Repetition (TENR) 1–3 syllable test and 1–4 syllable test (C-TENR; NC-TENR).

	NC-TENR	C-TENR			
Test	*M* (*SD*)	*M* (*SD*)	*F*	*p*	Cohen's *d* (effect size)
			TENR 1–3 syllable test		
Age	26 (0.52)	26 (0.95)	$F(1, 221) = 3.28$	> .05	
CDI	309 (178)	417 (139)	$F(1, 221) = 20.25$	< .0001	.67
MSEL–VR	30 (3.5)	33 (4.0)	$F(1, 187) = 15.79$	< .0001	.74
ROWPVT	22 (7.17)	29 (8.12)	$F(1, 210) = 25.91$	< .0001	.88
EOWPVT	20 (7.90)	25 (7.83)	$F(1, 212) = 17.10$	< .0001	.33
			TENR 1–4 syllable test		
Age	25 (1.45)	26 (1.32)	$F(1, 129) = 6.14$	.015	.05
CDI	263 (160)	401 (128)	$F(1, 125) = 35.94$	< .0001	.22
MSEL–VR	30 (3.43)	34 (4.14)	$F(1, 187) = 23.67$	< .0001	.16
ROWPVT	20 (5.95)	29 (8.12)	$F(1, 123) = 27.76$	< .0001	.18
EOWPVT	18 (8.46)	26 (7.98)	$F(1, 120) = 18.59$	< .0001	.13

Note. CDI = MacArthur-Bates Communicative Development Inventory (UK version; Klee & Harrison, 2001). MSEL–VR = Visual-Reception subscale of the Mullen Scales of Early Learning (Mullen, 1995). ROWPVT = Receptive One Word Picture Vocabulary Test (Brownell, 2000b). EOWPVT = Expressive One-Word Picture Vocabulary Test (Brownell, 2000a).

Table 4. Crosstabulation for the TENR16 and CDI16.

TENR test result	Diagnosis		Total
	Late talking	Typically developing	
Test positive	12	15	27
Test negative	4	141	145
Total	16	156	172

Note. TENR16 and CDI16 = 16th percentile cut-points by age group; 1–3 syllable test.

test result on the TENR than children who are TD, and children who are LT are about 0.28 times more likely to have a negative test result on the TENR than children who are TD. These results are considered to be "moderate LR+" and "moderate LR–" (Dollaghan, 2007). The resulting DOR was 27.86. To test the combined value of diagnostic indicators, LRs for family history of speech/language disorder (n = 16/172), parent concern for child speech/language development (n = 5/167), and gender of the child (male = 79; female = 93) were entered into the same analysis. First, sensitivity, specificity, and LRs were generated for these variables, with CDI16 as the outcome variable. No improvement in these figures was noted with the inclusion of a family history of speech/language impairment, parent concern about the child's speech/language development, or gender of the child in the classification calculations.

With these less than ideal results, TENR was rescored with 1 point awarded for each whole word correct. The 16th percentile was determined for each age group, and the crosstabulations were repeated. The results were worse than before, yielding a LR+ of 3.97. One further analysis using the 16th percentile was conducted, using different cut-points. The 16th percentiles for both the TENR and CDI were generated for the entire sample of children, not taking age into account. The 16th percentile for the CDI was 218, and for TENR it was 66% accuracy. Although specificity and LR+ improved, sensitivity and LR– results were unacceptable: sensitivity = 69% (95% CI = 44%–86%), specificity = 92% (95% CI = 87%–96%), LR+ = 8.94 (95% CI = 4.7–16.9), and LR– = .34 (95% CI = 0.16–0.70).

Classification Accuracy of the TENR 1–4 Syllable Test

Recall that the TENR 1–4 syllable test was only attempted with 131 children. Of these, 96 children completed the task (73%). One-way ANOVAs revealed that, as with the 1–3 syllable task, the noncompliant children scored significantly lower on the expressive and receptive vocabulary tests, the MSEL–VR test, and the CDI. Cut-points at the 16th percentile for the 1–4 syllable test were identified. Child scores were entered into a crosstabulation of TENR16 percentile × CDI16 percentile using SPSS to generate figures for entry into the calculation of sensitivity, specificity, and likelihood ratios. Crosstabulation results are shown in Table 5.

The sensitivity value shows the percentage of LTs who were correctly classified as LT by the TENR (88%; 95% CI = 53%–99%). The specificity value shows the percentage of TDs who were correctly classified as TD by the TENR (94%; 95% CI = 87%–98%). The LR+ shows the likelihood that a score below the 16th percentile on the NWR came from a child who was classified as LT (14.88; 95% CI = 6.1–36.2), and the LR– shows the likelihood that a score above the 16th percentile on the TENR came from a child coded as LT (.13; 95% CI = .02–.83). That is, children who are LTs are about 15 times more likely to have a positive test result on the TENR than children who are TD, and children who are LTs are about .13 times more likely to have a negative test result on the TENR than children who are TD. The DOR was 114. Thus, the results for the 1–4 syllable TENR were better than those of the 1–3 syllable test. Again, no improvement in these figures was noted with the inclusion of a family history of speech/language impairment, parent concern about the child's speech/language development, or gender of the child in the classification calculations. This is because of the degree of overlap in the variables. Of the 12 children failing the 1–4 syllable TENR, 10 were male, but only 1 had a family history of speech/language disorder and only 3 of the parents expressed concern about speech/language development.

Discussion

The LR+, LR–, and DOR results derived from published studies on NWR in children with language impairment encouraged us to explore the clinical utility of two versions of an NWR test with 2-year-old children. The usefulness of the TENR 1–3 syllable test was not as good as one would like, regardless of the method of scoring or method of calculating the cut-points. The TENR 1–4

Table 5. Crosstabulation for the TENR16 and CDI16.

TENR test result	Diagnosis		Total
	Late talking	Typically developing	
Test positive	7	5	12
Test negative	1	80	81
Total	8	85	93

Note. TENR16 and CDI16 = 16th percentile cut-points by age group; 1–4 syllable test.

syllable test demonstrated better diagnostic accuracy. However, the sample size for the LT group was only 8 for this last analysis, contributing to large CIs. Given this outcome, and the number of children who were not compliant during testing, it is tempting to conclude that there may be little value in pursuing the use of poor NWR performance as an indicator of early language delay in 2-year-olds. However, there are several caveats to this.

First, the results for the 1–4 syllable TENR test suggest that with a larger sample size of children, excellent results would be achieved. This is indicated by the good LRs generated, with caution suggested because of the large confidence intervals.

Second, the CDI was used to classify language delay (LT status). The CDI has been shown to have reasonable validity and reliability (see summary of research in Fenson et al., 2007). In our sample the CDI correlated significantly with an expressive vocabulary test (EOWPVT; Brownell, 2000a) at $r(205) = .54$ ($p < .001$), but as yet there is no consensus on the best method for diagnosing a language impairment in 2-year-old children. An analysis of the diagnostic accuracy of language tests for this age group is required before we repeat the attempt at determining the classification accuracy of NWR (Klee, 2008). This is an important way forward, as Dollaghan (2007) has suggested that moderate results indicate that the diagnostic/classification test needs to be administered in conjunction with another instrument. If NWR, even in conjunction with another as-yet-untested variable such as the proportion of optimal birth weight (Zubrick, Taylor, Rice, & Slegers, 2007), can be successfully used to identify possible cases of language delay in toddlers, then a tool could be developed that works across several languages, taking phonological systems into account. As our clinical populations become more multicultural, this would help to overcome the limitations of language-specific diagnostic tests in identifying toddlers who are more likely to develop language impairment. Further, the NWR test used in this study is quick to administer (less than 10 min), so cost savings should be considerable, and it would allow for very early identification of likely language impairment risk.

Third, the TENR test was administered 1 hr into a 2-hr assessment session with these children. It would be worthwhile to replicate the study with a comparable or larger sample size and administer only a language test and the NWR test in an effort to gain better compliance and avoid the effects of fatigue. Finally, alternative tasks used to gain child cooperation should be explored, as the ball and chute game used in this experiment may not have attracted every child, although it is worth noting that the children who did not attempt or complete the TENR test scored significantly lower on all other measures, indicating the need for follow-up.

In summary, the usefulness of a test of NWR to indicate the presence of language delay requires further exploration in both younger and older children. Favorable preliminary results for younger children from a test with one- to four-syllable nonwords were obtained.

Acknowledgments

This project was funded by Economic and Social Sciences Research Council Grant RES-000-22-0712. We thank Carmel Houston-Price and Graham Schafer for access to the Reading University research database, Jill Hearing and Sarah Fincham-Majumdar for excellent work as research assistants, and Northumbria University for their generous support of the project.

References

Adams, A.-M., & Gathercole, S. E. (1995). Phonological working memory and speech production in preschool children. *Journal of Speech and Hearing Research, 38,* 403–414.

Archibald, L. M. D., & Alloway, T. P. (2007). Comparing language profiles: Children with specific language impairment and developmental coordination disorder. *International Journal of Language and Communication Disorders, 43,* 165–180.

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

Bankson, N. W., & Bernthal, J. E. (1990). *Bankson-Bernthal Test of Phonology*. Chicago, IL: Riverside Publishing.

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of child language*. Oxford, England: Basil Blackwell.

Bercow, J. (2008). *The Bercow report*. Retrieved September 2, 2008, from http://www.dcsf.gov.uk/bercowreview.

Bishop, D. V. M. (1982). *Test for Reception of Grammar*. Oxford, England: Chapel Press.

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *The Journal of Child Psychology and Psychiatry, 37,* 391–403.

Brown, L., Sherbenou, R., & Johnson, S. (1990). *Test of Nonverbal Intelligence—2*. Austin, TX: Pro-Ed.

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

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

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing dependent measures. *Journal of Speech and Hearing Research, 40,* 519–525.

Centre for Evidence-Based Medicine. (2008). *Stats calculator, EBM toolbox*. Retrieved March 26, 2008, from http://www.cebm.utoronto.ca.

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.

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *The Journal of Child Psychology and Psychiatry, 49,* 635–645.

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.

Coady, J. A., Evans, J. L., & Kluender, K. R. (in press). The role of phonotactic frequency in nonword repetition by children with specific language impairments. *International Journal of Language and Communication Disorders*.

Conti-Ramsden, G. (2001). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research, 46,* 1029–1037.

Conti-Ramsden, G., & Hesketh, A. (2003). Risk markers for SLI: a study of young language-learning children. *International Journal of Language and Communication Disorders, 38,* 251–263.

D'Odorico, L., Assanelli, A., Franco, F., & Jacob, V. (2007). A follow-up study of Italian late talkers: Development of language, short-term memory, phonological awareness, impulsiveness, and attention. *Applied Psycholinguistics, 28,* 157–169.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore: Paul H. Brookes.

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

Dunn, L., & Dunn, L. (1981). *The Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: AGS.

Dunn, L., & Dunn, L. (1997). *The Peabody Picture Vocabulary Test–III*. Circle Pines, MN: AGS.

Dunn, L. M., Dunn, L. M., Whetton, C. W., & Burley, J. (1997). *The British Picture Vocabulary Scales* (2nd ed.). Windsor, England: NferNelson.

Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. *Applied Psycholinguistics, 19,* 279–309.

Edwards, S., Fletcher, P., Garmn, M., Hughes, A., Letts, C., & Sinka, I. (1997). *Reynell Developmental Language Scales—III (RDLS)*. Windsor, England: NferNelson.

Ellis Weismer, S., Tomblin, J. B., Zhang, X., Buckwalter, 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.

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

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

Gathercole, S. E., & Baddeley, A. D. (1996). *The Children's Test of Nonword Repetition*. London: Psychological Corporation.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology, 56,* 1129–1135.

Glascoe, F. P., & Squires, J. (2007). Issues with the new developmental screening and surveillance policy statement. *Pediatrics, 119,* 861–862.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation*—Second Edition. Circle Pines, MN: AGS.

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

Gray, S. (2003). Diagnostic accuracy and test-retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders, 36,* 129–151.

Hammill, D., & Newcomer, P. (1988). *Test of Language Development—Intermediate, Second Edition*. Austin, TX: Pro-Ed.

Kamhi, A. G., & Catts, H. W. (1986). Toward an understanding of developmental language and reading disorders. *Journal of Speech and Hearing Research, 51,* 337–347.

Kamhi, A. G., Catts, H. W., Mauer, D., Apel, K., & Gentry, B. F. (1988). Phonological and spatial processing abilities in language- and reading-impaired children. *Journal of Speech and Hearing Disorders, 53,* 316–327.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: AGS.

Klee, T. (2008). Considerations for appraising diagnostic accuracy studies in communication disorders. *Evidence-Based Communication Assessment and Intervention, 2,* 34–45.

Klee, T., & Harrison, C. (2001, July). *CDI words and sentences: Validity and preliminary norms for British English*. Paper presented at Child Language Seminar, University of Hertfordshire, England.

Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

McAlister, F. A., Straus, S. E., & Sackett, D. L. (1999). Why we need large, simple studies of the clinical examination: The problem and a proposed solution. *Lancet, 354,* 1721–1724.

McCauley, R. J. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Lawrence Erlbaum.

Meisels, S. J. (1988). Developmental screening in early childhood: The interaction of research and social policy. *American Review of Public Health, 9,* 527–550.

Montgomery, J. W. (1995). Sentence comprehension in children with specific language impairment: The role of phonological working memory. *Journal of Speech and Hearing Research, 38,* 187–199.

Montgomery, J. W. (2002). Understanding the language difficulties of children with specific language impairments: Does verbal working memory matter? *American Journal of Speech-Language Pathology, 11,* 77–91.

Mullen, E. M. (1995). *Mullen Scales of Early Learning*. Circle Pines, MN: AGS.

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 language impairments. *Journal of Speech, Language, and Hearing Research, 48,* 1033–1047.

Newcomer, P., & Hammill, D. (1988). *Test of Language Development—Primary, Second Edition*. Austin, TX: Pro-Ed.

Oetting, J., Cleveland, L. H., & Cope, R. F. (2008). Empirically derived combinations of tools and clinical cutoffs: An illustrative case with a sample of culturally/linguistically diverse children. *Language Speech and Hearing Services in Schools, 39,* 44–53.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language Speech and Hearing Services in Schools, 25,* 15–24.

Raven, J. C., Court, J. H., & Raven, J. (1986). *Raven's Coloured Matrices*. London: H. K. Lewis.

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.

Semel, E., Wiig, E., & Secord, W. (1995a). *Clinical Evaluation of Language Fundamentals, Third Edition*. San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (1995b). *Clinical Evaluation of Language Fundamentals—UK3*. London: Psychological Corporation.

Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry, 50,* 498–505.

Thal, D., Tobias, S., & Morrison, D. (1991). Language and gesture in late talkers: A 1-year follow-up. *Journal of Speech and Hearing Research, 34,* 604–612.

Wallace, G., & Hammill, D. (1994). *Comprehensive Receptive and Expressive Vocabulary Test (CREVT)*. Austin, TX: Pro-Ed.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition (WISC–III)*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1992). *Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI–R)*. Sidcup, UK: The Psychological Corporation.

Werner, E., & Krescheck, J. D. (1983). *Structured Photographic Expressive Language Test—II*. DeKalb, IL: Janelle.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3,* 1–13.

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals—Preschool (CELF–P)*. Sidcup, UK: Psychological Corporation.

Zubrick, S. R., Taylor, C. L., Rice, M. L., & Slegers, D. W. (2007). Late language emergence at 24 months: An epidemiological study of prevalence, predictors and covariates. *Journal of Speech, Language, and Hearing Research, 50,* 1562–1592.

Received April 30, 2008

Revision received July 13, 2008

Accepted November 26, 2008

DOI: 10.1044/1092-4388(2009/08-0030)

Contact author: Stephanie F. Stokes, Curtin University of Technology, School of Psychology, GPO Box U1987, Perth, Western Australia 6845, Australia.
E-mail: s.stokes@curtin.edu.au.

Appendix. Test of Early Nonword Repetition.

mad	kouɡə	moukəɹi	pəduləmeip
neit	dafi	doupəluf[b]	fɛnəraisɛk[c]
paim	ləpou	bæləkɒn	wuɡəlæmɪk[d]
bous[a]	fupɪm	fisaimɒt	lɒdənætɪʃ[e]

[a]This item was deleted because it closely resembled *bouz*. Equal stress is applied on each syllable, except in /pədulәmeip/, where the stress is on the second syllable, as in *perambulate*. [b]Modified from /dɒpəleit/. [c]Modified from /fɛnəraiz/. [d]Modified from /wʊɡəlæmɪk/. [e]Modified from /lɒdəeipɪʃ/, all from the CNRep (Gathercole & Baddeley, 1996).