Published in final edited form as:
*Lang Speech Hear Serv Sch*. 2010 July ; 41(3): 277–288. doi:10.1044/0161-1461(2009/08-0096).

# Extending Use of the NRT to Preschool-Aged Children with and without Specific Language Impairment

**Patricia Deevy**, **Lisa Wisman Weil**, **Laurence B. Leonard**, and **Lisa Goffman**
Purdue University, West Lafayette, Indiana

## Abstract

**Purpose—**The purpose of this study was to assess the diagnostic accuracy of the Nonword Repetition Test (NRT; Dollaghan & Campbell, 1998) in a sample of four- and five-year-olds with and without specific language impairment (SLI), and to evaluate its feasibility for use in universal screening.

**Method—**The NRT was administered to 29 children with SLI and 47 age-matched children with typical development (TD). Diagnostic accuracy was computed using alternative scoring methods, which treated out-of-inventory phonemes either as errors or as unscorable. To estimate accuracy in a universal screening context, probability of identifying a child at risk for language impairment was computed using the prevalence of SLI (7%) as the base rate.

**Results—**Diagnostic accuracy was acceptable using both scoring methods. The resulting likelihood ratios (LR+ = 22.66, 19.43; LR- = .05, .05) were similar to those reported for older children. The probability of accurate detection of children with SLI in the general population increased from 7% to 61%. However this value suggests that many false positives could be expected.

**Conclusions—**The NRT yielded results similar to those reported for older children. However, despite its strengths, the NRT is not sufficient for screening the general population of four- and five-year-olds.

## Introduction

In recent years, numerous studies have examined nonword repetition ability in children with specific language impairment (see Graf Estes, Evans, & Else-Quest, 2007, for a recent review). This ability has been measured in tasks that required children to repeat nonsense words from one (e.g., /nɑɪb/) or two (e.g., /teɪvɑk/) to four (e.g., /wugəlæmɪk/) or five syllables (e.g., /vɛsətreɪʃanɪst/) in ɪ length, with the child's accuracy measured in terms of the total number of phonemes repeated correctly, or the number of nonwords repeated correctly. It has been argued that nonword repetition is primarily a measure of phonological short-term memory capacity (that is, the capacity to temporarily store phonological information) because children's accuracy decreases as words get longer (Gathercole & Baddeley, 1989, 1990). However, these tasks also draw on a variety of skills involved in perception, encoding, and production, all or some of which might be weak in particular children and lead to lower nonword repetition scores (see Gathercole, 2006 for discussion).

The first studies of nonword repetition ability involving children with specific language impairment (SLI) emphasized group differences. In these studies, children with SLI performed significantly below the levels of children with typical language development (TD) who were the same chronological age. Both groups typically showed accurate repetition on the shortest words, but as the words increased in length, accuracy fell more sharply for the group with SLI than for the group with TD (Archibald & Gathercole, 2006; Bishop, North, & Donlan, 1996; Dollaghan & Campbell, 1998; Edwards & Lahey, 1998; Ellis Weismer, Tomblin, Zhang,

Buckwalter, Chynoweth, & Jones, 2000; Gathercole & Baddeley, 1990; Montgomery, 1995; Munson, Kurtz, & Windsor, 2005).

The studies of nonword repetition have employed a variety of nonword stimuli. However, two sets of nonword stimuli have been used most frequently. The first is the set of 40 nonwords developed by Gathercole and Baddeley (Gathercole & Baddeley, 1990) and later revised to form the Children's Test of Nonword Repetition (CNRep; Gathercole, Willis, Baddeley, & Emslie, 1994). The second is the set of 16 nonwords developed by Dollaghan and Campbell (1998), currently referred to as the Nonword Repetition Test (NRT). For both sets of nonwords, comparisons between children with SLI and same-age peers with TD have produced differences that reflect large effect sizes.

The potential of nonword repetition to serve as a clinical tool has been recognized by many researchers (Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis Weismer et al., 2000; Gray, 2003; Oetting & Cleveland, 2006; Washington & Craig, 2004). Bishop, North, and Donlan (1996) presented two compelling reasons to believe that such tasks tap into a deficit that is characteristic of individuals with language impairment. First, they found that even after language problems had resolved (according to scores on standardized tests) children with SLI showed significantly lower accuracy than children with TD on the task. Second, by comparing task performance in monozygotic and dizogotic twin pairs, in which one of the co-twins had a language impairment, they showed that there was a significant genetic component in the nonword repetition deficit and thus, that the task could serve as a behavioral marker of heritable forms of language impairment. However, as was pointed out by Bishop et al. (1996), neither group differences nor the genetic findings guarantee that a nonword repetition task will be a useful clinical tool. That is, while there may be significant differences between group means, the distributions of the scores of the two groups may overlap.

As a first step in determining the clinical utility of nonword repetition measures, subsequent studies have asked whether samples of children independently classified as exhibiting language impairment or as having typical language skills could be accurately indentified as "affected" or "unaffected" based on their performance on a task of nonword repetition. In these studies, the degree of match between the sorting of the affected and unaffected groups by the nonword repetition measure and the original classification has sometimes been referred to as "diagnostic accuracy" (e.g., Gray, 2003; Oetting & Cleveland, 2006). Use of this term follows a tradition reflected in earlier studies that have compared the degree of match between the results of a newly developed test and the original classification according to some gold standard (see Greenslade, Plante, & Vance, in press for a recent example). Given this precedent, we have employed the term "diagnostic accuracy" in this paper to refer to a nonword repetition task's accuracy in sorting samples of children relative to their classification based on a gold standard.

The study reported in this paper had three goals, discussed in greater detail in subsequent sections. First, we evaluated the diagnostic accuracy of one nonword repetition task, the NRT, when used with children four and five years of age, and hence younger than the age levels that have been studied in earlier investigations of the diagnostic accuracy of this measure. Second, given that younger children may exhibit phonological limitations that could influence their performance on measures such as the NRT, we explored whether diagnostic accuracy changes as a function of how phonological errors are scored. Third, we determined the feasibility of using the NRT as a screening measure administered to the wider population of unidentified four- and five-year-olds by estimating its accuracy given the presumed prevalence rate of SLI.

## Diagnostic Accuracy

Studies of the diagnostic accuracy of nonword repetition tasks have employed samples of children who were independently classified as either exhibiting a language impairment or

typical language skills based on a gold standard of enrollment in speech-language services. These studies have often used roughly equal numbers of children in each group. Accuracy has been measured in terms of sensitivity, that is, the degree to which children independently classified as showing SLI are identified as affected by the nonword repetition measure, and specificity, that is, the degree to which children independently classified as displaying typical language skills are identified as unaffected by the nonword repetition measure. Following recommendations by Plante and Vance (1994), sensitivity and specificity values of 80% to 89% have been considered adequate, and values of 90% and higher have been considered good.

Two studies have reported sensitivity and specificity values for preschool-aged children using the CNRep. Gray (2003) administered the CNRep to 22 four- and five-year-old children with SLI and 22 age-matched children with typical language development. Initial classification was determined by treatment status; a child was placed in the group with SLI for this study if he or she had qualified for treatment (i.e., had scored below −1.5 *SD* of the mean on two standardized language tests). CNRep scores were then used to classify the same children into two groups (affected and unaffected). The proportion of children with SLI who were correctly classified (sensitivity) by the CNRep was found to be 95%; the proportion of children with TD who were correctly classified (specificity) was 100%. Conti-Ramsden (2003) tested two groups of 32 children each, aged four and five years. All children with SLI were in language treatment at the time of testing. Although specificity was high at 100%, sensitivity was below levels of adequacy at 66%.

One potential limitation of the CNRep test is that some of the nonwords in this set include phoneme sequences which match those of actual morphemes of English (e.g., “pen” in /pɛnl/, “ing” in /slɑdiŋ/). The NRT, on the other hand, was designed to minimize wordlikeness by using sounds only in syllable positions in which they occur infrequently in English. It has been argued that repeating nonwords with “low wordlikeness” places greater demands on processing ability. In particular, it requires a greater reliance on phonological working memory since there is little or no support from stored lexical representations (Gathercole, 1995). This characteristic may reduce the potential bias due to cultural or economic differences inherent in measures based on language knowledge, and thus may allow for more accurate identification (Campbell, Dollaghan, Needleman, & Janosky, 1997; Rodekohr & Haynes, 2001).

Thus far, the diagnostic accuracy of the NRT has been evaluated only in children aged six years and older. Dollaghan and Campbell (1998) compared 20 six- to nine-year-old children with SLI to 20 age-matched peers with TD. Children were initially classified as SLI if they were currently enrolled in speech-language therapy. These investigators evaluated the NRT’s ability to accurately classify the children using likelihood ratios (LRs). The positive likelihood ratio (LR+) is computed as sensitivity/(1-specificity) and reflects the odds that a score within the range designated as the “affected range” came from an affected child. As a rough guide, an LR + of 10 (or greater) for a test score in the affected range indicates that the odds are 10 (or more) to 1 that the score came from a child with a language disorder and thus allows “ruling in” the diagnosis with high confidence (see Sackett, Haynes, Guyett, & Tugwell, 1991). The negative likelihood ratio (LR-) is computed as (1-sensitivity)/specificity and reflects the odds that a score in the unaffected range came from an affected child. An LR- of 0.10 or less reflects odds low enough to allow “ruling out” the diagnosis with high confidence. Dollaghan and Campbell identified score ranges for the NRT that could rule in or rule out language impairment with a high degree of confidence. They found an LR+ of 25 for scores of 70% or less total percentage of phonemes correct (TPPC); that is, the scores within this range were 25 times more likely to have come from a child with SLI than a child with TD. They found an LR- of 0.03 for TPPC scores 81% or higher, thus having less than a 1 in 20 chance of coming from a child with SLI.

Ellis Weismer et al. (2000) examined diagnostic accuracy of the NRT in a population-based sample of 581 seven- and eight-year-olds. These investigators used three different ways of initially classifying children into affected and unaffected groups and found that the NRT yielded the best discrimination when classification was based on whether a child was or was not currently receiving treatment (rather than on test scores). Computing LRs for the same levels of scores used by Dollaghan and Campbell (1998), Ellis Weismer et al. found an LR+ of 6.71 for TPPC scores of 70% or less and an LR- of 0.29 for scores of 81% or higher. When the more extreme score levels of 60% or less and 90% or more were used, the resulting LR+ and LR- were 10.0 and 0, respectively.

A third study of diagnostic accuracy employing the NRT was conducted by Oetting and Cleveland (2006) and focused exclusively on six-year-olds. These investigators tested 16 children with SLI and 36 children with TD. Classification as SLI was based on treatment status as well as scores below −1 *SD* of the mean on the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981) and on the syntactic quotient of the Test of Language Development-Primary (TOLD-P:2; Newcomer & Hammill, 1988). Although LRs were not reported, based on the sensitivity and specificity values provided by Oetting and Cleveland, the LR+ can be calculated as 7.0 and the LR- as 0.48. (If the LRs were calculated using more extreme score levels rather than the single cut point provided by the discriminant function analysis used by Oetting and Cleveland, it is likely that the LR+ would be higher and the LR- would be lower.)

One goal of the present study was to evaluate the diagnostic accuracy of the NRT in a sample of younger four- and five-year-old children. Group differences between preschoolers with and without language impairment have already been reported for the NRT (Gray, 2004, 2006; Thal, Miller, Carlson, & Vega, 2005; Washington & Craig, 2004). However, it was not clear how this measure would fare in terms of diagnostic accuracy in this younger age range. As in many earlier studies, we computed sensitivity and specificity values. However, we followed Dollaghan and Campbell (1998) in using LRs as the principal metric for evaluating diagnostic accuracy. LRs can provide better information about the clinical utility of a test than can sensitivity and specificity (Dollaghan, 2007; Sackett et al., 1991). For example, the range of scores yielded by a measure can be divided into several levels rather than just two, with LRs computed at each level. This provides the clinician with information about the degree of abnormality of a score at different points in the range, rather than categories of normal or abnormal (Sackett et al, 1991). In addition, LRs are not affected by prevalence rate to the extent seen for sensitivity and specificity.

## Accounting for Phonological Limitations

The nonwords in the NRT were designed to minimize the phonological demands for children ages six years and older (Dollaghan & Campbell, 1998). However, when assessing the nonword repetition performance of four- and five-year-olds, phonological ability may prove to be an important factor. At these ages, some errors might reflect a child's limitation in producing a phoneme accurately, even in words already known by the child. This can be true even for the consonants included in the NRT, especially in the case of speech delays (e.g., Shriberg & Kwiatkowski, 1994), which can co-occur with SLI (e.g., Shriberg, Tomblin, & McSweeney, 1999). If these phonemes are required in a nonword repetition task, the child's score might be lower, not simply because of limitations in phonological memory, but also because of difficulty with the phonemes themselves. A second goal of the present study was to determine if the diagnostic accuracy of the NRT differed appreciably depending upon whether misarticulations were scored as errors.

Researchers who have examined the nonword repetition abilities of preschoolers have approached the issue of phonological limitations in a variety of ways. Some have tried to

accommodate young children's speech patterns by liberalizing the scoring system, most commonly by counting as correct those phoneme substitutions that reflect normal phonological processes and that are revealed elsewhere either in the child's spontaneous speech (Chiat & Roy, 2007; Conti-Ramsden, 2003) or on a standardized test of articulation (Gray, 2003, 2004). Others have excluded particular segments such as fricatives and/or affricates from scoring (Thal et al., 2005). Washington and Craig (2004) simply documented that the speech of children in their study was within normal limits on a standardized test of articulation. Bishop, Adams, and Norbury (2006) used a score derived by adjusting a child's nonword repetition score on longer words based on their score on the shortest words, with the rationale that errors on the shortest words were more likely to reflect limitations in phonological accuracy than limitations in phonological memory.

When alternative scoring methods have been used in group comparisons between children with SLI and their peers with TD, the scoring methods have not substantially altered the findings; group differences favoring the peers with TD have remained (e.g., Thal et al., 2005). However, stable findings from simple group comparisons cannot be taken to mean that diagnostic accuracy with preschoolers is unaffected by phonological limitations. Accordingly, in the present study, we evaluated diagnostic accuracy with two scoring methods. In the first method, we treated all phoneme substitutions as errors; in the second, phoneme substitutions were allowed, provided that the misarticulated phoneme was not in the child's inventory.

## Feasibility as a Screening Instrument for the Wider Population of Children

To date, diagnostic accuracy of the NRT has been computed based on children who actually served as study participants. In some studies (e.g., Dollaghan & Campbell, 1998), an equal number of children with SLI and with typical language have participated. In others, a smaller percentage of children constituted the group with language impairment (15.49% in Ellis Weismer et al., 2000; 30.77% in Oetting & Cleveland, 2006). Calculations based on such samples are important first steps in determining the clinical utility of a measure such as the NRT. However, in the present study, we attempted a further application.

A third goal of the present study was to assess the feasibility of using the NRT as a screening measure for the wider unidentified population of four- and five-year-olds. Despite the value of accurately identifying children from the general population who might be at risk for language impairment and therefore in need of further testing, the current evidence does not support the use of universal screening for language impairment, at least with the screening instruments that have been evaluated thus far (see Law, Boyle, Harris, Harkness, & Nye, 2000). One of the challenges facing current universal screening tools is the high percentage of children who screen positive but who, upon further testing, prove not to have a language impairment (see Klee, Carson, Cagin, Hall, Kent, & Reese, 1998 for an example). Given that the next step following failure on screening is a more extensive language evaluation, a high false positive rate could translate into added costs in time and expense for the child, the family, and society.

For purposes of evaluating the feasibility of a measure to serve as a useful screening tool for the general population, calculations must go beyond those that are focused strictly on a small sample in which the prevalence of the disorder is as high as 50%. With a prevalence rate of 50%, even an uninformed guess will be correct in half of the instances. However, as prevalence decreases, the probability of accurately identifying a child as being at risk for disorder necessarily decreases; conversely, the probability of accurately "passing" a child with typical language development increases (Sackett et al., 1991). Based on the epidemiological study of Tomblin, Records, Buckwalter, Zhang, Smith, and O'Brien (1997), the prevalence of SLI among unidentified five-years-olds is approximately 7%. Therefore, if the NRT is to be used as a screening measure to identify those children in the general population who truly warrant further testing, it should be relatively successful in selecting those 7% who are at risk for a

language impairment, while also "passing" the remaining 93% of children. To evaluate the NRT in this way, it was necessary to estimate the probability of receiving a particular score, given both the LRs ascertained from the sample of participants and the base rate of language impairment in the population as a whole. We employed this type of calculation in the present study.

In summary, the three goals of the present study were: 1) to determine the diagnostic accuracy of the NRT with a younger sample of children with and without SLI, using LRs as the principal metric; 2) to assess the degree to which phonological limitations affect diagnostic accuracy; and 3) to evaluate the potential utility of the NRT as a screening measure, given the LRs found in our sample of study participants and the estimated prevalence of SLI in the population.

## Method

### Participants

A total of 92 children initially participated in this study, with the final sample consisting of 29 children with SLI and 47 children with TD. Children with TD were recruited through fliers at preschools, advertisements in a local parent newsletter, and through a newsletter reporting on lab activities. Children with SLI were also recruited through these methods, but primarily were referred by local speech-language pathologists. Informed consent to participate was obtained from each child's parent or guardian in accord with the policies of the human subjects review board of the authors' institution. Data from four children were lost due to technical failure (two from each group). One child with SLI chose not to participate, and 11 additional children who had been identified as exhibiting a language impairment and were enrolled in language intervention were excluded because their test scores did not meet our gold standard.

As described below, children included in the group with SLI were those that met the standard exclusionary criteria for SLI (i.e., normal hearing, normal non-verbal intelligence, and no history of neurological impairment). Our gold standard for identifying children with SLI was the Structured Photographic Expressive Language Test, II (SPELT-II; Werner & Kresheck, 1983). We used a cutoff score of 3.25 *SD* below the mean, the cutoff identified by Plante and Vance (1994) as providing the best sensitivity and specificity in a sample of 20 four- and five-year-olds with SLI and 20 age-matched children with TD. (The 11 children excluded for not meeting the gold standard earned scores of 1.6 *SD* to 3.2 *SD* below the mean for their age.)

The average age of the children with SLI was 5;0 (years; months) (*SD* = 0;5; Range = 4;1 to 5;9). This group included 18 boys and 11 girls; 28 were White and 1 was Hispanic; all were monolingual English speakers. All children with SLI scored ≥ 3.3 *SD* below the mean for their age on the SPELT-II. All had hearing within normal limits bilaterally (20 dB HL) at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz, passed an oral-mechanism examination following the protocol of Robbins and Klee (1987) and scored above 85 on the Columbia Mental Maturity Scales (CMMS; Burgemeister, Blum, & Lorge, 1972), a test of non-verbal intelligence. No child had a history of seizures or showed any evidence of neurological dysfunction according to parent report. Twenty-two of the 29 children with SLI were enrolled in a language intervention program, and six children with SLI were scheduled to begin treatment. A summary of participant characteristics can be found in Table 1.

The Bernthal-Bankson Test of Phonology (BBTOP; Bankson & Bernthal, 1990) was used to evaluate the phonological abilities of all of the children with SLI. The Consonant Inventory score was determined for each of these 29 children. The mean standard score was 74 (*SD* = 12; Range = 64 – 107); 25 of the children had scores below 85. Thus, 86% (25/29) of the children with SLI had below-average ability based on Consonant Inventory score. An articulation probe was also administered to determine whether the children could produce the

consonants /s/, /z/, /t/, and /d/ in the final position of monomorphemic words (e.g., *fox, hand*). These four consonants often appear in word-final position as grammatical morphemes that are included in items on the SPELT-II. All children except one earned scores of at least 75% correct. This one child marked final consonants clearly, but substituted /t/ for /s/ and /z/.

Spontaneous language samples consisting of at least 100 utterances were obtained from each child with SLI. In these sessions, children played with toys while interacting with an examiner. The child was encouraged to initiate topics and determine the direction of conversation. The examiner asked questions or commented to maintain conversation. The sample was transcribed using the Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 2004) coding system. The language samples were not used for selecting children. However, they served as an additional means of determining the phonemes in the children's inventories, as described subsequently.

The age of each child with TD fell within two months of that of a child with SLI. The average age of the children with TD was 5;0 (*SD* = 0;6; Range = 4;1 to 5;11). This group included 28 boys and 19 girls; 42 were White, 1 was Hispanic, 2 were African-American, and 2 were Asian/Pacific Islander; all were monolingual English speakers. These children scored above the 19$^{th}$ percentile on the SPELT–II, and passed the hearing screening, oral-mechanism exam, and the CMMS.

Children with TD were not given a test of phonology unless there was concern on the part of the parent or researcher. Two children were administered the Goldman-Fristoe Test of Articulation – 2 (Goldman & Fristoe, 2000); they received standard scores of 79 and 96. A third child was administered the BBTOP and received a standard score of 93 on the Consonant Inventory. Thus only one child or 2% of the group with TD was known to have below-average phonological ability. No child was excluded because of concerns about phonology. Spontaneous speech samples were also obtained from the children in the group with TD, following the procedures used with the group with SLI.

## Procedure and Materials

The NRT consists of 16 nonwords, four of each word length, from one to four syllables (see Appendix for the complete list). As described in Dollaghan and Campbell (1998), the stimuli were controlled in various ways to help ensure that phonological working memory would be tested rather than other skills. The nonwords are relatively simple phonologically, excluding later-developing consonants and consonant clusters. Tense vowels and diphthongs were used to increase perceptibility and reduce stress effects on production accuracy. To avoid effects of reliance on vocabulary knowledge (rather than on phonological working memory), the NRT nonwords were constructed so that no syllable corresponds to an English word. Phonotactic probabilities were indirectly controlled by using sounds only in syllable positions in which they occur infrequently. For the present study, biphone frequency was directly measured using an on-line phonotactic probability calculator (Vitevitch & Luce, 2004). Not surprisingly, the average biphone frequency for each nonword was quite low (*M* = .00064, *SD* = .00043). Overall, the potential contribution of lexical effects was minimized.

Children were administered the NRT in individual sessions conducted as part of two larger research projects. The 16 nonwords were recorded by a female speaker and digitized for presentation. Stimuli were presented using a computer and an external speaker. Nonwords were presented in a fixed order, from shortest to longest (see Appendix). Children were told that they were going to hear some funny, made-up words and that they were to try to say them "just like the lady said them". Non-contingent positive verbal reinforcement (e.g., "You're doing so well!") was given throughout the task. Nonwords were presented once unless the child talked over the nonword or some other noise interfered with the child's ability to hear the presentation

of the nonword. Some children, upon hearing a nonword, refused to respond or responded inaudibly. They were encouraged to "do your best" and the nonword was presented again. Responses to re-presentations constituted only 2% of all possible responses, affecting both groups equally. These responses were not included in the data. Responses were recorded on a digital audio recorder.

## Scoring

Scoring was completed by a trained graduate student or the first or second author. Only children's responses to the first presentation of a target nonword were scored; if a child refused to respond to the first presentation of the target nonword or spoke too quietly to be heard, he or she received a score of zero phonemes correct for that nonword. On three occasions (for two children with SLI and one child with TD), a computer problem resulted in the presentation of the nonword [naib] as [aib]. These children's responses were scored for the two remaining phonemes, and the denominator for the one-syllable nonwords was adjusted from 12 to 11 for purposes of computing percentage correct. For one child with SLI, a three-syllable nonword was inadvertently not presented; the denominator for this set was adjusted from 28 to 21 for computing the percentage correct for this child.

**Scoring Method 1: No allowances for out-of-inventory phonemes—**In computing percentage of phonemes correct (PPC), we followed the method of Dollaghan and Campbell (1998). Each phoneme was compared to its target and scored as incorrect if the child omitted it or substituted another phoneme. Additions or distortions (productions whose phonetic values deviated from expectations but did not cross phoneme boundaries) were not counted as errors. Viewing nonword repetition as a measure of the ability to retain information in phonological short-term memory, substitutions reflect a loss of information, whereas additions and distortions do not. Distortions in our data affected affricates and, to a lesser degree, diphthongs; by definition, these productions were still recognizable tokens of the target phoneme. Finally, if one or more syllables were omitted in a nonword, the remaining syllables were aligned to the target nonword using the vowels as anchors; once aligned, scoring of each phoneme proceeded as described. The number of phonemes repeated correctly was divided by the number of target phonemes to yield a PPC for each nonword length and a total percentage of phonemes correct (TPPC) for all nonwords.

**Scoring Method 2: Exclusion of out-of-inventory phonemes—**Given the potential influence of phonological limitations on the children's performance, we employed a second system of scoring, in which we eliminated from scoring those phonemes in the NRT items that appeared to be absent from the children's inventories. To determine whether a phoneme should be removed, we set a criterion of zero correct productions out of three opportunities to produce the phoneme. As a first step, we examined the child's productions on the NRT itself where most of the phonemes occur in target nonwords multiple times. The spontaneous speech sample was checked for further evidence if a phoneme was attempted fewer than three times on the NRT. For example, if /g/ was produced as /d/ once on the NRT and omitted twice, we checked the first two instances of a /g/ in the speech sample. If both productions were incorrect, the phoneme was removed from analysis for that child; if one production was correct, the phoneme was not removed.

We chose to use a more generous criterion of zero correct out of three attempts before ruling a phoneme "out of inventory" because our sources of information about children's phoneme inventories were limited. Because children in the group with TD were rarely administered a test of articulation, our only sources across all children were the NRT itself and the speech sample. Neither is an ideal source since the nonword task may induce errors that might not be seen in a familiar word while spontaneous speech is biased to some degree toward words the

child knows well and may be less likely to show the error. As a result of this verification, 10 children with SLI and 3 children with TD had their percentage correct scores adjusted in Scoring Method 2. For the group with SLI, an average of 11 phonemes out of 96 were eliminated from the denominator (*SD* = 4; range: 6 – 18); for the group with TD, an average of 4 phonemes out of 96 were eliminated (*SD* = 3; range: 1 – 6). The adjustment changed individual children's TPPC scores an average of 7 percentage points for the group with SLI (*SD* = 2, range = 4 – 11) and an average of 3 percentage points for the group with TD (*SD* = 2, range = 1 – 5).

### Reliability

Recordings from 10 randomly selected children with TD (21%) and 8 randomly selected children with SLI (20%) were transcribed independently by a second trained graduate student or the first or second author. For the group with SLI, phoneme-by-phoneme percentages of agreement for judgments between researchers ranged from 89% to 100%, with an average interrater reliability of 92%. For the group with TD, phoneme-by-phoneme percentages of agreement for judgments between researchers ranged from 93% to 100%, with an average interrater reliability of 96%. Thus, total interrater reliability for the groups with SLI and TD was 94%. These reliability rates were consistent with the levels reported by Dollaghan and Campbell (1998).

## Results

### Preliminary Analysis: Group Comparisons

Prior to determining the diagnostic accuracy of the NRT for this sample of children, it was important to establish differences at the group level. This was especially important for Scoring Method 2, given the large proportion of children with SLI who showed below-average scores on the BBTOP.

**Scoring method 1: No allowances for out-of-inventory phonemes—**The first scoring method made no allowances for phonological errors (other than distortions of phonemes, as described in *Scoring*). A mixed model analysis of variance (ANOVA) was performed on PPC with participant group as a between-subjects variable and nonword length as a within-subjects variable. The ANOVA revealed significant main effects of Group, $F$ (1, 74) = 99.14, $p < .0001$, $\eta_p^2 = .57$; Length, $F$ (3, 222) = 180.04, $p < .0001$, $\eta_p^2 = .71$; and a significant interaction between Group and Length, $F$ (3, 222) = 10.17, $p < .0001$, $\eta_p^2 = .12$. Tukey honestly significant difference (HSD) testing at the .01 level was then used to investigate the significant interaction. Effect sizes were calculated through *d*; values of 0.80 and larger were considered large effect sizes, and those between 0.50 and 0.79 were considered medium effect sizes (Cohen, 1988). Observed means for each syllable length are reported in Table 2. There were significant differences in accuracy between groups for two-syllable nonwords ($d = 1.56$), three-syllable nonwords ($d = 2.26$), and four-syllable nonwords ($d = 1.60$), but not for one-syllable nonwords. There were also significant differences in accuracy within each group between two- and three-syllable nonwords (SLI: $d = 1.34$; TD: $d = .92$) and between three- and four-syllable nonwords (SLI: $d = 1.29$; TD: $d = 1.45$) but not between one- and two-syllable nonwords. A *t*-test comparing groups on TPPC showed a significant difference, $t$ (74) = −9.74, $p < .001$, $d = 2.26$. Group means for TPPC are reported in Table 2.

**Scoring method 2: Out-of-inventory phonemes excluded—**In Scoring Method 2, phonemes that appeared to be absent from a child's inventory were treated as unscorable, and the child's PPC for each nonword length was computed after the out-of-inventory phonemes were deleted from the numerator and denominator. A mixed model ANOVA was used to evaluate differences between participant groups across nonword lengths. As in Scoring Method

1, we found significant main effects of Group, $F$ (1, 74) = 91.21, $p < .0001$, $\eta_p^2 = .55$; Length, $F$ (3, 222) = 191.73, $p < .0001$, $\eta_p^2 = .72$; and a significant interaction between Group and Length, $F$ (3, 222) = 11.87, $p < .0001$, $\eta_p^2 = .14$. Observed means for each syllable length are reported in Table 3. The significant interaction was examined further through post-hoc testing (Tukey HSD) at the .01 level. We found significant differences in accuracy between groups for two-syllable nonwords ($d$ = 1.46), three-syllable nonwords ($d$ = 2.10), and four-syllable nonwords ($d$ = 1.50), but not for one-syllable nonwords. There were also significant differences in accuracy within groups between two- and three-syllable nonwords (SLI: $d$ = 1.47; TD: $d$ = .97) and between three- and four-syllable nonwords (SLI: $d$ = 1.34; TD: $d$ = 1.44) but not between one- and two-syllable nonwords. The TPPC scores were significantly different for groups, $t$ (74) = −9.10, p < .0001, $d$ = 2.12. Group means for TPPC are reported in Table 3.

Comparing the differences between Scoring Methods 1 and 2 in mean PPC across nonword lengths for the group with SLI (Table 2 and Table 3), it appears that the score adjustments in Scoring Method 2 had only a minimal effect at each length. There remained a large and significant decrement as nonwords got longer, with significant group differences favoring the children with TD for nonwords of two-, three- and four-syllables in length.

## Diagnostic Accuracy

**Scoring method 1: No allowances for out-of-inventory phonemes—**Given that large differences were found between the two groups, we proceeded to an analysis of the diagnostic accuracy of the NRT for the sample of 29 children with SLI and 47 children with TD. As reviewed here, many studies have reported the diagnostic accuracy of nonword repetition in terms of sensitivity and specificity or LRs. To allow for comparison of metrics across studies we performed both types of calculation. Given the relatively small sample size in this study, confidence intervals were also computed to show the range within which the true diagnostic accuracy was expected to fall with a probability of 95%. We also computed positive predictive value (ppv) and negative predictive value (npv) for the sensitivity and specificity analysis and post-test probability for the LR analysis. The measures ppv and npv provide information about a particular test result, specifically, the proportion of children with (or without) SLI whose scores fall within a specified range of scores. Post-test probability simply reflects the conversion of the LR from odds to a percentage; this is computed by dividing the LR/(LR + 1) and multiplying by 100.

As noted earlier, the children were assigned to the two groups on the basis of their scores on the SPELT-II, with 3.25 *SD* below the mean serving as the dividing point, following Plante and Vance (1994). We then submitted the TPPC scores to a logistic regression analysis to determine a cutoff score on the NRT that would yield the greatest separation into groups. Based on this analysis, children with TPPC scores of ≤ 66% were classified as SLI; children with scores > 66% were classified as TD. Classification resulted in 86% sensitivity and 91% specificity, with 95% confidence intervals of 73%–99% and 83%-99%, respectively (see Table 4). Following Plante and Vance (1994), the resulting sensitivity value was considered adequate, and the resulting specificity value was considered good.

As shown in Table 4, based on the sample of children participating in the study, the ppv was 86%. That is, for this sample, the percentage of children falling below the cutoff determined by the logistic regression analysis who were classified as SLI by the SPELT-II was 86%. The npv was 91%. That is, the percentage of children scoring above the cutoff who were classified as within typical limits by the SPELT-II was 91%.

We then computed LRs for four levels of TPPC, following the method of Sackett et al. (1991). As noted earlier, an LR+ represents the odds that a score within a given range of scores came from an affected individual; it is computed by dividing the sensitivity found within that

range by (1 - the specificity) found within that range. The LR- is computed as (1-sensitivity)/specificity and represents the odds that a score in the unaffected range came from an affected child. The LRs for four levels of TPPC for Scoring Method 1 are reported in Table 5. Cutoff scores were chosen to maximize the LR+ and minimize the LR-. Assuming the lowest level of scores (≤54%) to represent the "affected" range, the LR+ (22.66) showed that a score in this range was over 22 times more likely to come from a child with SLI than from a child with TD. Assuming the highest level of scores (≥77) to represent the "unaffected" range, the LR- (0.05) showed that a score in this range was one-twentieth as likely to come from a child with SLI as from a child with TD. Note that the cut point for the lowest range (≤54%) was near that found by Ellis Weismer et al. (2000) for seven- and eight-year-olds (≤60%) while the highest range (≥77%) was similar to that found by Dollaghan and Campbell (1998) for six- to nine-year-olds (≥81%).

We computed post-test probability of disorder for the lowest (≤54%) and highest (≥77%) score ranges. This measure simply converts an LR from odds to a percentage reflecting the probability that a score in this range came from a child with SLI as opposed to a child with TD. For this sample, the post-test probabilities were 96% and 5% for the lowest and highest levels of scores, respectively. That is, a TPPC score of 54% or lower had a 96% probability of having come from a child who had been classified as SLI by the SPELT-II, whereas TPPC scores of 77% and higher had only a 5% probability of having come from a child who had been classified as exhibiting SLI by the SPELT-II.

**Scoring method 2: Out-of-inventory phonemes excluded—**In the second scoring method, we adjusted the scores of children for whom a limited phonetic inventory may have directly contributed to lower PPC scores. For each child, we searched for evidence of the ability to produce each phoneme used in the NRT; if such evidence was not found, the phoneme was removed from scoring and PPC was calculated using the revised denominator.

Using logistic regression, we determined diagnostic accuracy for TPPC scores in this data set. Based on this analysis, children with TPPC scores of ≤ 68 were classified as SLI; children with scores > 68 were classified as TD. A comparison of Table 6 with Table 4 shows that sensitivity and specificity were somewhat lower with Scoring Method 2, with sensitivity (79%) falling slightly below levels of adequacy. The ppv and npv were 82% and 88%, respectively. Two of the children who were correctly classified as SLI according to Scoring Method 1 and who had their scores adjusted were misclassified as TD using Scoring Method 2. For the group with TD, one child who was misclassified according to Scoring Method 1 was correctly classified using Scoring Method 2 after having his score adjusted.

Results from the LR analysis of these data are reported in Table 7. We used the same score ranges as in Scoring Method 1 to compare results. At a TPPC of 54% or lower the odds for correctly ruling in SLI were quite good (19 times more likely); likewise, at a TPPC of 77% or higher the odds of correctly ruling out SLI were quite good (one-twentieth as likely). The post-test probabilities indicated that TPPC scores of 54% or lower had a 95% probability of having come from a child who had been classified as SLI by the SPELT-II, whereas TPPC scores of 77% and higher had only a 5% probability of having come from a child who had been categorized as exhibiting SLI by the SPELT-II.

Removing from analysis those phonemes which were not in the inventory of children did affect classification, although not to a great degree. Of the 10 children with SLI who had their scores adjusted, five remained in the lowest score range, two moved from the lowest to the middle low score range, and three moved from the middle low to the middle high range. For the three children with TD, two remained in the highest score range; the one child with a below-average

score on the GFTA-2 moved from the middle low to the middle high range. Thus, the scores of six children, or 8% of the sample, changed in a way that had an impact on the LR analysis.

**Feasibility as a Screening Instrument for the Wider Population of Children**

**Scoring method 1: No allowances for out-of-inventory phonemes—**The third goal of the present study was to explore the feasibility of using the NRT as a screening tool for the wider population of four- and five-year-olds. For this purpose, it was necessary to consider the prevalence rate of SLI in the general population. The finding by Tomblin et al. (1997) that approximately 7% of five-year-olds meet the criteria for SLI appeared to be the most appropriate basis for selecting a prevalence rate. Accordingly, 7% was used as the pre-test probability.

We computed the post-test probability of disorder for scores in the lowest (≤54%) and highest (≥77%) ranges, multiplying the LH+ and LH- found for this sample by the prevalence rate/1-prevalence rate (.07/.93). The adjusted ratios were each divided by (1+ the adjusted ratio). The resulting post-test probabilities were 61% and 0.3% for the lowest and highest levels of scores, respectively. That is, if the NRT were administered to the general population of preschoolers, 61% of the children with TPPC scores of 54% or lower would be classified as SLI by the SPELT-II, whereas only 0.3% of the children with TPPC scores of 77% and higher would be classified as SLI by the SPELT-II.

**Scoring method 2: Out-of-inventory phonemes excluded—**Using the 7% prevalence rate, we computed post-test probabilities for the LRs found when scoring excluded out-of-inventory phonemes. We computed the post-test probability of disorder for scores in the lowest and highest ranges, multiplying the LH+ and LH- found for this sample by .07/.93; the adjusted ratios were divided by (1+ the adjusted ratio). The resulting post-test probabilities were 58% for TPPC scores ≤54% and 0.3% for scores ≥77%. That is, if the NRT were administered to the general population of preschoolers, and out-of-inventory phonemes were not considered in the scoring, 58% of the children with TPPC scores of 54% or lower would be classified as SLI by the SPELT-II, whereas only 0.3% of the children with TPPC scores of 77% and higher would be classified as SLI by the SPELT-II.

## Discussion

Three goals were pursued in this investigation. The first goal was to determine whether diagnostic accuracy of the NRT with a sample of four- and five-year-olds would provide results comparable to those seen in earlier studies of older children. Based on our sample of 29 children classified as exhibiting SLI and 47 children classified as possessing typical language skills according to the SPELT-II, we found LR+ and LR- values that were in line with those of previous studies. For example, the LR+ of 22.66 and the LR- of 0.05 resemble the Dollaghan and Campbell findings of an LR+ of 25 and an LR- of 0.03.

These comparable findings suggest that the NRT can be extended to younger ages than have been employed in earlier studies. However, two other differences between the present study and earlier studies should also be acknowledged. First, the cut points found to be most discriminating were not the same across studies. We would argue that this is to be expected given the age differences involved. For example, in the Dollaghan and Campbell (1998) study of children ages six to nine years, a higher cut point for LR+ (70% and lower) and LR- (81% and higher) proved most satisfactory, whereas in the present study with four- and five-year-olds, the cut points for LR+ and LR- had to be lower (54% and below and 77% and above, respectively).

A second difference across studies is the gold standard employed. Previous studies of the diagnostic accuracy of the NRT have used treatment status as the gold standard. In the present study, SPELT-II scores above and below −3.25 *SD* were employed as the gold standard, with the rationale that earlier studies of the SPELT-II (Plante & Vance, 1994) have found acceptable sensitivity and specificity levels for this cut point. Given the different gold standards employed, we cannot claim that the present study differed from previous studies only in the ages of the participants. On the other hand, as noted earlier, 22 of the 29 children classified as exhibiting SLI on the SPELT-II were in fact enrolled in language intervention, and six children were scheduled to begin treatment. No child with TD was participating in an intervention program.

The second goal of the present study was to determine if diagnostic accuracy changes substantially as a function of how phonological errors were treated. We employed two scoring methods. The first method allowed no errors of any type. The second method treated as unscorable any instance in which the substituted phoneme appeared to be absent from the child's inventory. The LR+ values for the two methods at a cut point of 54% or below were 22.66 and 19.43, respectively. For both scoring methods, the LR- was 0.05, using a cut point of 77% or above. Given that both LR+ values were well above 10 and both LR- values were well below 0.10 (see Sackett et al., 1991), it would appear that satisfactory results were obtained with either scoring method.

However, such a conclusion cannot be viewed as definitive because we examined the influence of phonological factors in only one way, by using an alternative method of scoring. One might argue that these methods could only reduce but not eliminate the contribution of phonological factors given that our two groups of children were clearly different in the number of phonological errors made. An alternative method would be to employ a research design that included, for example, a group of children with both SLI and speech delay, a group of children with SLI only, a group of children with speech delay only, and a group of children with typical language development and no speech delay. Such a design might allow us to determine more conclusively whether the NRT's success in distinguishing preschool-aged children with SLI from typically developing peers is based in part on the fact that preschoolers with SLI are more likely to exhibit problems with phonology along with their deficits in language.

The third goal of the present study was to assess the feasibility of using the NRT as a screening measure for the general population of four- and five-year-olds. This assessment required a calculation that incorporated the presumed prevalence of the disorder in the general population. For our purposes, the presumed prevalence of SLI of 7% among five-year-olds seemed most appropriate. We found that an NRT score of 54% or lower resulted in a notable increase in the probability of correctly identifying children at risk for language impairment relative to the base rate. That is, the probability of accuracy increased from 7% to 61%. Using the alternative method that treated errors on out-of-inventory phonemes as unscorable, the probability of accuracy increased from 7% to 58%. Probability shifts of this magnitude can be regarded as clinically important (see Straus, Richardson, Glasziou, & Haynes, 2005).

Although we believe that an approximately eight-fold increase in accuracy is noteworthy, we acknowledge that probability levels of 58% to 61% would be insufficient for application of the NRT as a sole language screening tool for four- and five-year-olds. These calculations suggest that many children would fail the screening yet be found to exhibit age-appropriate language skills after further testing.

One important qualification is that our prevalence estimate of 7% is based on the presumed prevalence of SLI, not of language impairment more generally. If the intent of screening is to identify children for further testing who might be at risk for any type of language impairment, a higher prevalence estimate could be employed. For example, Klee, Pearce, and Carson

(2000) employed a prevalence estimate of 13% based on the large-scale studies of Beitchman, Nair, Clegg, and Patel (1986) and Tomblin, Records, and Zhang (1996) that were not restricted to SLI. Using a prevalence estimate of 13% for the data in the present study, the post-test probability of disorder with a score of 54% or less would be considerably higher, at 77% (or 74% if out-of-inventory phonemes were treated as unscorable). However, even with this improvement, the resulting probability suggests that a large number of the children identified for further testing would prove to exhibit language abilities within age-appropriate levels.

Perhaps it is not surprising that a single measure such as the NRT is insufficient as a screening tool. In a large twin study, Bishop, North, and Donlan (2006) found that the ability reflected in a task of tense/agreement morpheme use and the ability reflected in nonword repetition were both heritable but genetically separable. That is, some identical twin pairs at risk for language impairment shared deficits in tense/agreement morpheme use but did not exhibit limitations in nonword repetition, whereas other identical twin pairs showed concordance in poor nonword repetition ability but adequate tense/agreement morpheme use. These findings suggest that nonword repetition ability reflects only one type of language deficit and for this reason may not serve as an adequate means of screening if used alone.

The assumption that nonword repetition limitations represent only one type of language deficit may seem at odds with the finding that diagnostic accuracy was relatively high when children with SLI constituted approximately 38% (29 of 76) of the participants in the study. Shouldn't a larger proportion of children with SLI have been misclassified by the NRT, given that other types of language deficits are genetically separable from nonword repetition? We suspect that our findings can be traced to the high degree to which genetically separable deficits co-occur in the population of children with SLI. Bishop et al. (2006) reported a disproportionately high percentage of children showing the "double deficit" of tense/agreement problems and problems with nonword repetition. The gold standard used in the present study – the SPELT-II – is largely a grammatical measure that includes items requiring the use of tense/agreement morphemes. Thus, many of the children with SLI may have had weaknesses in both nonword repetition and tense/agreement use.

We believe our findings regarding the NRT as a potential screening tool should prompt careful examination of other language screening measures. For the screening tools that have been subjected to calculations of post-test probability using the prevalence of the disorder as the base rate, it is common to find estimates suggesting that more than half of the children who fail screening would be found to have age-appropriate language skills upon further testing (see Klee et al., 2000; Law et al., 2000). Thus, it does not appear sufficient to choose an alternative screening test given our findings for the NRT if the post-test probability of the alternative is either unknown or found to be unsatisfactory in earlier studies. Recall that, in the present study, when based on the participant sample of 29 in the group with SLI and 47 in the group with TD, diagnostic accuracy was relatively high, with sensitivity at 86% and specificity at 91%. Thus, similar diagnostic accuracy values reported in a test manual that are based on a participant sample containing a large percentage of children with SLI should not be interpreted to mean that the test is sufficient for screening children in the general population. We hope that one outcome of the present study will be an effort by other clinical researchers to evaluate available screening tools and, where necessary, develop new ones that are more satisfactory.

Another potential direction for future clinical research would be to determine whether referrals for further testing following failure on a screening measure might be prioritized based on the presence of other risk factors. For example, Klee et al. (2000) determined that the percentage of children identified for further testing and then found to be age-appropriate would drop from approximately 49% to approximately 23% if it were assumed that a child would be referred for further testing only if, in addition to failing the language screening test, the child had

experienced multiple ear infections or the parents had expressed concern about the child's language development. Studies of this type that employ the NRT as the screening measure might yield post-test probabilities that are substantially higher than the ones reported in the present study.

In summary, the findings of the present study suggest that the diagnostic accuracy of the NRT for four- and five-year-olds is similar to that reported in studies with older children when calculations are based on samples of children with SLI and TD who were independently classified according to scores on the SPELT-II. The precise contribution of phonological limitations warrants further investigation because our group with TD exhibited rather accurate phonology, whereas our group with SLI included many children with phonological limitations. Based on this sample of children, at least, diagnostic accuracy did not vary substantially according to whether substitutions of out-of-inventory phonemes were treated as errors or as unscorable. Finally, the NRT does not appear to be sufficient as the sole language screening tool to identify children from the general population who may be at risk for language impairment. Although the probability of identifying a child with a true language impairment on the basis of a low score (≤54%) on the NRT increased from 7% (the prevalence rate of SLI) to 61%, a higher probability rate is needed. One clear implication from this finding is that other screening instruments should be evaluated according to the same standards, to avoid the potential clinical use of an alternative whose accuracy as a screening tool for the general population is unknown.

## Acknowledgments

The research in this paper was supported in part by Research Grants R01 DC00458 to Laurence Leonard and R01 DC04826 to Lisa Goffman from the National Institute on Deafness and Other Communication Disorders. We thank the children and families who participated and Abigail Bormann, Rachel Brunner, and Kelsey Pithoud for their assistance during this project.

## References

Archibald L, Gathercole S. Nonword repetition: A comparison of tests. Journal of Speech, Language, and Hearing Research 2006;49:970–983.

Bankson, N.; Bernthal, J. Bernthal-Bankson Test of Phonology. Chicago: Riverside; 1990.

Beitchman J, Nair R, Clegg M, Patel P. Prevalence of speech and language disorders in 5-year-old kindergarten children in the Ottawa-Carleton region. Journal of Speech and Hearing Disorders 1986;51:98–110. [PubMed: 3702369]

Bishop D, North T, Donlan C. Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. Journal of Child Psychology and Psychiatry 1996;37:391–403. [PubMed: 8735439]

Bishop D, Adams C, Norbury C. Distinct genetic influences on grammar and phonological short-term memory deficits: Evidence from 6-year-old twins. Genes, Brain, and Behavior 2006;5:158–169.

Burgemeister, B.; Blum, L.; Lorge, I. Columbia Mental Maturity Scale. New York: Harcourt Brace Jovanovich; 1972.

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

Chiat S, Roy P. The Preschool Repetition Test: An evaluation of performance in typically developing and clinically referred children. Journal of Speech, Language, and Hearing Research 2007;50:429–443.

Cohen, J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum; 1988.

Conti-Ramsden G. Processing and linguistic markers in young children with specific language impairment. Journal of Speech, Language, and Hearing Research 2003;46:1029–1037.

Dollaghan, C. The handbook for evidence-based practice in communication disorders. Baltimore, MD: Brookes; 2007.

Dollaghan C, Campbell T. Nonword repetition and child language impairment. Journal of Speech, Language, and Hearing Research 1998;41:1136–1146.

Dunn, LM.; Dunn, LM. Peabody Picture Vocabulary Test-Revised. Circle Pines, MN: American Guidance Service; 1981.

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

Ellis Weismer S, Tomblin JB, Zhang X, Buckwalter P, Chynoweth J, Jones M. Nonword repetition performance in school-age children with and without language impairment. Journal of Speech, Language, and Hearing Research 2000;43:865–878.

Gathercole S. Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. Memory and Cognition 1995;23:83–94.

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

Gathercole, S.; Baddeley, A. The role of phonological memory in normal and disordered language development. In: von Euler, C.; Lundberg, I.; Lennerstrand, G., editors. Brain and reading. New York: MacMillan; 1989. p. 336-360.

Gathercole S, Baddeley A. Phonological memory deficits in language disordered children: Is there a causal connection? Journal of Memory and Language 1990;29:336–360.

Gathercole S, Willis C, Baddeley A, Emslie H. The Children's Test of Nonword Repetition: A test of phonological working memory. Memory 1994;2:103–127. [PubMed: 7584287]

Goldman, R.; Fristoe, M. Goldman-Fristoe Test of Articulation - 2. Pines, MN: American Guidance Association; 2000.

Graf Estes E, Evans J, Else-Quest N. Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. Journal of Speech, Language, and Hearing Research 2007;50:177–195.

Gray S. 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 2003;36:129–151. [PubMed: 12609578]

Gray S. Word learning by preschoolers with specific language impairment: Predictors and poor learners. Journal of Speech, Language, and Hearing Research 2004;47:1117–1132.

Gray S. The relationship between phonological memory, receptive vocabulary, and fast mapping in children with specific language impairment. Journal of Speech, Language, and Hearing Research 2006;49:955–969.

Greenslade K, Plante E, Vance R. The diagnostic accuracy and construct validity of the Structured Photographic Expressive Language Test – Preschool: Second Edition (SPELT-P2. Language, Speech, and Hearing Services in Schools. (in press)

Klee T, Carson D, Gavin W, Hall L, Kent A, Reece S. Concurrent and predictive validity of an early language screening program. Journal of Speech, Language, and Hearing Research 1998;41:627–641.

Klee T, Pearce K, Carson D. Improving the positive predictive value of screening for developmental language disorder. Journal of Speech, Language, and Hearing Research 2000;43:821–833.

Law J, Boyle J, Harris F, Harkness A, Nye C. The feasibility of universal screening for primary speech and language delay: Findings from a systematic review of the literature. Developmental Medicine and Child Neurology 2000;42:190–200. [PubMed: 10755459]

Miller, J.; Chapman, R. Systematic Analysis of Language Transcripts (Version 8.0) [Computer software]. Madison: Language Analysis Laboratory, Waisman Center, University of Wisconsin; 2004.

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

Munson B, Kurtz B, Windsor J. 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 2005;48:1033–1047.

Newcomer, P.; Hammill, D. Test of language development-Primary. Second edition. Austin, TX: ProxEd; 1988.

Oetting J, Cleveland L. The clinical utility of nonword repetition for children living in the rural South of the U.S. Clinical Linguistics and Phonetics 2006;20:553–561. [PubMed: 17056486]

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

Robbins J, Klee T. Clinical assessment of oropharyngeal motor development in young children. Journal of Speech and Hearing Disorders 1987;52:271–277. [PubMed: 3455449]

Rodekohr R, Haynes. Differentiating dialect from disorder: A comparison of two processing tasks and a standardized language test. Journal of Communication Disorders 2001;34:255–272. [PubMed: 11409607]

Sackett, D.; Haynes, R.; Guyatt, G.; Tugwell, P. Clinical Epidemiology. Boston: Little, Brown; 1991.

Shriberg L, Kwiatkowski J. Developmental phonological disorders I: A clinical profile. Journal of Speech and Hearing Research 1994;37:1100–1126. [PubMed: 7823556]

Shriberg L, Tomblin B, McSweeney J. Prevalence of speech delay in 6 year-old children and comorbidity with language impairment. Journal of Speech, Language, and Hearing Research 1999;42:1461–1481.

Straus, S.; Richardson, WS.; Glasziou, P.; Haynes, R. Evidence-based medicine: How to teach and practice EBM. Third Edition. London, England: Elsevier; 2005.

Thal D, Miller S, Carlson J, Vega M. Nonword repetition and language development in 4-year-old children with and without a history of early language delay. Journal of Speech, Language, and Hearing Research 2005;48:1481–1495.

Tomblin JB, Records N, Zhang X. A system for the diagnosis of specific language impairment in kindergarten children. Journal of speech and Hearing Research 1996;39:1284–1294. [PubMed: 8959613]

Tomblin JB, Records N, Buckwalter P, Zhang X, Smith E, O'Brien M. The prevalence of specific language impairment in kindergarten children. Journal of Speech, Language, and Hearing Research 1997;40:1245–1260.

Vitevitch M, Luce P. A Web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, & Computers 2004;36:481–487.

Washington J, Craig H. A language screening protocol for use with young African American children in urban settings. American Journal of Speech-Language Pathology 2004;13:329–340. [PubMed: 15719899]

Werner, EO.; Krescheck, JD. Structured Photographic Expressive Language Test – II. DeKalb, IL: Janelle Publications; 1983.

**Table 1**

Summary of Participant Characteristics

| | SLI | | | TD | | |
|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Range** | **Mean** | **SD** | **Range** |
| Age | 5;0 | 0;5 | 4;1 – 5;9 | 5;0 | 0;6 | 4;1 – 5;11 |
| Spelt-II[a] | < 1st %-ile[b] | | ---[c] | 63rd %-ile[b] | | 19th %-ile – 99th %-ile |
| CMMS[d] | 106 | 13 | 85 – 135 | 118 | 10 | 98 – 143 |

[a] Percentile ranks on the Structured Photographic Expressive Language Test-II.

[b] Median percentile ranks.

[c] All scores for this group were <1st %-ile.

[d] Standard scores on the Columbia Mental Maturity Scale.

**Table 2**

Scoring Method 1 – Means (and Standard Deviations) for Percentage of Phonemes Correct (PPC) for Each Syllable Length and for Total Percentage of Phonemes Correct (TPPC)

| | PPC 1 syllable | PPC 2 syllables | PPC 3 syllables | PPC 4 syllables | TPPC |
|---|---|---|---|---|---|
| SLI (n = 29) | 80.31 (14.07) | 76.03 (15.49) | 56.31 (13.91) | 37.27 (15.48) | 56.08 (11.21) |
| TD (n = 47) | 92.22 (7.05) | 94.57 (6.50) | 85.70 (12.01) | 63.84 (17.65) | 79.76 (9.70) |

**Table 3**

Scoring Method 2 – Means (and Standard Deviations) for Percentage of Phonemes Correct (PPC) for Each Syllable Length and for Total Percentage of Phonemes Correct (TPPC)

| | PPC 1 syllable | PPC 2 syllables | PPC 3 syllables | PPC 4 syllables | TPPC |
|---|---|---|---|---|---|
| SLI (n = 29) | 83.42 (10.84) | 79.24 (14.07) | 58.60 (13.91) | 38.71 (15.78) | 58.47 (10.62) |
| TD (n = 47) | 92.40 (7.14) | 94.95 (5.85) | 85.83 (11.96) | 63.98 (17.71) | 79.97 (9.60) |

## Table 4

Scoring Method 1 – Percentages of Children Correctly Classified by Total Percentage of Phonemes Correct (TPPC)

| | | Clinical classification | | |
|---|---|---|---|---|
| | | **SLI** | **TD** | **Predictive Values** |
| NRT Score | SLI (≤ 66%)[a] | 25 | 4 | ppv = 86% |
| | TD (> 66%)[a] | 4 | 43 | npv = 91% |
| | Percentage Classified Correctly | Sensitivity = 25/29 = 86% (73 – 99%)[b] | Specificity = 43/47 = 91% (83 – 99%)[b] | |

[a] Cutoff scores.

[b] 95% confidence intervals.

**Table 5**

Scoring Method 1 – Likelihood Ratio Analysis for Total Percentage of Phonemes Correct (TPPC)

| | SLI (n = 29) | | TD (n = 47) | | | |
|---|---|---|---|---|---|---|
| | **No.** | **Prop.** | **No.** | **Prop.** | **Likelihood Ratio (95% CI)** | **Post-Test Probability of Disorder** |
| ≤ 54 | 14 | .4827 | 1 | .0213 | 22.66 (3.15–163.56) | 96% |
| 55 – 66 | 11 | .3793 | 3 | .0638 | 5.94 (1.81–19.53) | 86% |
| 67 – 76 | 3 | .1034 | 10 | .2128 | .48 (0.15–1.62) | 32% |
| ≥77 | 1 | .0345 | 33 | .7021 | .05 (0.007–0.34) | 5% |

## Table 6

Scoring Method 2 – Percentages of Children Correctly Classified by Total Percentage of Phonemes Correct (TPPC)

| | | Clinical classification | | |
|---|---|---|---|---|
| | | **SLI** | **TD** | **Predictive Values** |
| NRT score | SLI (≤ 68%)[a] | 23 | 5 | ppv = 82% |
| | TD (> 68%)[a] | 6 | 42 | npv = 88% |
| | Percentage Classified Correctly | Sensitivity = 23/29 = 79% (65–93%)[b] | Specificity = 41/47 = 89% (81–97%)[b] | |

[a] Cutoff scores.

[b] 95% confidence intervals.

**Table 7**

Scoring Method 2 – Likelihood Ratio Analysis for Total Percentage of Phonemes Correct (TPPC)

| | SLI (n = 29) | | TD (n = 47) | | | |
|---|---|---|---|---|---|---|
| | **No.** | **Prop.** | **No.** | **Prop.** | **Likelihood Ratio (95% CI)** | **Post-Test Probability of Disorder** |
| ≤ 54 | 12 | .4138 | 1 | .0213 | 19.43 (2.67–141.82) | 95% |
| 55 – 66 | 10 | .3448 | 2 | .0425 | 8.11 (1.91–34.41) | 89% |
| 67 – 76 | 6 | .2069 | 11 | .2340 | .88 (0.37–2.13) | 47% |
| ≥77 | 1 | .0345 | 33 | .7021 | .05 (0.007–0.34) | 5% |

## Table 8

Post-Test Probabilities for Scoring Method 1 Likelihood Ratios, Assuming a Prevalence Rate for SLI of 7%

| Pre-Test Probability of Disorder | Likelihood Ratios | | Post-Test Probability of Disorder |
|---|---|---|---|
| 7% | LH+[a] | 22.66 | 61% |
| | LH-[b] | .05 | .3% |

[a] Note: Likelihood ratios for TPPC scores on the NRT ≤ 54.

[b] Likelihood ratios for TPPC scores on the NRT ≥77.

## Table 9

Post-Test Probabilities for Scoring Method 2 Likelihood Ratios, Assuming a Prevalence Rate for SLI of 7%

| Pre-Test Probability of Disorder | Likelihood Ratios | | Post-Test Probability of Disorder |
|---|---|---|---|
| 7% | LH+[a] | 19.43 | 58% |
| | LH-[b] | .05 | .3% |

[a] Note: Likelihood ratios for TPPC scores on the NRT ≤ 54.

[b] Likelihood ratios for TPPC scores on the NRT ≥77.

## Appendix

Appendix Order of presentation of NRT stimuli

| One syllable | Two syllables | Three syllables | Four syllables |
|---|---|---|---|
| 1. /nɑɪb/ | 5. /teɪvɑk/ | 9. /tʃinɔɪtɑʊb/ | 13. /veɪtatʃɑɪdɔɪp/ |
| 2. /voʊp/ | 6. /tʃoʊvæg/ | 10. /nɑɪtʃoʊveɪb/ | 14. /dævoʊnɔɪtʃig/ |
| 3. /tɑʊdʒ/ | 7. /vætʃɑɪp/ | 11. /dɔɪtɑʊvæb/ | 15. /nɑɪtʃɔɪtɑʊvʊb/ |
| 4. /dɔɪf/ | 8. /nɔɪtɑʊf/ | 12. /teɪvɔɪtʃɑɪg/ | 16. /tævatʃinɑɪg/ |