Nonword Repetition and Sentence Repetition as Clinical Markers of Specific Language Impairment: The Case of Cantonese

Stephanie F. Stokes
School of Linguistics and Applied Language Studies, University of Reading, Reading, England

Anita M-Y. Wong
University of Hong Kong

Paul Fletcher
University College Cork, Cork, Ireland

Laurence B. Leonard
Purdue University, West Lafayette, Indiana

Purpose: Recent research suggests that nonword repetition (NWR) and sentence repetition (SR) tasks can be used to discriminate between children with SLI and their typically developing age-matched (TDAM) and younger (TDY) peers.

Method: Fourteen Cantonese-speaking children with SLI and 30 of their TDAM and TDY peers were compared on NWR and SR tasks. NWR of IN nonwords (CV combinations attested in the language) and OUT nonwords (CV combinations unattested in the language) were compared. SR performance was compared using 4 different scoring methods.

Results: The SLI group did not score significantly lower than the TDAM group on the test of NWR (overall results were TDAM = SLI > TDY). There were nonsignificant group differences on IN syllables but not on OUT syllables. The results do not suggest a limitation in phonological working memory in Cantonese-speaking children with SLI. The SR task discriminated between children and their TDAM peers but not between children with SLI and their TDY peers matched for mean length of utterance.

Conclusions: SR but not NWR discriminates between children with SLI and their TDAM peers. Poorer NWR for English-speaking children with SLI might be attributable to weaker use of the redintegration strategy in word repetition. Further cross-linguistic investigations of processing strategies are required.

KEY WORDS: language disorders, children

Over the last two decades research has consistently shown that English-speaking children with specific language impairment (SLI) score significantly lower than their age-matched typically developing peers (TDAM) and language-matched typically developing peers (TDLM) on tests of working memory.1 Research has focused on both phonological working memory—for example, nonword repetition (NWR; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990) and sentence repetition (SR; e.g., Kamhi & Catts, 1986)—and functional working memory via tests of simultaneous processing and storage (Just & Carpenter, 1992; see Montgomery, 2002, for a review). Because NWR and SR have been described as clinical markers for SLI (e.g., Bishop, North, & Donlan, 1996; Campbell, Dollaghan, Needleman, & Janosky, 1997; Conti-Ramsden, Botting, & Faragher, 2001), here we focus on those aspects of working memory.

1We use the acronym SLI for all uses of language impairment, specific language impairment, language disorder, and so on in the literature review. Likewise, we use the acronyms TDAM for typically developing age-matched peers and TDY for typically developing younger (or language-matched) peers.
NWR

Two early studies investigated the working memory skills of children with SLI. Kamhi, Catts, Mauer, Apel, and Gentry (1988) compared NWR skills in 10 children with SLI (mean age = 7;9 [years;months]) against those of 10 children with reading impairments and 10 TDAM peers. The children with SLI scored significantly lower than their TDAM peers on monosyllabic and multisyllabic nonwords. Gathercole and Baddeley (1990) compared the NWR skills of 6 children with SLI (aged 8;6) with those of 6 TDAM and 6 typically developing younger (TDY) children. The children with SLI scored significantly lower than their TDAM and TDY peers on three- and four-syllable nonwords. The mean performance of the children with SLI was approximately 4 years below their chronological age. Gathercole and Baddeley claimed that working memory deficits in these children were not attributable to language status, as the children performed worse than their TDY peers. These studies paved the way for several studies on the use of NWR to identify children with SLI. The studies varied in sample sizes and the inclusion of control groups: either TDAM or TDY (and sometimes both).

Marton and Schwartz (2003) examined the NWR and sentence comprehension skills of 13 children aged 7;0 with SLI and 13 TDAM peers. The children repeated 24 nonwords: 8 each of two, three, and four syllables. The nonwords had a trochaic pattern, and none of the syllables were real words in English. The mean percentage accuracy of NWR across the tasks was 46.36 for the children with SLI and 65.46 for the TDAM group. The children with SLI scored significantly lower than their TDAM peers on three- and four-syllable nonwords. Marton and Schwartz also included tasks of sentence processing (comprehension), having the child repeat the nonsense word as well as answering questions about content. For example, after repeating the nonword at the end of “Bill liked the party given by /kklot/,” the child was asked “What did Bill like?”

The children with SLI were worse not only at producing nonwords but also at recalling the sentence content, leading the authors to suggest a reduced working memory capacity, due to difficulties with simultaneous processing in children with SLI rather than specific problems with phonological data encoding, representation, storage, or retrieval. Recent research found support for a reduced working memory capacity but also shows specific difficulties with phonological storage (Archibald & Gathercole, in press).

Archibald and Gathercole (in press) assessed the working memory skills of 16 children with SLI aged 9;11 (6;9–11;10). Although six children had deficits in central executive functions (described as coordination of multiple tasks, shifting between tasks, selective attention and inhibition, and updating information) without deficits in the phonological loop, the majority of children (10) had deficits in both central executive functions and the phonological loop. In a comparison of these scores with those from a sample of 636 TDAM children, Archibald and Gathercole, who used likelihood ratios, reported that the children with SLI were nine times more likely to have reduced central executive composite scores than the TDAM group and were four times more likely to have low scores on phonological loop functioning.

The interpretation of these results was that the lower performance on tests of working memory in children with SLI are mostly captured by an explanation of poor storage and processing of phonological information. This deficit in turn is reflected in low vocabulary scores and impoverished morpho-syntax in children with SLI (e.g., Baddeley, Gathercole, & Papagno, 1998). This apparent relationship between reduced phonological working memory capacity and language ability led researchers to suggest that tasks such as NWR may serve as a method of identifying children with language impairments. The first of such studies was conducted by Dollaghan and Campbell (1998).

Dollaghan and Campbell (1998) reported that 20 children aged 7;10 with SLI scored significantly lower than 20 TDAM peers on three- and four-syllable nonwords. A second study evaluated the classificatory power of NWR for a group of 44 children with SLI and 41 TDAM children. The test classified the children as TDAM or SLI almost perfectly (98% accuracy). Of particular interest in this study is the finding of low within-group variability and the total absence of an overlap of scores for the SLI and TDAM groups. This lack of overlap is particularly important in tools that one would want to label clinical markers.

In a study of younger children, Conti-Ramsden (2003) examined the classificatory power of NWR, digit span, past-tense markers, and noun plurals in 32 five-year-old children with SLI and their TDAM peers. NWR had a sensitivity of 66% and a specificity of 100% in distinguishing between the two groups of children (at the 25th percentile cut point); the past-tense marker had 71% sensitivity and 91% specificity. In a companion article, Conti-Ramsden and Hesketh (2003) compared the performance of 5-year-old children with SLI with that of TDLM peers (aged 3;0), on the same four possible clinical markers: (a) a past-tense task, (b) a noun plural task, (c) an NWR task, and (d) digit span. The children with SLI performed significantly below the TDLM children on digit span and NWR, but not on past tense or plural tasks. The sensitivity and specificity values were 53% and 90% for digit span and 66% and 85% for NWR, both at the 25th percentile cut point.
Ellis Weismer et al. (2000) assessed the utility of Dollaghan and Campbell’s (1998) test of NWR as an identifier of SLI in a much larger sample. Eighty children with SLI were compared with 359 TDAM children, 84 age-matched children with language impairment plus low nonverbal cognition, and 58 age-matched children with low nonverbal cognition but normal range language performance. Both the SLI group and the group with language impairment/low nonverbal cognition performed significantly worse than the TDAM group, and the group with language impairment/low nonverbal cognition performed worse than the group with low nonverbal cognition/normal range language.

Although NWR does discriminate between children with SLI and their TDAM peers, it is not known what underpins the difference. It is claimed that working memory deficits underlie the poor performance of children with SLI, but there are at least two other factors to consider: (a) the lexicality effects of NWR stimuli and (b) the phonological structure of the language under investigation.

A test of NWR should be free of lexicality effects (Gathercole, Frankish, Pickering, & Peaker, 1999). This means that the nonword stimuli should be constructed of syllables that do not occur as real words in the ambient language and that CV combinations should not be highly predictable. In addition, the segments should be early developing, to avoid articulatory constraints, and the syllables should be perceptually salient (Dollaghan & Campbell, 1998). Lexicality and sublexical effects (e.g., high phonotactic probability of onsets) can result in the use of long-term language knowledge to facilitate performance on nonword repetition. This occurs through the cognitive process of redintegration, whereby high-probability structures, such as onset biphone frequencies (CV), are used to reconstruct nonword traces in the retrieval process for recall (Gathercole, 1999).

The process of redintegration would provide an advantage for more able language users over less able users (e.g., children with SLI), as nonword stimuli can be reconstructed if the biphones are of sufficiently high frequency in the ambient language (Gathercole et al., 1999). Thus, even though the stimuli are nonwords, if the biphones have a high distributional probability in the ambient language, then it may be possible for able language users to use redintegration to re-create the nonword in output. This would mean that groups of children that varied by high versus low language ability should respond differently to high- and low-probability nonwords. The low-ability group should be worse than the high-ability group on high-probability words but should be equivalent to the high-ability group on low-probability words, for which neither group is able to use redintegration for reconstruction. In a task of serial recall (recalling lists of nonwords, not repeating individual nonwords, as is typical of NWR tasks), Gathercole et al. (1999) found that high-probability nonwords were recalled more accurately than low-probability nonwords; however, both groups showed a similar performance with regard to probability, performing better on high-probability than low-probability words. The authors claimed that therefore the differences between the groups were due to the differences in phonological loop functioning and not to redintegration as a function of language ability. It remains unclear how the sublexical nature of NWR stimuli affects NWR rather than serial recall.

To our knowledge, NWR in children with SLI has been explored in only two languages other than English: (a) Spanish (Calderon & Guiterrez-Clellen, 2003) and (b) Swedish (Hansson, Forsberg, Lofqvist, Maki-Torkko, & Sahlen, 2004; Sahlen, Reinerskold-Wagner, Nettelbladt & Radeborg, 1999). Calderon and Guiterrez-Clellen (2003) compared NWR performance of 21 children with SLI with that of 21 TDAM children (aged 4;11). All of the children were either Spanish monolingual or Spanish–English bilingual speakers. The results were similar to those of English; the SLI group performed significantly worse than the TDAM group, and performance decreased for both groups as syllable length increased. The results are not addressed here further, as some of the children were bilingual.

The methods of the Swedish studies render them difficult to compare directly with findings for English; however, one very important finding has emerged. Sahlen et al. (1999) found that children with SLI were six times more likely to omit unstressed syllables in weak–strong syllable combinations than strong–weak syllable combinations on an NWR task. In addition, a small pilot study of English-speaking children with SLI (N = 4, age range: 12–14 years) revealed possible prosodic influences on poor NWR (Marshall, Ebbels, Harris, & van der Lely, 2002). This begs the question of whether difficulties with complex prosodic structures underpin problems with NWR in English and Swedish.

We sought to explore the issues of prosodic structure and lexicality effects by investigation of NWR in Cantonese, where prosodic structure is not an issue (see the Method section). Whereas English and Swedish have complex phonotactic structures, variable stress patterns, and difficult-to-articulate consonants, Cantonese has a small phonetic inventory (16 singleton onsets and 6 singleton offsets), a simple phonotactic structure (CV(C)), and invariable stress. Two-, three-, and four-syllable words have equal stress on each syllable, and there is no stress reduction. Although Cantonese is a tonal language, children use tone correctly in words from very early in development, as young as 2 years (So & Dodd, 1995). Tone errors are also rare in a test of receptive vocabulary (Lee,
Lee, & Cheung, 1996), even though adult-like ability to discriminate among tones at the single-syllable level does not develop until around 10 years of age (Ciocca & Lui, 2003). To avoid introducing a further complicating variable, all of the stimuli were controlled for tone (see Method section). Furthermore, the nature of the Cantonese syllabary means that it is possible to construct NWR stimuli that are devoid of high-probability features (see Method section).

SR

Using likelihood ratios, Ellis Weismer et al. (2000) suggested that although NWR can be used to help identify children with SLI, the test should be used in conjunction with other measures for identification purposes. Such other measures have been the use of grammatical morphemes and additional tests of psycholinguistic processing (SR/immediate imitation). Other studies use the term recalling sentences, which is a subtest within the Clinical Evaluation of Language Fundamentals—Revised (CELF–R), or the Test of Language Development—Primary (TOLD–P; Semel, Wiig, & Secord, 1989, and Newcomer & Hammill, 1997, respectively), which is in fact an SR/imitation task. Still other studies use the term sentence recall, tapping the ability to recall the sentences after a delay, or recalling parts of the sentence, often including judgment regarding the veracity of the sentence. The latter task differs from immediate SR of the type used here. In reviewing the research that has used any type of sentence imitation/repetition task, we refer to the task as sentence repetition to avoid confusion.

SR tasks have long been part of a screening battery for the identification of language impairment (e.g., the CELF–R [Semel et al., 1989] and the TOLD–P [Newcomer & Hammill, 1997]) or general abilities (Wechsler Preschool and Primary Scale of Intelligence—Revised; Wechsler, 1989) and have been used to explore the processing abilities of children with SLI and other conditions and in languages other than English (e.g., Cantonese: Stokes & Fletcher, 2003; Italian: Vicari, Caselli, Gagliardi, Tonucci, & Volterra, 2002; Volterra, Caselli, Capirci, Tonucci, & Vicari, 2003; Dutch: Rispens, 2004). To avoid confusion.

Botting and Conti-Ramsden (2003) suggested that although NWR can be used to help identify children with SLI, the test should be used in conjunction with other measures for identification purposes. Such other measures have been the use of grammatical morphemes and additional tests of psycholinguistic processing (SR/immediate imitation). Other studies use the term recalling sentences, which is a subtest within the Clinical Evaluation of Language Fundamentals—Revised (CELF–R), or the Test of Language Development—Primary (TOLD–P; Semel, Wiig, & Secord, 1989, and Newcomer & Hammill, 1997, respectively), which is in fact an SR/imitation task. Still other studies use the term sentence recall, tapping the ability to recall the sentences after a delay, or recalling parts of the sentence, often including judgment regarding the veracity of the sentence. The latter task differs from immediate SR of the type used here. In reviewing the research that has used any type of sentence imitation/repetition task, we refer to the task as sentence repetition to avoid confusion.

SR tasks have long been part of a screening battery for the identification of language impairment (e.g., the CELF–R [Semel et al., 1989] and the TOLD–P [Newcomer & Hammill, 1997]) or general abilities (Wechsler Preschool and Primary Scale of Intelligence—Revised; Wechsler, 1989) and have been used to explore the processing abilities of children with SLI and other conditions and in languages other than English (e.g., Cantonese: Stokes & Fletcher, 2003; Italian: Vicari, Caselli, Gagliardi, Tonucci, & Volterra, 2002; Volterra, Caselli, Capirci, Tonucci, & Vicari, 2003; Dutch: Rispens, 2004). However, few studies have explored the use of SR as a clinical marker.

Conti-Ramsden et al. (2001) compared the performance of 160 eleven-year-old children with SLI with that of 100 age-matched children (aged 10;9) on four potential clinical markers: (a) NWR, (b) tense marking, (c) a third-person singular task, and (d) an SR task. SR was the best identifier of SLI, followed by NWR, then past tense, and, last, third-person singular. In this study, the sensitivity and specificity values for SR were 90% and 85% at the 16th percentile cut point; those for NWR were 78% and 87%. Concurrent research found significant differences between 20 children with SLI (7;2–13;0) and both TDAM and TDLM groups on tests of both NWR and SR (Briscoe, Bishop, & Norbury, 2001). Additionally, Conti-Ramsden et al. (2001) reported that a group of 11-year-old children whose language impairments had resolved (they did not score below the 16th centile on language tests) were identified with at least 50% accuracy by NWR and SR as having had a history of language impairment.

Botting and Conti-Ramsden (2003), in a comparison of language and processing skills in four groups of children with impaired language (autism spectrum disorder, SLI, and two groups of children with primary pragmatic language impairment), found that SR was better than NWR and a past-tense task at discriminating children with SLI from both other impaired groups and TDAM children. The authors suggested that the source of the difficulties with SR could be either general working memory deficits or a “deficit in linguistic processing” (p. 523), suggestions also made by Eadie, Fey, Douglas, and Parsons (2002).

The underlying mechanism for deficits in SR may be a language impairment, reduced working memory, or both. As Conti-Ramsden et al. (2001) pointed out, SR must tap a child's language knowledge, whereas NWR does not (providing the NWR test is constructed to avoid wordlike nonwords). Other authors have reported the role of language knowledge in facilitating sentence recall/repetition (e.g., MacWhinney, Feldman, Sacco, & Valdes-Perez, 2000; Willis & Gathercole, 2001). One would expect SR to have a better identification status than NWR, because by definition children with SLI have poorer language skills than their TDAM peers. However, at present it seems that SR tasks do not distinguish between children with SLI and children with other types of language impairments with greater than moderate success (Botting & Conti-Ramsden, 2003). In addition, a recent study showed that both children with SLI and children with attention deficit/hyperactivity disorder performed significantly poorer on an SR task (a subtest of the TOLD–P;3; Newcomer & Hammill, 1997) than their TDAM peers (Redmond, 2005).

Redmond (2005) suggested that finer scoring procedures be used in SR tasks if the task is to be used as part of an inclusion battery for the detection of SLI. Therefore, in this study, we compared four different scoring methods for the SR task (see below). Redmond also noted that the SR stimuli must be sufficiently complex (but not long), to avoid ceiling scores in the TDAM group. For this reason, we chose Cantonese sentences that were of an appropriate length but that included structures that would avoid ceiling performance. These sentences included passive constructions (as did Redmond) and aspect markers (which are not obligatory in the language and are later to develop than some other forms; see Stokes & Fletcher, 2003).
It is clear that research that directly compares psycholinguistic and linguistic markers as identifier variables favors psycholinguistic markers (NWR, digit span, and SR). These results are encouraging for cross-linguistic research, as specific linguistic markers (e.g., past tense and third person singular) do not occur in all languages. A clinical marker ideally should have high sensitivity and specificity (over 80%; Plante & Vance, 1994) and a high positive likelihood ratio (LR). However, although the LRs reported by Dollaghan and Campbell (1998), Archibald and Gathercole (in press), and Ellis Weismer et al. (2000) are encouraging, the sensitivity and specificity values reported by Conti-Ramsden and colleagues are disappointing (Conti-Ramsden, 2003; Conti-Ramsden et al., 2001; Conti-Ramsden & Hesketh, 2003). It is clear that LRs and sensitivity/specificity measures are more robust for older children. As the age of participants falls, so too does the classificatory power of NWR/SR, no doubt because of the difficulty of firmly identifying language impairment in very young children (Leonard, 1998). A similar pattern of decreasing classificatory power with decreasing age has been reported for mean length of utterance (MLU) and lexical diversity (Klee, Stokes, Wong, Fletcher, & Gavin, 2004; Klee, Gavin, & Stokes, in press).

If children with SLI were shown, across languages, to have deficits in NWR and/or SR, then we would be one step closer to uncovering the cognitive mechanisms that contribute to language impairments. If Cantonese-speaking children with SLI have no difficulty with tests of NWR and SR, then it is possible that the problems identified in Swedish and English may have more to do with the structure of the language than with a processing deficit. Alternatively, differences between TDAM and SLI groups may be due to differences in language ability rather than working memory. It is also possible that both processing and language limitations contribute to a decreased performance on these tasks.

**Method**

**Participants**

Cantonese monolingual children ($N = 44$) participated in this study. Their characteristics have been reported previously (Wong, Leonard, Fletcher, & Stokes, 2004; Fletcher, Leonard, Stokes, & Wong, 2005). Fourteen of the children, aged 4:2 to 5:7 ($M = 4:11, SD = 5.6$ months) had been previously diagnosed as having deficits in language ability by a speech-language pathologist in the community, and they met the criteria for SLI. They scored at more than $1.20 SD$ below the mean on the Receptive subtest of the Cantonese version of the Reynell Developmental Language Scales (CRDLS; Reynell & Huntley, 1987). Except for 1 child who scored 83, all of the children scored at least 84, that is, no less than $1 SD$ below the mean ($M = 97.43, SD = 11.80$), on a test of nonverbal intelligence, the CMMS (Columbia Mental Maturity Scales; Burgemeister, Blum, & Lorge, 1972). The mean MLU for the group was $3.70 (SD = 0.70)$. All of the children passed tests of oral-motor function, articulation, and hearing, and had no reported neurological or psychosocial dysfunction.

The first control group was composed of 15 TDAM children, aged between 4:1 and 6:9 ($M = 5:0, SD = 7.7$ months), who scored no lower than $0.67 SD$ below the mean for their ages on the receptive subtests of the CRDLS ($M = 56.33, SD = 2.70$) and no lower than $1.00 SD$ below the mean on the CMMS ($M = 110.67, SD = 7.72$). The mean MLU value for this group was $4.49 (SD = 0.77)$; this was significantly higher than that of the SLI group, $t(27) = 2.89, p < .01$.

The second control group was composed of 15 TDY children aged from 2:11 to 3:6 ($M = 3:3, SD = 4.7$ months). They were significantly younger than the children with SLI, $t(27) = 10.75, p < .001$, but they had very similar MLU values ($M = 3.97, SD = 0.73$). These children scored no lower than $0.67 SD$ below the mean for their age on the CRDLS and no lower than $1 SD$ below the mean on the CMMS (or, for the children under 3:0, the Leiter International Performance Scale [Leiter, 1979]; $M = 112.47, SD = 15.72$). The raw scores for this group on the CRDLS ($M = 41.73, SD = 6.10$) were not significantly different from those of the SLI group ($M = 42.64, SD = 5.79$), $t(27) = 0.41, p = .68$. All typically developing children passed tests of oral-motor function and hearing and had no reported neurological or psychosocial dysfunction.

The only matching criterion for the TDAM group with the SLI group was age, and that for the TDY group was receptive grammar scores. There was no attempt to match the groups on general intelligence, although a selection criterion was that the children with SLI all scored within normal limits. Consequently, a univariate analysis of variance (ANOVA) for Group × CMMS scores revealed a significant effect of group, $F(2, 41) = 6.49, p < .01, \eta^2 = 0.24$. Post hoc analyses showed that both typically developing groups had significantly higher CMMS scores than the SLI group ($Ms = 111, 112, and 97; SDs = 7.72, 15.72,$ and $11.80$, for the TDAM, TDY, and SLI groups, respectively). The confidence interval (CI) for the TDAM–SLI comparison was $4.08–22.40$, and for the TDY–SLI comparison it was $5.88–24.20$. The range of CMMS scores for the TDAM group was $98–126$, for the TDY group it was $89–142$, and for the SLI group it was $83–114$. The TDAM group scored significantly higher than both the TDY and SLI groups on all language measures (see Table 1). There was no significant
Table 1. Means and between-group differences on the language measures.

<table>
<thead>
<tr>
<th>Language</th>
<th>TDAM</th>
<th>TDY</th>
<th>SLI</th>
<th>F(2, 41)</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRDLS–R</td>
<td>56.53</td>
<td>41.73</td>
<td>42.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRVT</td>
<td>60.13</td>
<td>47.07</td>
<td>50.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLU</td>
<td>4.49</td>
<td>3.97</td>
<td>3.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. MANOVA = multivariate analysis of variance; TDAM = typically developing age matched; TDY = typically developing younger; SLI = specific language impairment; CRDLS–R = Receptive scale of the Cantonese version of the Reynell Developmental Language Scales (Reynell & Hunley, 1987); CRVT = Hong Kong Cantonese Receptive Vocabulary Test (Lee et al., 1996); MLU = mean length of utterance.

Twenty-four nonwords, 6 for each of the four syllable lengths, were constructed for this study (see Appendix A). The syllable structure for the one-, two-, three-, and four-syllable nonwords was CVC, CVCVC, CVCVCVC, and CVCVCVCVC, respectively. The Cantonese syllabary (Bauer & Benedict, 1997) was used to select syllables that do not occur in the language. This is equivalent to selecting a syllable like /pom/ in English—it is theoretically possible, given the phonotactic constraints of English, but it does not occur in the language. The Cantonese syllabary is a syllable matrix of all of the combinations of 20 initials and 56 rimes, plus the two nasal syllables, that are theoretically possible in the language. The syllabary contains a total of 1,122 possible syllables; of these combinations, 372 are nonoccurring. Native Cantonese speakers do not recognize these latter combinations as words they know. The nonwords did not contain late-developing consonants (/kʷ/, /kʰ/, /tʰ/, and /tsʰ/). To investigate sublexical effects, the component CV syllables in the nonwords were further selected on the basis of whether the CV combinations did or did not occur in the language. This is equivalent to selecting English CV combinations that do occur in the language (like /put/, as in put) and those that do not occur in the language, such as /zut/. This selection process resulted in CV structures for Cantonese that were either nonoccurring (N = 34) or occurring (N = 26) in the Cantonese syllabary (Bauer & Benedict, 1997). CV structures that appear in the Cantonese syllabary are described as IN syllables; CV structures that do not appear in the syllabary are labeled OUT syllables.

Cantonese words tend to be monosyllabic (Bauer & Benedict, 1997), but it is also common to find words that are composed of two or more monosyllabic syllables. These polysyllabic words are word compounds (fan3gaaau3 = sleep); English loan words (mako6dong2noo4 = McDonald’s); or combinations such as gaa6zaat (cockroach), in which each individual syllable has no meaning and stative verb phrases such as fei1tat1tat1 (“adorably fat”; Bauer & Benedict, 1997) that function as one unit syntactically. (Throughout this article, Cantonese syllables are written in romanized form. Numerals provided with the Cantonese syllables represent tone values, following the system adopted by the Linguistic Society of Hong Kong, 1974.) However, polysyllabic words containing more than three syllables are uncommon. Unlike stress-timed languages, such as English, syllables in a Cantonese polysyllabic word carry relatively equal stress, with only a few exceptions. Although tones are sometimes neutralized in the unstressed syllables of a polysyllabic word in Putonghua (standard Chinese), neutralization of tones is uncommon in Cantonese.

The same tone or tonal pattern was assigned for nonwords of the same syllable length. All one-syllable nonwords carried a low-level tone (e.g., weng6), two-syllable words took a low falling + high-level tone (e.g., wo5kaa1), three-syllable words took tones of mid level + high level + low level (e.g., mo3woe1paam6), and four-syllable words had the structure of high level + high rising + low falling + low level (e.g., keu1wan2hu4fen6). Tonal patterns for multisyllabic nonwords were modeled after real Cantonese words of the same syllable length and hence would sound familiar to Cantonese speakers.

To ensure that the nonwords were neither too close to real words nor too foreign to be similar to Cantonese real words, four adult native speakers of Cantonese were asked to rate their wordlikeness on a scale from 1 (very much like a Cantonese word) to 5 (very much unlike a Cantonese word). Two 1-syllable nonwords, receiving average ratings below 2, were replaced, and the procedure was repeated. The final list of nonwords used in this study appears in Appendix A. A female native speaker of Cantonese recorded the nonwords onto a minidisk.

Each child scored one point for each correct consonant and nucleus. We use the term nucleus because vowels and diphthongs were scored as one unit, so /bei/ → /bi/ would be scored as 1 (nucleus incorrect). The maximum score for one-syllable words was 17 (1 word was mistakenly of CV structure), for two-syllable words it
was 30, for three-syllable words it was 42, and for four-syllable words it was 54 (total possible score = 143).

**SR Stimuli**

Two sentence types were used for this task (see Appendix B). All sentences were either 9 or 10 syllables in length. The first set of eight sentences contained the elements of subject, verb, aspect marker, and object, within a sentence frame. An example is:

Go3 naam4zai2 hou2 daai6lik6 caat3 gan2 deoi3 haai4

(subject cl noun advi adj vt asp cl nn
cl boy very forceful rub asp cl shoe
(The boy is polishing the shoe very hard.)

The second set of eight sentences contained the passive elements bei2 + agentNP + V; an example is:

Ceong4 tau4faat3 go3 naam4zai2 bei2 zek3 gau2 zeoi1

(Adi nn cl nn prep cl nn vt
Long hair cl boy prep cl dog chase
(The boy with long hair was chased by the dog.)

There were four methods of scoring for this task. In the first (hereafter complete sentence correct), the child scored one point for each sentence repeated exactly, with the inclusion of all sentence elements, for a total score of eight for each of the aspect marker and passive tasks. This scoring procedure is the most common in use and mirrors that used for the TOLD–P3 Sentence Imitation subtest (Newcomer & Hammill, 1997) and a test of sentence imitation in Dutch-speaking children (Rispens, 2004). However, this procedure may place the TDY children at a disadvantage in that they may not have the necessary sentence-length skills to complete the task. For this reason, a second scoring procedure was used. In this second procedure (hereafter core elements correct), one point per sentence was awarded for any repetition that included all of the core elements of the two stimulus types, the aspect marker sentences and the passive sentences, where the core elements are defined as the verb + aspect marker + noun for the former, and bei + agent + verb for the latter, for a maximum score of eight for each sentence type. This results in exact repetition of three to four syllables, with these syllables marked by italics in the first example above. The third scoring method was used to equate our results with those for English (Conti-Ramsden et al., 2001). This method (hereafter the error method) was the calculation from the CELF–R (Semel et al., 1989), where a score of three points is given for complete sentence accuracy, two points for one error, one point for two or three errors, and zero points for four or more errors. The last method was the calculation of the percentage of syllables correct, where one point is given for each correct syllable and additions and transpositions are ignored (hereafter percentage syllables correct).

**Procedure**

The NWR and SR tasks were randomized, so that some children received the NWR task first and some received the SR task first. The NWR task was administered via free-field speakers. The child was told that he or she would hear some “funny made up words” and asked to listen carefully and repeat them exactly as they were heard. The trial items were presented first (two words and two nonwords). A trial was repeated once if the child’s response was incorrect. No feedback was given on experimental items, but encouragement was given as required. Each experimental item was presented once only. The nonwords were presented in order of increasing difficulty (all one-syllable nonwords, followed by two-syllable nonwords, etc.). All responses were recorded verbatim and audiorecorded for later transcription.

The SR tasks also were administered via free-field speakers. The three trial items were administered first. The child was asked to listen carefully and repeat the sentence exactly as heard. One repeat was permitted for the trial items if the child’s response was incorrect. No test items were repeated. All responses were recorded verbatim and audiorecorded for later transcription.

**Transcription Reliability**

Interrater reliability of the transcription of NWR and SR was conducted on 16% of the data. There were six original native Cantonese-speaking trained transcribers: four in the first pass of transcriptions and two validators. The percentage of interrater agreement on transcription of NWR segments was 84%, 76%, 77%, and 77% for one-, two-, three-, and four-syllable nonwords, respectively, for an average agreement of 79%. Disagreements were on the status of final plosives, which are unreleased in Cantonese, and diphthongs, but exact counts of each disagreement type were not kept. The reliability of SR transcription was 93% for aspect sentences and 95% for passive sentences.

**Results**

**NWR for the Total Test**

The distributions of the data were not significantly different from a normal distribution, for all groups, for all variables (all Kolgomorov–Smirnov p values > .05); therefore, we used parametric statistics to explore differences among the groups. Means and standard deviations were derived for one-, two-, three-, and four-syllable nonwords and average accuracy across the four word lengths. The means (and standard...
deviations) for the TDAM children, the TDY children, and the children with SLI are shown in Table 2.

A one-way ANOVA revealed significant differences among the groups on total percentage correct for all syllable lengths combined (the total number of phonemes correct divided by the total number of phonemes × 100), F(2, 43) = 8.22, p < .001, η² = .29. Because the test of homogeneity of variances was not significant, Levene statistic (2, 41) = 0.91, p = .41, we used the least significant difference (LSD) method for post hoc comparisons. The TDAM group and the SLI group scored significantly higher than the TDY group (68.35% and 62.99% accuracy vs. 55.76% accuracy, p < .001 and p < .03, respectively). There was no significant difference between the TDAM and the SLI groups. This finding shows that the test of NWR was age sensitive but did not distinguish between the children with the SLI and their age-matched peers.

A 3 (group) × 4 (length) repeated measures ANOVA revealed a significant main effect of group, F(2, 41) = 8.55, p < .002, partial η² = .29, and length, F(3, 39) = 34.99, p < .001, partial η² = .73. Effect sizes greater than .8 are regarded as high, and those between .50 and .79 are regarded as moderate (Cohen, 1988). The Group × Length interaction was not significant. A post hoc analysis (LSD at the .05 level) showed that there was a decreasing order of accuracy of production of nonwords from one-syllable to four-syllable structures. One- and two-syllable words were of equal difficulty, and both were significantly easier than three- and four-syllable words. Three-syllable words were significantly easier than four-syllable words (p < .01 for all comparisons).

On post hoc LSD comparisons, the TDAM group scored significantly higher than the TDY group on two- (76%–66%), three- (69%–57%), and four-syllable (62%–45%) nonwords. The children with SLI scored significantly higher than the TDY children on one-syllable and two-syllable nonwords (77%–70% and 76%–66%; p < .01 for all comparisons). There were no significant differences between the TDAM children and the children with SLI. The mean differences, significance values, and 95% CIs for these comparisons are shown in Table 3. Although the children with SLI had higher scores than the TDY children on three- and four-syllable words, the difference was not significant.

**Table 2.** Means (in percentages) for all syllable lengths on the nonword test for the three participant groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>One syllable</th>
<th>Two syllables</th>
<th>Three syllables</th>
<th>Four syllables</th>
<th>Total % correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDAM</td>
<td>74.51</td>
<td>76.22</td>
<td>68.73</td>
<td>61.73</td>
<td>68.35</td>
</tr>
<tr>
<td></td>
<td>(6.55)</td>
<td>(8.054)</td>
<td>(7.07)</td>
<td>(11.20)</td>
<td>(6.40)</td>
</tr>
<tr>
<td>TDY</td>
<td>69.80</td>
<td>65.77</td>
<td>56.98</td>
<td>44.81</td>
<td>55.76</td>
</tr>
<tr>
<td></td>
<td>(9.13)</td>
<td>(10.27)</td>
<td>(14.01)</td>
<td>(15.43)</td>
<td>(9.22)</td>
</tr>
<tr>
<td>SLI</td>
<td>77.31</td>
<td>75.95</td>
<td>61.56</td>
<td>52.38</td>
<td>62.99</td>
</tr>
<tr>
<td></td>
<td>(8.58)</td>
<td>(9.16)</td>
<td>(16.22)</td>
<td>(13.44)</td>
<td>(9.69)</td>
</tr>
<tr>
<td>Total</td>
<td>73.79</td>
<td>72.57</td>
<td>62.44</td>
<td>52.98</td>
<td>62.35</td>
</tr>
<tr>
<td></td>
<td>(8.55)</td>
<td>(10.26)</td>
<td>(13.57)</td>
<td>(14.94)</td>
<td>(9.86)</td>
</tr>
</tbody>
</table>

Note. Numbers in parentheses are standard deviations.

**Table 3.** Means, significance levels, and confidence intervals (CIs) for between-group comparisons of performance on nonword repetition.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Groups</th>
<th>Mean difference in scores</th>
<th>p</th>
<th>95% CI for differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-syllable word</td>
<td>SU &gt; TDY</td>
<td>7.50</td>
<td>.02</td>
<td>1.38–13.62</td>
</tr>
<tr>
<td>Two-syllable word</td>
<td>TDAM &gt; TDY</td>
<td>10.44</td>
<td>.00</td>
<td>3.65–17.23</td>
</tr>
<tr>
<td>Three-syllable word</td>
<td>SU &gt; TDY</td>
<td>10.17</td>
<td>.00</td>
<td>17.08–6.26</td>
</tr>
<tr>
<td>Four-syllable word</td>
<td>TDAM &gt; TDY</td>
<td>11.75</td>
<td>.02</td>
<td>2.19–21.29</td>
</tr>
</tbody>
</table>

Note. Only significant results are reported.

**NWR for IN and OUT words**

CV structures that appear in the Cantonese syllable are described as IN syllables, and CV structures that do not appear in the syllable are labeled OUT syllables. A 3 (group) × 2 (syllable) repeated measures
ANOVA revealed a significant main effect of group, $F(2, 41) = 8.66, p < .002$, partial $\eta^2 = .29$, and syllable, $F(1, 41) = 132.21, p < .001$, partial $\eta^2 = .76$. The Group $\times$ Syllable interaction was not significant. IN syllable nonwords were easier to repeat than OUT syllable nonwords. On multivariate comparisons, the TDAM group scored significantly higher than the TDY group on IN syllables (difference = 17.75, $p < .01$, CI: 8.88–26.62) and OUT syllables (difference = 10.98, $p < .02$, CI: 2.23–19.73).

The children with SLI scored significantly higher than the TDY children on IN syllables (difference = 10.75, $p < .03$, CI: 1.73–19.78) and OUT syllables (difference = 10.56, $p < .03$, CI: 1.66–19.46). Although the TDAM group scored higher than the SLI group on IN syllables, the result was not statistically significant. Follow-up multivariate analyses revealed that although there were group differences on total IN and OUT syllables, there was no significant difference among the groups on individual syllable lengths for IN nonwords (i.e., the differences in average IN syllable scores could be attributable not to, say, three-syllable IN nonwords but rather to overall performance on IN syllables; see Table 4). On the other hand, the differences in OUT syllables were attributable to the differences in performance on two-syllable OUT words, $F(2, 41) = 5.23, p < .01, \eta^2 = .20$.

In summary, the TDAM and the SLI group scored significantly higher than the TDY group on NWR and on repetition of IN nonwords and repetition of OUT nonwords. There were no significant differences between the TDAM and the SLI groups.

**Table 4.** Means and standard deviations (in parentheses) for percentage of accuracy of repetition of IN and OUT CV structures by word length for the three participant groups.

<table>
<thead>
<tr>
<th>CV structure</th>
<th>TDAM</th>
<th>TDY</th>
<th>SLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-syllable IN</td>
<td>88.89</td>
<td>66.67</td>
<td>83.33</td>
</tr>
<tr>
<td>(16.27)</td>
<td>(33.33)</td>
<td>(28.49)</td>
<td></td>
</tr>
<tr>
<td>One-syllable OUT</td>
<td>51.11</td>
<td>48.88</td>
<td>57.14</td>
</tr>
<tr>
<td>(24.77)</td>
<td>(21.33)</td>
<td>(24.20)</td>
<td></td>
</tr>
<tr>
<td>Two-syllable IN</td>
<td>78.89</td>
<td>58.89</td>
<td>71.43</td>
</tr>
<tr>
<td>(19.38)</td>
<td>(27.36)</td>
<td>(23.04)</td>
<td></td>
</tr>
<tr>
<td>Two-syllable OUT</td>
<td>55.56</td>
<td>37.78</td>
<td>59.52</td>
</tr>
<tr>
<td>(19.59)</td>
<td>(16.02)</td>
<td>(22.37)</td>
<td></td>
</tr>
<tr>
<td>Three-syllable IN</td>
<td>62.86</td>
<td>51.43</td>
<td>54.08</td>
</tr>
<tr>
<td>(21.46)</td>
<td>(19.32)</td>
<td>(17.88)</td>
<td></td>
</tr>
<tr>
<td>Three-syllable OUT</td>
<td>48.47</td>
<td>34.55</td>
<td>41.56</td>
</tr>
<tr>
<td>(21.07)</td>
<td>(21.79)</td>
<td>(22.18)</td>
<td></td>
</tr>
<tr>
<td>Four-syllable IN</td>
<td>63.33</td>
<td>46.00</td>
<td>57.14</td>
</tr>
<tr>
<td>(14.96)</td>
<td>(20.63)</td>
<td>(21.99)</td>
<td></td>
</tr>
<tr>
<td>Four-syllable OUT</td>
<td>26.19</td>
<td>16.19</td>
<td>21.42</td>
</tr>
<tr>
<td>(15.89)</td>
<td>(13.07)</td>
<td>(16.34)</td>
<td></td>
</tr>
</tbody>
</table>

**SR**

Performance on SR was assessed in four ways: (a) one point if the entire sentence was correct ("complete sentence correct"), (b) one point if the core elements were correct, regardless of omission/error of the other elements ("core elements correct"), (c) the error method (three points given for complete sentence accuracy, two points for one error, one point for two or three errors, and zero points for four or more errors), and (d) the percentage of syllables correct. Means and standard deviations for SR are shown in Table 5. Note that the error score procedure is essentially reverse scored, so that a higher score means fewer errors.

We ran complete statistical analyses for each scoring method, and three of the four methods yielded very similar results. Therefore, we present the first analysis in full here, and the results of subsequent analyses appear in Table 5. The first question was "Is there a significant difference among the groups in repetition of complete sentences?" Because two sentence types were used, we compared the children's performance on these two types (aspect markers and passives), using a multivariate analysis of variance of the form Group (3) $\times$ SR (2). As might be suspected from the means, on complete aspect SR there was a significant effect of group, $F(2, 41) = 9.08, p < .01, \eta^2 = .31$ (Ms = 43.33, 16.67, and 15.18 for TDAM, TDY, and SLI, respectively). The error variances across the three groups were not assumed to be equal, and Dunnet's C test was used for post hoc comparisons. The TDAM group scored significantly higher than both the SLI group ($p < .01$, CI: 12.97–43.34) and the TDY group ($p < .01$, CI: 11.75–14.59). There was no difference between the SLI and TDY groups. There were no interactions.

On complete passive SR, there was a significant effect of group, $F(2, 41) = 6.68, p < .05$, $\eta^2 = .25$. Dunnett's C test was used for post hoc comparisons. The TDAM group scored significantly higher than both the SLI group ($p < .01$, CI: 10.47–41.91) and the TDY group ($p < .01$, CI: 6.72–37.62). There was no difference between the SLI and TDY groups. There were no interactions.

The second question was "Is there a significant difference among the groups in repetition of core elements?" As with the first method, the TDAM participants outscored both the SLI and TDY groups for the aspect sentences, but there were no significant differences on the passive sentences, and there were no interactions (see Table 5). The third question was "Is there a significant difference among the groups in repetition of sentences using the error method?" As with the first method, the TDAM participants outscored both the SLI and TDY groups on both aspect and passive sentences, and there were no interactions. The fourth question was "Is there a significant difference among the groups in repetition of sentences using the percentage of syllables correct scoring
method (syllables correct method)?” As with the first method, the TDAM participants outscored both the SLI and TDY groups on both aspect and passive sentences, and there were no interactions.

A repeated measures ANOVA, Groups (3) × Scoring Method (4) × Sentence Type (2), showed that the children achieved highest scores on the percentage of syllables correct scoring method, $F(3, 123) = 290.55$, $p < .001$, $\eta^2 = .88$. The overall mean for this method was 82%; for core elements correct it was 70%, for the error method it was 53%, and for complete sentence correct it was 27%. There were no interactions.

The core elements correct method for passive SR resulted in much higher scores for the children than the complete passive SR scoring method. Indeed, the core passive method resulted in 13 of the 44 children achieving ceiling scores of 100% (5 in the TDAM group, 4 in the TDY group, and 4 in the SLI group), whereas no child scored 100% on complete passive sentences. Similarly, 5 children scored 100% on the core aspect method (4 in the TDAM group and 1 in the TDY group), whereas none achieved 100% in the complete aspect sentences.

At the other end of the scale, 9 children scored 0% on the complete passives (5 in the TDY group and 4 in the SLI group), whereas none scored 0% on the core passives. Similarly, 15 children scored 0% accuracy on the complete aspect method (1 in the TDAM group, 7 in the TDY group, and 7 in the SLI group), whereas only 1 child in the SLI group scored 0% on the core aspect method.

Using the error method, 1 TDY child and 1 child with SLI scored zero on both the aspect and passive tests, and no child reached ceiling. Using the percentage of syllables correct method, 2 TDAM children scored at ceiling on the aspect test, and 1 TDY child scored at ceiling on the passive test. No child scored zero, obviously, as this would mean no syllables correct.

In summary, three of the four scoring methods yielded very similar results for both aspect and passive sentences in terms of group differences (TDAM > TDY = SLI). When only the core elements were required, the TDAM group outperformed both groups on the aspect sentences, but there were no differences on passive sentences. The core and complete sentence correct scoring methods yield too many basal and ceiling scores. The error and percentage of syllables correct method are the best scoring methods.

### Identifying SLI

Given the significant differences between the TDAM and the SLI groups on SR, it is worthwhile to explore the use of the test as a clinical marker. Discriminant function analyses were conducted to see whether the SR measures have potential clinical application to distinguish between children developing typically and those with SLI. Given that clinicians, teachers, parents, and so on, take age into account when assessing a developmental ability such as language, only the TDAM children are relevant. Also, because SR scores do not differ between

<table>
<thead>
<tr>
<th>Scoring method</th>
<th>Sentence type</th>
<th>TDAM (15)</th>
<th>TDY (15)</th>
<th>SLI (14)</th>
<th>$F(2, 41)$</th>
<th>$\eta^2$</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete sentence correct</td>
<td>Passives</td>
<td>45.83</td>
<td>23.67</td>
<td>19.64</td>
<td>6.68$^a$</td>
<td>.25</td>
<td>10.47–41.91$^b$</td>
</tr>
<tr>
<td></td>
<td>(21.47)</td>
<td>(23.24)</td>
<td>(17.48)</td>
<td></td>
<td></td>
<td></td>
<td>6.72–37.62$^c$</td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>43.33</td>
<td>16.67</td>
<td>15.17</td>
<td>9.08$^a$</td>
<td>.31</td>
<td>12.97–43.34$^b$</td>
</tr>
<tr>
<td></td>
<td>(22.59)</td>
<td>(19.86)</td>
<td>(17.80)</td>
<td></td>
<td></td>
<td></td>
<td>11.75–41.59$^c$</td>
</tr>
<tr>
<td>Core elements correct</td>
<td>Passive</td>
<td>89.17</td>
<td>70.33</td>
<td>71.43</td>
<td>ns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10.42)</td>
<td>(25.28)</td>
<td>(28.77)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>83.33</td>
<td>61.67</td>
<td>46.43</td>
<td>7.81$^a$</td>
<td>.29</td>
<td>15.30–49.29$^b$</td>
</tr>
<tr>
<td></td>
<td>(14.68)</td>
<td>(24.31)</td>
<td>(26.60)</td>
<td></td>
<td></td>
<td></td>
<td>4.53–37.14$^c$</td>
</tr>
<tr>
<td>Error scoring</td>
<td>Passives</td>
<td>73.61</td>
<td>44.72</td>
<td>44.94</td>
<td>8.85$^a$</td>
<td>.30</td>
<td>10.23–47.12$^b$</td>
</tr>
<tr>
<td></td>
<td>(14.14)</td>
<td>(26.28)</td>
<td>(22.36)</td>
<td></td>
<td></td>
<td></td>
<td>8.72–49.06$^c$</td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>73.33</td>
<td>44.44</td>
<td>38.69</td>
<td>10.06$^a$</td>
<td>.33</td>
<td>14.60–54.68$^b$</td>
</tr>
<tr>
<td></td>
<td>(11.77)</td>
<td>(26.62)</td>
<td>(26.07)</td>
<td></td>
<td></td>
<td></td>
<td>9.23–48.56$^c$</td>
</tr>
<tr>
<td>Percentage of syllables correct</td>
<td>Passives</td>
<td>93.24</td>
<td>74.70</td>
<td>76.81</td>
<td>7.56$^a$</td>
<td>.27</td>
<td>5.74–27.12$^b$</td>
</tr>
<tr>
<td></td>
<td>(3.21)</td>
<td>(17.73)</td>
<td>(17.03)</td>
<td></td>
<td></td>
<td></td>
<td>8.04–29.04$^c$</td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>95.17</td>
<td>76.33</td>
<td>77.86</td>
<td>6.32$^a$</td>
<td>.24</td>
<td>5.26–29.36$^b$</td>
</tr>
<tr>
<td></td>
<td>(2.87)</td>
<td>(21.82)</td>
<td>(17.06)</td>
<td></td>
<td></td>
<td></td>
<td>6.99–30.67$^c$</td>
</tr>
</tbody>
</table>

$^a$A significant difference between children with SLI and their age-matched peers (TDAM) at $p < .01$. $^b$A significant difference between TDAM children and their younger peers (TDY). $^c$A significant difference between children with SLI and their TDY peers.
the children with SLI and their TDY peers, there is no value in attempting to distinguish between these groups on this measure. Two scoring methods were examined.

The first discriminant function analysis explored the use of the error method for aspect and passive sentences (entered together) as a classification function. The overall Wilks's lambda was significant, \( \Lambda = .52, \chi^2(2, N = 29) = 16.93, p < .01 \). The predictors successfully differentiated between the two groups. Ninety percent of the children were correctly classified in the first pass, and 76% were correctly classified using the leave-one-out method (cross-validation). One hundred percent of the TDAM children were correctly classified. Of the children with SLI, 11 were correctly classified as SLI, and 3 were misclassified as TDAM. These were children who scored greater than 67% accuracy on the aspect marker sentences and greater than 55% on the passive sentences. When a second discriminant function was conducted to assess the power of the variables to independently assign group membership, the aspect test misclassified one of the TDAM group as SLI and 5 of the SLI group as TDAM. The passive sentence test misclassified 3 of the TDAM group as SLI and 3 of the SLI group as TDAM. Classification power is higher when the aspect and passive sentences are used together.

The results of the discriminant function analysis are shown in Table 6. Because one of the cells is zero, an adjustment is made by adding .5 to each cell before computing LRs (Walter, 1985). The adjusted sensitivity (true positive/[true positive + false negative]) = 77%. The specificity (true negative/[false positive + true negative]) = 97%. The positive predictive value (true positive/[true positive + false positive]) = 97%. The negative predictive value (true negative/[false negative + true negative]) = 81%. The positive LR (LR+ = sensitivity/[1 – specificity]) = 25.66. The negative LR (LR– = [1 – sensitivity/specificity]) = .24. The LR+ indicates that children with SLI are 25 times more likely to have more difficulty with SR than their TDAM peers (Attia, 2003). Dollaghan (2004) stated that an LR– at or below .20 indicates that the test is very likely to correctly identify individuals who are free of the condition.

Table 6. Classification of children with SLI and their TDAM peers as typically developing or language impaired on a sentence repetition task.

<table>
<thead>
<tr>
<th>Test results</th>
<th>SLI</th>
<th>TDAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>3</td>
<td>15</td>
</tr>
</tbody>
</table>

Note. True positive = 78%, true negative = 100%, false positive = 0%, false negative = 21%.

Scores on the passive and aspect sentences were combined into a total SR score. There was no significant correlation between NWR and SR for any of the groups \((rs = .32, .37, and .25, all at p > .05, for the TDAM, TDY, and SLI groups, respectively)\). Relationships among the processing measures (NWR and SR) and the language measures (receptive grammar CRDLS–R, receptive vocabulary CRVT, and MLU) were explored using Pearson correlations. For the TDAM group, both NWR and SR were not significantly correlated with any of the language measures. For the TDY group, NWR was significantly correlated with receptive vocabulary scores \((r = .65, p < .01)\). SR was correlated with receptive grammar \((r = .90, p < .01)\), receptive vocabulary \((r = .59, p < .05)\), and MLU \((r = .52, p < .05)\). For the SLI group, NWR did not correlate with any language variable. SR correlated with the receptive grammar \((r = .58, p < .05)\).

Scores on the test of nonverbal intelligence were not correlated with NWR or SR scores for both of the typically developing groups. Scores were correlated with SR scores in the SLI group \((r = .79, p < .01)\), but not with the NWR scores.

**Overall Summary**

The SLI group was no different from the TDAM group on NWR. Both groups scored significantly higher than the TDY group. Although the TDAM group scored higher than the SLI group on IN nonwords, the...
difference was not statistically significant. These two groups scored exactly the same on OUT nonwords. On an SR task, the TDAM children scored significantly higher than the SLI and TDY groups, who did not differ. When methods of scoring SR were compared, the best measure was points awarded on the basis of number of errors made. An SR task using an error scoring method successfully distinguished between TDAM and SLI groups and should be further evaluated as a clinical marker of SLI in Cantonese.

Discussion

NWR

Research with English-speaking children has consistently shown a significant difference between children with SLI and their TDAM and TDLM peers on tests of NWR. This poorer performance has been attributed to a language-independent working memory constraint. One might expect, then, that a working memory constraint might be identified in other groups of children with SLI, regardless of the ambient language. However, although NWR is a sensitive developmental test in Cantonese (at least in this cohort), with older children scoring significantly higher than younger children, there is no difference in performance between children with SLI and their TDAM peers. At first glance, this suggests that a working memory capacity limitation of the type believed to underlie NWR skills does not underpin language impairment in Cantonese-speaking children. There are other possible interpretations.

Participant characteristics. First, one might argue that these children do not have language impairment. However, these same children have a poorer performance than both TDAM and TDY children on a host of grammatical makers, including aspect markers (Fletcher et al., 2005), wh-questions (Wong et al., 2004), and passives (Leonard, Wong, Deevy, Stokes, & Fletcher, 2005). The children also meet internationally accepted criteria for classification as language impaired (>1 SD below the mean on language tests with age-appropriate nonverbal intelligence). Given the superior ability of younger TD children on grammatical markers, we are confident that this group of children with SLI is representative of the condition.

Prosodic factors. A second possible culprit is the nature of the language. Both English and Swedish have complex phonotactic structures, variable stress patterns, and difficult-to-articulate consonants. Cantonese has none of these. It is possible that although NWR in Cantonese does differ between TD children between the ages of 5;0 and 3;6, it does not tax working memory in the same way that nonwords do in other languages. Candidates for investigation would be the complex syllable structures and timing patterns of some languages (e.g., Swedish and English). Although researchers are careful to consider these factors in test design (see Santos and Beuno's, 2003, development of a test of Brazilian Portuguese as a good example), it may be that it is these same language factors that prove taxing for children with SLI. This suggestion has been made in small-scale studies of English-speaking and Swedish-speaking children with SLI (Marshall et al., 2002; Sahlen et al., 1999), both of which have implicated prosodic variables. There has been limited work on this topic, but contrary evidence comes from Marton and Schwartz (2003), who reported that stress errors were rare in NWR: Only 4% of all responses from the children with SLI had stress changes, and the TDAM peers had <1% stress errors. Gathercole (1999) outlined possible temporal and sequencing skills required to adequately perform an NWR task. Investigation of these processing skills, in addition to the redintegration strategy outlined below, may prove fruitful in the exploration of processing limitations in SLI.

Redintegration. A third factor, also rooted in language, is sublexical processing. As outlined above, the Cantonese syllabary contains a total of 1,122 possible syllables, constructed from the combinations of 20 initials and 56 rimes. Of these combinations, 372 do not occur in Cantonese words. This means that only 750 syllable structures are possible and therefore syllables that make up all Cantonese words come from a small set of 750 initial and rime combinations that vary by six contrastive tones. We constructed our OUT stimuli by choosing an onset (initial consonant + nucleus vowel) that does not occur, ever, in Cantonese. This is equivalent to choosing a CV combination of /z + u/ in English. It is possible to imagine words beginning with this CV combination in English, given neighbors such as book, look, took, but the onset does not occur in the language (and it is extremely difficult to think of CV combinations that do not appear in English—this is the nature of the language). This may be where the results for English and Cantonese part company. Although Dollaghan and Campbell (1998) stressed that nonword stimuli must not be composed of predictable sequences, their test includes onsets such as /nau/, /tei/, /dæ/, all of which occur in English words. So too do other tests (lists used by Gathercole & Baddeley, 1990; Marton & Schwartz, 2003). Is NWR therefore as language independent as we had thought? Is it more language dependent than processing dependent in English?

It is possible that the structure of English NWR stimuli renders them more similar to English real words than had been previously assumed. Storkel and Young (2004) showed that new words that are of the same phonotactic structure as an existing word (homophones) are easier to learn than words of an unknown structure.
It is possible that sublexical factors, other than complete syllable structure, such as frequency of onsets, may be important. Perhaps in English the better vocabulary skills of TDAM children allow them to use common CV onset patterns as an aid in NWR. The suspicion that onset patterns (CV combinations) may be playing a role in these tasks was alluded to indirectly by Marton and Schwartz (2003), who found that the most common syllable substitution errors were combining the CV of one syllable with the offset of another syllable, yielding the right onset and offset but in the wrong combination. These results suggest that TDAM children may be using the strategy of redintegration.

Redintegration occurs in repetition tasks (and online word production) when an incomplete trace of the stimulus is retrieved, and the long-term memory language store uses lexical and phonotactic information to “fill in the blanks” of the skeletal score (the CVC pattern), creating either an accurate response or a close approximation to the target word. A full discussion of redintegration processes is beyond the scope of this article, but see Gathercole (1999) and Schweickert (1993). It is possible that the English targets used thus far have allowed able language users to use redintegration to achieve an accurate production, a skill that may not be available to children with SLI, resulting in TDAM groups outscoring SLI groups on English NWR tasks.

A test of this premise is to see what happens when CV combinations that are attested (IN) versus those that are unattested (OUT) in the language are imitated. It is not possible to use redintegration with unattested syllables, because there is no existing CV form to retrieve from the long-term store. A successful score on unattested syllables can be achieved only by automatic repetition of the stimulus (true repetition without calling on the long-term store). If the attested syllables are produced at higher success rates by the TDAM group, but the unattested syllables are not, then one can conclude that the greater success of the TDAM group with attested syllables reflects their ability to use redintegration more successfully than the SLI group.

The results presented here for Cantonese support this conclusion. Small sample sizes resulted in non-significant results, but the IN (attested) scores for the TDAM group were higher than those for the SLI group (74% vs. 67%), whereas the OUT (unattested) scores were identical (45%). On the basis of this finding, we suggest that children with SLI do not differ from their TDAM peers on the ability to repeat nonsense syllables when only phonological working memory skills are tested. When the stimuli allow the participant to call on the long-term language store, the TDAM peers do better than the children with SLI. We stress again that the differences in the scores were not statistically significant and encourage the use of larger participant groups to better answer this question.

The poorer performance of English-speaking children with SLI compared with their age-matched peers requires further investigation. The TDAM advantage may be rooted in superior specific processing abilities (e.g., redintegration), or it may reflect a language advantage, rather than some nonspecific advantage in NWR. A problem with this interpretation is that English-speaking children with SLI also score lower than their TDY peers.

Comparisons with younger children. Only three studies have included younger control groups. Conti-Ramsden and Hesketh (2003) matched their groups on scores on the Reynell Developmental Language Scales—III (Edwards et al., 1997). Although this test does measure receptive and expressive language abilities, it may not provide a sufficiently detailed view of receptive and expressive vocabulary skills. This is important, because vocabulary ability has been positively correlated with NWR skills (Gathercole & Baddeley, 1990).

In Briscoe et al.’s (2001) study, the children with SLI formed two groups: The older participants (N = 6) were aged 11;89–13;0 (M = 12;11), and the younger participants (N = 14) were aged 7;20–10;91 (M = 8;96). A group of 15 children aged 4;97–9;17 (M = 7;40) formed the younger, language-matched control group. Although the TDY group was matched to the SLI group to within 3 months on the British Picture Vocabulary Scale ([BPVS] Dunn & Dunn, 1982; raw scores of 71.2 and 68.8), comparisons of the groups across language tasks reveal some interesting findings. Both groups of children with SLI (older and younger) scored significantly lower on the BPVS standard scores than both the TDAM and the TDY groups. In addition, Briscoe et al. reported that both SLI groups had significant word finding deficits. It would appear that although the TD and SLI groups were matched in one way, deficits in another area may be clouding the issue.

Gathercole and Baddeley (1990) matched six children with SLI (aged 7;02–8;10) to six younger children on the basis of their very close performance on the BPVS (Dunn & Dunn, 1982). Of the six children with SLI, five had a mean difference between chronological age and performance on the BPVS of 20 months, and a difference of 18 months on the Test of Reception of Grammar (Bishop, 1983). One child was within 12 months for both vocabulary and receptive grammar. On two tests of NWR, the children with SLI (N = 5) scored significantly lower than both TDAM and TDY peers. The authors stated that the phonological working memory deficits of children with SLI must reflect difficulties over and above their language skills as they were outperformed by the TDY groups. Thus far, though, it appears that
this finding has not been replicated with very closely matched SLI and TDY groups in a large sample size. Further work is needed.

**SR**

Given the range of scoring methods used for SR tasks across studies, it was important to compare the effects of using one method over another. Awarding points according to the number of errors per sentence proved to be the most effective method, as it successfully differentiated between children with SLI and their TDAM peers. It is important, too, to consider the types of sentences chosen for the test. Application of two sentence types (aspect and passive) together proved to be the best identifier of SLI in this cohort while ensuring sufficient sentence complexity to avoid ceiling performance in the TDAM group (Redmond, 2005).

SR has the potential to be a useful clinical marker of SLI in Cantonese, as it is in English (Conti-Ramsden et al., 2001). To be useful, cutoff points must be established so that the test can be administered (perhaps to preschool children) as a screen. Space limitations prevent pursuit of this topic here, but see Conti-Ramsden et al. (2001) for an example of use of cutoff points for clinical markers. Apart from field testing and validation, theoretical issues also remain. What does a deficit in SR imply? Conti-Ramsden et al. (2001) suggested that poor performance by children with SLI on SR might be attributable to "limitations in short-term memory" (p. 747), a mechanism common with impairments in NWR. The Cantonese results do not support this suggestion. Although other research has found a significant correlation between performance on NWR and performance on SR (Bishop et al., 1996; Conti-Ramsden et al., 2001; Kamhi & Catts, 1986), there was no correlation in this study.

The Cantonese results suggest that NWR and SR tests may depend on short-term memory and language abilities to varying degrees across the preschool period. The scores on these two tests for the TDAM group (being the most able language users of the three groups) were not related to their scores on any of the language measures. However, the TDY children's scores on both NWR and SR were strongly correlated with their receptive vocabulary scores ($r_s = .65$ and .90). SR in the TDY group was also significantly correlated with MLU as well as with receptive grammar. Processing tasks are not independent of language performance at 3 years of age but may be independent by age 5.

Although the NWR scores were not related to language scores in the SLI group, SR scores were moderately correlated with receptive grammar scores ($r = .58$). The SLI group outscored the TDY group on NWR, but not on SR, suggesting better performance for the SLI group when short-term memory mechanisms are implicated, but when language ability is involved, the SLI and TDY groups performed equitably. It is notable that the two groups did not differ on any of the language measures. The parsimonious explanation is that the ability to repeat sentences must to some degree be affected by language ability, unlike NWR, which is assumed to be language independent. It remains possible that there is a third mediating factor, not explored here, that underlies performance on both SR and grammatical comprehension, perhaps sequencing or temporal processing factors.

Our results do not suggest a limitation in phonological working memory in Cantonese-speaking children with SLI, with the caveat that this may be due to the relatively simple structure of Cantonese words (and nonwords). Poor NWR by children with SLI in other languages might be due to the prosodic (temporal and sequential) properties of the languages or superior redintegration skills of TDAM children. Further work might explore prosodic and processing factors in cross-linguistic studies.

**Acknowledgments**

This work was supported by Research Grant R01 DC00458 from the National Institute on Deafness and Other Communication Disorders. We thank the Hong Kong Christian Service Choi Wan and Yuen Long Early Education and Training Centres; the Speech and Language Clinic at the Division of Speech and Hearing Sciences, University of Hong Kong; the Pamela Youde Child Assessment Centre; and the Yan Chai Hospital. Special thanks are extended to the children and their families.

**References**


Marshall, C., Ebbels, S., Harris, J., & van der Lely, H. K. J. (2002). Investigating the impact of prosodic complexity on
the speech of children with specific language impairment. In R. Vermeulen & A. Neeleman (Eds.), *UCL working papers in linguistics* (Vol. 14, pp. 43–66), Department of Phonetics and Linguistics, University of London.


Received January 12, 2005
Revision received May 8, 2005
Accepted July 21, 2005
DOI: 10.1044/1092-4388(2006/019)

Contact author: Stephanie F. Stokes, who is now at Speech and Language Sciences, School of Education, Communication and Language Sciences, King George VI Building, University of Newcastle upon Tyne, NE7 1RU, United Kingdom. E-mail: s.stokes@ncl.ac.uk
**Appendix A.**

**Trial Items**
- maa1mi4 (word)
- bei2gaa1ciu1 (word)
- keon6
- boe5haat1

**Experimental Items**

<table>
<thead>
<tr>
<th>One syllable</th>
<th>In CV</th>
<th>Out CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>tet6</td>
<td>te</td>
<td></td>
</tr>
<tr>
<td>weng6</td>
<td></td>
<td>we</td>
</tr>
<tr>
<td>faap6</td>
<td>faap</td>
<td></td>
</tr>
<tr>
<td>kou6</td>
<td></td>
<td>kou</td>
</tr>
<tr>
<td>pam6</td>
<td>pa</td>
<td></td>
</tr>
<tr>
<td>myun6</td>
<td></td>
<td>myu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Two syllable</th>
<th>In CV</th>
<th>Out CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>nu5pim1</td>
<td>pi</td>
<td>nu</td>
</tr>
<tr>
<td>gyuStak1</td>
<td>gyu ta</td>
<td></td>
</tr>
<tr>
<td>woi5kaan1</td>
<td>kaa</td>
<td>woi</td>
</tr>
<tr>
<td>dui5faap1</td>
<td>fa</td>
<td>dui</td>
</tr>
<tr>
<td>jou5moeng1</td>
<td></td>
<td>moe jou</td>
</tr>
<tr>
<td>te5lun1</td>
<td>te</td>
<td>lu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Three syllable</th>
<th>In CV</th>
<th>Out CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>leoi3beoi1</td>
<td>leoi</td>
<td>beoi jo</td>
</tr>
<tr>
<td>moi3woe1</td>
<td>paa</td>
<td>moi woe</td>
</tr>
<tr>
<td>taau3feu1</td>
<td>ma</td>
<td>taau feu</td>
</tr>
<tr>
<td>noe3foi1</td>
<td>noe toe</td>
<td></td>
</tr>
<tr>
<td>kyu3jou1boeng6</td>
<td>kyu</td>
<td>boe jou</td>
</tr>
<tr>
<td>feoi3tei1het6</td>
<td>he</td>
<td>feoi tei</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Four syllable</th>
<th>In CV</th>
<th>Out CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>keu1wau2hui4feng6</td>
<td>keu fe</td>
<td>wau hui</td>
</tr>
<tr>
<td>hu1leoi2peoi4nen6</td>
<td>leoi ne</td>
<td>hu peoi</td>
</tr>
<tr>
<td>lui 1 daau2kou4bem6</td>
<td>be</td>
<td>lui daau kou</td>
</tr>
<tr>
<td>feoi1lyu2te4faap6</td>
<td>lyu te faa</td>
<td>feoi</td>
</tr>
<tr>
<td>kou1boe2moi4lun6</td>
<td>kou boe moi lu</td>
<td></td>
</tr>
<tr>
<td>woi 1 tei2nyu4faam6</td>
<td>nyu faa</td>
<td>woi tei</td>
</tr>
</tbody>
</table>
Appendix B.

Sentence Repetition: Aspect Marker

Trial Items
1. ngo5dei6 heoi3 gung1jyun2
   we go park
   We go to the park.

2. sai3lo2 hai2 fong2 dou6 duk6syu1
   Little brother at room PRT study
   Little brother is studying in the room.

3. Wong4 lou5si1 gam1jat6 hou2 leng3
   Wong teacher today very pretty
   Miss Wong is very pretty today.

Experimental Items
1. go3 naam4za2i2 hou2 daai6lik6 caat3 gan2 deoi3 haai4
   CL boy very forceful polish ASP CL shoe
   The boy is polishing the shoes very hard.

2. zaat3 bin1 go3 neoi5za2i2 zing2 di3 zo2 byun2 syu1
   tie pigtail CL girl make fall ASP CL book
   The girl with the pigtail has dropped the book.

3. go4go1 m4 siu2sum1 dou2 se2 zo2 bu1 seoi2
   brother not careful pour spilled ASP CL water
   Big brother has spilled the glass of water by accident.

4. daai3 ngaan5geng2 go3 naam4za2i2 daa2 gan2 cin1 cau1
   wear spectacles CL boy play ASP swing
   The boy wearing spectacles is playing on the swing.

5. go3 neoi5za2i2 hou2 daai6seng1 gong2 gan2 gu3si6
   CL girl very loudly tell ASP story
   The girl is telling a story very loudly.

6. zeok3 fu3 go3 naam4za2i2 daa2 i aan6 zo2 go3 wun2
   wear pants CL boy break broken ASP CL bowl
   The boy wearing pants has broken the bowl.

7. go3 neoi5za2i2 hou2 fai3 gam2 sai2 zo2 gin6 saam1
   CL girl very fast ADVI wash ASP CL shirt
   The girl has washed the shirt very quickly.

8. zeok3 dyun2 kwun4 go3 neoi5za2i2 cai3 gan2 cai3tou4
   wear short skirt CL girl put-together ASP puzzle
   The girl wearing a miniskirt is putting together a puzzle.

Sentence Repetition: Passive

Trial Items
1. keoi5 zung1 ji3 sik6 fan6 s/he
   likes eat rice
   S/he likes to eat rice.

2. ngo5 soeng2 fung4 maa1 mi4 haang4 gaaai1
   I want with mummy shopping
   I want to go shopping with mummy.

3. hok6haau6 jau5 hou2 do1 je5 waan2
   school have very many things play
   There are many playthings in the school.

Experimental Items
1. jau5 go3 neoi2za2i2 bei2 go3 naam4za2i2 daa2
   have CL girl PASS CL boy hit
   A girl was hit by the boy.

2. mui4mu2 bei2 go4go 1 ceong2 zo2 baa2 so 1
   little sister PASS brother grab ASP CL comb
   Little sister’s comb was grabbed by the big brother.

3. daai3 mou2 go3 neoi5za2i2 bei2 zek3 maa1 ngaau5
   wear hat CL girl PASS CL cat bite
   The girl wearing a hat was bitten by the cat.

4. ze4ze1 bei2 saam1 zek3 baan1 dim2 gau2 fai6
   big sister PASS three CL spot dog bark
   Big sister was barked at by three Dalmatians.

5. ceong4 tau4 faat3 go3 naam4za2i2 bei2 zek3 gau2 zeoi1
   long hair CL boy PASS CL dog chase
   The boy with long hair was chased by the dog.

6. jau5 go3 naam4za2i2 bei2 go3 neoi5za2i2 tek3
   have CL boy PASS CL girl kick
   A boy was kicked by the girl.

7. mui4mu2 bei2 leong5 zek3 hak1 faa1 maau1 zit1
   little sister PASS two CL black mix-colored cat tickle
   Little sister was tickled by the black mix-colored cat.

8. ze4ze1 bei2 daai4 daai2 lo2 zo2 go3 doi2
   Older sister PASS little brother take ASP CL bag
   Older sister’s bag was taken by the little brother.