Brain & Language 257 (2024) 105450

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Research Article

Revisiting nonword repetition as a clinical marker of developmental language disorder: Evidence from monolingual and bilingual L2 Cantonese

Nga Ching Fu[a,b], Angel Chan[a,b,c,*], Si Chen[a,b,c,d], Kamila Polišenská[e,f], Shula Chiat[e]

[a] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR
[b] Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Hong Kong SAR
[c] The Hong Kong Polytechnic University – Peking University Research Centre on Chinese Linguistics, Hong Kong SAR
[d] Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong SAR
[e] Department of Language and Communication Science, City St George's, University of London, United Kingdom
[f] Division of Psychology, Communication and Human Neuroscience, The University of Manchester, United Kingdom

ARTICLE INFO

Keywords:
Nonword repetition
Cantonese Chinese
Developmental Language Disorder
L2 Cantonese
Specific Language Impairment
Bilingualism
Clinical Marker
Crosslinguistic nonword repetition test

ABSTRACT

Cross-linguistically, nonword repetition (NWR) tasks have been found to differentiate between typically developing (TD) children and those with Developmental Language Disorder (DLD), even when second-language TD (L2-TD) children are considered. This study examined such group differences in Cantonese. Fifty-seven age-matched children (19 monolingual DLD (MonDLD); 19 monolingual TD (MonTD); and 19 L2-TD) repeated language-specific nonwords with varying lexicality levels and Cantonese-adapted quasi-universal nonwords. At whole-nonword level scoring, on the language-specific, High-Lexicality nonwords, MonDLD scored significantly below MonTD and L2-TD groups which did not differ significantly from each other. At syllable-level scoring, the same pattern of group differentiation was found on quasi-universal nonwords. These findings provide evidence from a typologically distinct and understudied language that NWR tasks can capture significant TD/DLD group differences, even for L2-Cantonese TD children with reduced language experience. Future studies should compare the performance of an L2-DLD group and evaluate the sensitivity and specificity of Cantonese NWR.

1. Introduction

Developmental Language Disorder (DLD; previously also known as Specific Language Impairment or SLI), is a neurodevelopmental disorder where children have significant difficulties in understanding and using language that are not associated with other biomedical conditions such as hearing impairment, intellectual disability, autism spectrum disorder (ASD), and brain injury (Bishop et al., 2016, 2017). The task of identifying DLD in bilingual children has been described as a major challenge (Armon-Lotem, 2012; Schwob et al., 2021), due to similarities in language limitations and errors observed in children with transient difficulties that arise from insufficient exposure to a language and resolve with increased exposure, and children experiencing more persistent language difficulties that arise from a language disorder (so observed in monolingual as well as bilingual children with DLD; Camilleri & Law, 2007), where language difficulties are associated with poor prognosis, are unlikely to resolve on their own, and require specialist support. Nonword repetition (NWR) has been advocated as an important tool within test batteries for assessing bilingual children, as children with DLD have been found to perform below typically developing (TD) children in NWR, even in a bilingual context (Schwob et al., 2021).

1.1. NWR in bilingual children

Because children with DLD often have deficits in some or all of the skills required to support the accurate repetition of nonwords, including impaired phonological working memory capacity (Gathercole & Baddeley, 1990; Montgomery, 2002), vocabulary skills (Gray et al., 1999; McGregor, 2009; Watkins et al., 1995), and motor planning deficits (Stark & Blackwell, 1997), the demands of NWR are disproportionally challenging for children with DLD. This may explain why NWR has been found to discriminate between children with and without DLD. Bilingual TD children are, on the one hand, expected to have processing capabilities (e.g. working memory capacity, motor planning) comparable to

[*] Corresponding author at: Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR.
E-mail addresses: ngaching.fu@connect.polyu.hk (N.C. Fu), angel.ws.chan@polyu.edu.hk (A. Chan), sarah.chen@polyu.edu.hk (S. Chen), kamila.polisenska@city.ac.uk (K. Polišenská), shula.chiat.1@city.ac.uk (S. Chiat).

https://doi.org/10.1016/j.bandl.2024.105450
Received 31 August 2023; Received in revised form 13 July 2024; Accepted 31 July 2024
Available online 20 September 2024

those of monolingual TD (MonTD) children, so should perform similarly on NWR. On the other hand, evidence has shown that vocabulary knowledge plays an important role in NWR and that children draw on their existing lexical-phonological knowledge when repeating nonwords, perhaps because this helps them to reconstruct a target nonword as the memory trace decays in working memory in a process known as redintegration. Like children with DLD, bilingual children may have weaker lexical and sub-lexical representations, especially in their L2, to support NWR due to reduced language experience and proficiency, putting them at a disadvantage relative to MonTD children and hence at risk of being misclassified as having DLD. Empirical evidence is in line with these different possibilities: some studies have reported that bilingual TD children performed just as well as MonTD children in NWR (Armon-Lotem, 2018; Armon-Lotem & Chiat, 2012; Lee & Gorman, 2013; Messer et al., 2010), while others have reported that bilingual TD children performed below MonTD children (Boerma et al., 2015; Kohnert et al., 2006; Sharp & Gathercole, 2013). Moreover, bilingual TD groups have been reported to perform above monolingual DLD (MonDLD) groups (e.g. Boerma et al., 2015; Thordardottir & Brandeker, 2013), as well as bilingual DLD groups (Schwob et al., 2021), though other studies show overlapping NWR performance in L2-TD and MonDLD groups (e.g., Kohnert et al., 2006; Windsor et al., 2010).

1.2. Language-Specific vs. Quasi-Universal NWR tests

Whether bilingual TD children are disadvantaged in NWR in comparison to MonTD children appears to be partly related to the characteristics of the nonword stimuli used. A *meta*-analysis (Schwob et al., 2021) reported that differences in NWR performance between monolingual and bilingual TD groups tend to be minimized in studies that used quasi-universal NWR tests, rather than language-specific NWR tests. Unlike traditional NWR tests that are created with reference to phonotactic possibilities of a given language (i.e. language-specific NWR tests), quasi-universal NWR tests are designed to be maximally compatible with the lexical phonology of diverse languages, hence being “quasi-universal”. One such example is the crosslinguistic NWR task (CL-NWR; Chiat, 2015), where nonword stimuli (e.g. *bamudi*) were designed to contain only crosslinguistically frequent consonants and vowels in simple CV syllables, and are set to be articulated with neutral prosody, through applying even length and stress on all syllables equally. By removing elements of nonwords that are relatively uncommon across languages and may only be familiar to speakers of certain languages (e.g. lexical tones in Mandarin and Cantonese, consonant clusters in English, Polish and French), nonword stimuli in CL-NWR reduce the likelihood of disadvantaging bilingual children with reduced experience of a given language, therefore minimizing the gap between monolingual and bilingual TD groups.

The CL-NWR test was evaluated by Boerma et al. (2015) in its ability to differentiate TD and DLD children in both monolingual and bilingual groups. The study compared four groups of five- to six-year-old children – monolingual Dutch-speaking TD children, monolingual Dutch-speaking children with DLD, bilingual Dutch-speaking TD children, and bilingual Dutch-speaking children with DLD (all bilingual children were second language (L2) learners of Dutch with different first languages (L1s)) – on their NWR performance in the Dutch version of the CL-NWR task versus a Dutch language-specific NWR task. Results indicated that children with DLD, regardless of monolingual or bilingual status, performed below TD children on both the language-specific and quasi-universal NWR tasks. Bilingual TD children performed below monolingual TD children in the language-specific NWR task, but not the quasi-universal CL-NWR task; and monolingual and bilingual children with DLD performed at a similar level in both language-specific and quasi-universal NWR tasks. Aside from group comparisons, the study also examined the diagnostic accuracy of the language-specific NWR task and the Dutch CL-NWR task, and found that both tasks had adequate levels of sensitivity and specificity in the monolingual sample, but the language-specific task fell short in classifying L2 learners of Dutch with and without DLD. These results suggest that the quasi-universal CL-NWR test may have higher clinical accuracy than language-specific NWR tasks when used in a bilingual population, providing support that a quasi-universal NWR task may be better able to separate the effects of language impairment and bilingualism, making CL-NWR a promising tool for identifying DLD in bilingual children.

More recent studies on the CL-NWR task have reported less clear-cut findings. In a follow-up longitudinal study (Boerma & Blom, 2021), the same groups of children included in Boerma et al. (2015) were retested using the Dutch CL-NWR task at one-year intervals over a two-year period. At age six- to seven- years (second wave of testing) and seven- to eight-years (third wave of testing), sensitivity and specificity of the Dutch CL-NWR task was found to drop to clinically unacceptable levels (<80 % accuracy) for both monolingual and bilingual children, although significant and large group differences were still yielded between TD and DLD groups, while bilingual L2 children were not found to perform below their monolingual counterparts. Furthermore, diagnostic accuracy reached clinically acceptable levels when the Dutch CL-NWR task was used in combination with a narrative assessment, suggesting that while the CL-NWR task may not be suitable to be used as a standalone diagnostic tool, especially for children aged six years or older, it can be an informative measure when used within test batteries to assess monolingual and bilingual children.

In another study (Öberg & Bohnacker, 2022), NWR performance on the Swedish CL-NWR task was examined in four- to seven-year-old bilingual children acquiring L1-Arabic and L2-Swedish. Due to a large discrepancy in the sample size of the L2-TD (N=99) and L2-DLD (N=11) groups in this study, group means could not be compared, but z-scores (derived using the TD group as the reference population) revealed substantial overlap in performance between the bilingual-TD and bilingual-DLD groups, thus it was concluded that NWR performance on the Swedish CL-NWR task could not reliably distinguish bilingual children with DLD from bilingual TD children. Therefore, the degree of TD/DLD group differentiation and clinical accuracy of the CL-NWR task, particularly on bilingual children, may depend on the specific languages (and L1-L2 language combinations) examined. Other factors, such as criteria for recruitment and identification of the DLD sample, which could bear on the profiles of cognitive and linguistics abilities of samples (see also suggestions for future research in the conclusions section), could also contribute to different findings. As such, it is important for the CL-NWR task to be examined in typologically diverse languages.

1.3. NWR in monolingual and bilingual Cantonese-Speaking children

Until recently, Cantonese has been a rare cross-linguistic exception, with a study documenting that NWR did not differentiate between TD children and those with DLD among monolingual Cantonese-speaking children (Stokes et al., 2006). More recently, our team (Fu et al., 2024a) reported that a novel set of language-specific Cantonese nonword stimuli with varying lexicality levels was able to capture significant TD/DLD group differences among predominantly monolingual[1] Cantonese-speaking children, suggesting that Cantonese is not a true exception, and that NWR has potential to serve as a cross-linguistic clinical marker of DLD. The study also found that nonwords of higher lexicality and sub-lexicality levels (i.e., those with a greater resemblance to real words, specifically in terms of morphemicity and

[1] These children acquire Cantonese as L1/home language and Cantonese is the medium of instruction used in the schools they attend for Chinese and other subjects (except English and Mandarin subjects). They are different from, for example, heritage speakers of Cantonese, in that they are exposed to additional languages no more than 20% of their waking hours, thus we describe these children as “predominantly monolingual” rather than bilingual L1 speakers of Cantonese, following common operational definitions of mono-/bi-lingualism.

consonant–vowel combination attestedness) captured greater TD/DLD group differences in Cantonese-speaking children, suggesting that lexicality and sub-lexicality effects must be taken into consideration in the design of NWR stimuli for generating TD/DLD group differences.

To further examine the potential utility of NWR for TD/DLD differentiation in Cantonese-speaking children, this study explored whether our new NWR tasks are still able to accurately capture TD/DLD group differences in Cantonese-speaking children when bilingual L2-TD learners with reduced exposure to and knowledge of the language of testing are taken into account. When assessing bilingual children using NWR, an ideal scenario would be to have nonwords that do not disadvantage bilingual L2 Cantonese TD children with reduced target language experience (i.e. that minimize the gap between monolingual and bilingual children sharing the same TD status), and are able to capture significant group differences between DLD and TD groups even where the TD group is L2 (i.e. that maximize the gap between MonDLD and L2-TD as well as MonTD). While our findings on predominantly monolingual Cantonese-speaking children suggest that high lexicality and sub-lexicality nonwords are better able to capture TD/DLD group differences, this may not be the case when bilingual L2 learners are taken into account, as L2 learners may be less able to benefit from the increase in lexicality due to weaker lexical and sub-lexical representations in their L2 compared to MonTD children. Therefore, nonwords that are similar to real words in the ambient language may disadvantage rather than support L2-TD relative to MonTD children. Thus, it is also important to examine the influence of lexicality of nonwords on the pattern and degree of group differentiation.

In this study, we examine this potential in language-specific NWR stimuli (reported in Fu et al., 2024a), as well as quasi-universal, Cantonese-adapted CL-NWR stimuli (Chiat, 2015), by comparing performance in a group of TD, L1-Urdu-L2-Cantonese-speaking children residing in Hong Kong[2] with performance in predominantly-monolingual Cantonese-speaking TD children and their peers with DLD. The choice of this particular bilingual group was influenced by three factors. First, people of Pakistani origin constitute 23.9 % of the large South Asian communities residing in Hong Kong, according to the Hong Kong 2021 Population Census. Poon (2010) noted that Urdu, rather than the local official languages of Hong Kong (i.e. Cantonese, English, and Mandarin), is the preferred language in these Pakistani families, as many parents are not proficient in Cantonese. As a result, most of these Pakistani children are bilingual learners using the minority language Urdu as first language (L1) at home, while acquiring Cantonese as L2 in a school and community context, reducing exposure to both languages. Second, assessing these children in their L1 is challenging, given that little is known about developmental expectations for this bilingual group, with language assessment tools in Urdu being unavailable or inaccessible for local speech and language therapists (SLT). Third, these children often come from families with low socio-economic status (SES), where parents may sometimes lack sensitivity to the possibility of language disorder when their child presents with language difficulties, and may be less likely to seek support from professionals for assessment. Therefore, relative to other bilingual groups of children in Hong Kong, such as English-Cantonese and Mandarin-Cantonese bilingual children, testing L1-Urdu-L2-Cantonese-speaking bilingual children provides a particular motivation and opportunity to examine whether our novel Cantonese NWR stimuli disadvantage bilingual L2 children likely to be socioeconomically disadvantaged and have weak language skills in the testing language (i.e. Cantonese). Moreover, there is a greater and more urgent need for suitable assessment tools to be developed for Urdu-Cantonese bilingual children, given the current lack of suitable assessment tools.

1.4. *Scoring methods in NWR*

As a further extension to the line of work on Cantonese NWR, this study will examine the effects of using different scoring methods in NWR on group differentiation. Previous studies on other languages have found different levels of clinical accuracy when different scoring methods were used. The two most commonly used approaches were percentage of items correct (PIC; i.e. scoring NWR at whole-item level), and percentage of phonemes correct (PPC; i.e. scoring NWR at phoneme level). Studies have found that both methods discriminated between TD and DLD groups in both monolingual and bilingual children (Schwob et al., 2021), with some studies finding that PIC generated higher levels of diagnostic accuracy than PPC (Dispaldro et al., 2013; Guiberson & Rodríguez, 2013), and others reporting no difference between PIC and PPC, especially in bilingual populations (Boerma et al., 2015; le Clercq et al., 2017). Practically speaking, studies have also noted that PIC is faster and easier to score, making it more suitable for use in speech and language therapy clinics than PPC (Dispaldro et al., 2013; Pham & Ebert, 2020). While findings on PIC vs. PPC scoring are not definitive, it is clear that scoring methods do affect the power of TD/DLD group differentiation (see Ortiz, 2021; and Schwob et al., 2021). As no studies have examined how scoring approaches in Cantonese NWR tasks affected group differentiation between TD children and those with DLD, we aim to also examine whether two scoring approaches – scoring Cantonese NWR at a syllable level vs. whole-nonword level – produce a different pattern or degree of group differentiation in Cantonese MonTD, MonDLD, and L2-TD children. We opted to compare whole-nonword level scoring with syllable level scoring, rather than PPC, as syllable level scoring allowed for a more fine-grained method of scoring (relative to whole-nonword scoring) to be examined, whilst still offering advantages over PPC in being faster and easier to score.

1.5. *The present study*

This study aims to examine the ability of language-specific (with varying lexicality levels) and quasi-universal Cantonese NWR stimuli to minimize the differences between MonTD and L2-TD children, whilst maximizing differences between MonTD and MonDLD groups, and between L2-TD and MonDLD groups, in L1-Urdu-L2-Cantonese-speaking children.

Specifically, two research questions (RQ) are addressed in this study:

RQ1: When NWR accuracy is scored on whole-nonwords correct, can language-specific Cantonese nonwords (including High-Lexicality, Low-Lexicality, High-lexicality-Vowel-Matched nonwords), and Cross-linguistic nonwords capture significant group differences between MonTD and MonDLD children, and between L2-TD and MonDLD children, while minimizing group differences between MonTD and L2-TD children?

RQ2: When NWR accuracy is scored on syllables correct, can language-specific Cantonese nonwords (including High-Lexicality, Low-Lexicality, High-lexicality-Vowel-Matched nonwords), and Cross-linguistic nonwords capture significant group differences between MonTD and MonDLD children, and between L2-TD and MonDLD children, while minimizing group differences between MonTD and L2-TD children?

Addressing these RQs will lay important foundations for future studies examining the clinical utility and accuracy of NWR for identifying DLD in bilingual L2 Cantonese-speaking children, and add to the understanding of how lexicality and language-specificity of nonword

[2] We were unable to include an additional group of L2 Cantonese-speaking children with DLD, as assessment tools for L2 Cantonese-speaking children are currently under development – these assessment tools have been demonstrated to be useful for identifying DLD in a case study by our team members (Hamdani et al., 2024), but findings have to be replicated in larger scale studies to confirm the diagnostic potential of these new assessment tools. Hence, it is currently very difficult to identify a group of L2 Cantonese-speaking children with DLD, who also have to have comparable language experience to the L2-TD group.

stimuli, as well as scoring approaches, affect TD/DLD group differentiation.

2. Methods

2.1. Participants

Fifty-seven Cantonese-speaking children participated in this study.[3] The children were either recruited online or re-invited to take part in this research study, after previously participating in other projects. All children attended local schools in Hong Kong, where Cantonese was the medium of instruction (MOI).

2.1.1. Monolingual DLD group

The first group of children (*N*=19, fourteen male), aged 8;1 to 11;10[4] (*M*=9;8, *SD*=1;0), were predominantly monolingual[5] Cantonese-speaking children, who met the criteria for DLD (MonDLD), on the basis that they demonstrated poor language skills in the norm-referenced, Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS; T'sou et al., 2006), and that their language difficulties had a negative functional impact on daily social interactions or educational progress, as reported by parents and/or school personnel. In HKCOLAS, seventeen of these children scored at 1.25 SD below age means in two or more out of six subtests, and two scored at 1.25 SD below age means in one subtest and 1 SD below age means in another. One child had co-occurring attention deficit hyperactivity disorder (ADHD), and another child had co-occurring dyslexia, neither of which are considered differentiating conditions of DLD, under the CATALISE diagnostic guidelines (Bishop et al., 2017).

2.1.2. Monolingual TD group

The second group of children (*N*=19, fourteen male), were predominantly monolingual TD children (MonTD), who were individually matched to each child in the DLD group in age (within four months of age difference on the day of testing, *M*=9;8, *SD*=1;1, range = 8;0 to 11;9), gender, and grade in school. These children scored age-appropriately in HKCOLAS and there were no parental concerns over any areas of development.

2.1.3. Bilingual TD group

The third group of children (*N*=19, seven male) were L1-Urdu-L2-Cantonese-speaking TD children (L2-TD), who were acquiring Cantonese as a second, school and community language, while using Urdu as a heritage language at home. These children were not individually matched to each monolingual DLD-TD pair in age, gender, and grade in school, but as a group, they had a comparable age range with the two groups of predominantly monolingual children (*M*=9;7, *SD*=1;1, range = 8;0 to 11;7).[6] As there are no available norm-referenced tests to assess the language profiles of these L2-Cantonese-speaking children, their TD status was established through a parental questionnaire which was adapted from the Language Impairment Testing in Multilingual Settings – Parents of Bilingual Children Questionnaire (LITMUS-PaBiQ; Tuller, 2015). Parents of fifteen children completed the questionnaire – parents of the remaining four were unreachable, but these children had previously participated in a research study in which they were assessed by an experienced Urdu-speaking SLT in a range of language assessment tasks tapping lexical, morphosyntactic and narrative competence in their strongest language (i.e. Urdu). Their performance on these language tasks was considered developmentally appropriate based on the clinical judgment of the experienced SLT. Responses on the questionnaire confirmed that all participants were born at full-term, did not have significantly delayed one-word and word-combination stages, did not have any other developmental delays, did not have a history of receiving speech therapy, did not have hearing impairments, had not been diagnosed with any other developmental disorders (including ADHD, dyslexia, ASD) and did not have family history of language impairments. Parents also expressed no concerns over the children's development in any areas. Urdu was reported to be the strongest language of all children, except for one child who was reported to be slightly stronger in Cantonese. Parents rated their children's Cantonese proficiency in comprehension and production, on a scale of 1 (poor) to 7 (excellent). In comprehension, the mean of rated proficiencies was 5.47/7 (*SD*=1.41, range = 2 to 7); and in production, the mean of rated proficiencies was 5.33/7 (*SD*=1.50, range = 2 to 7), indicating good overall levels of Cantonese proficiency. All children were born in Hong Kong, apart from one (8;00 at testing) who had moved to Hong Kong at 5 years, and there was missing information on one child. No participants had received regular exposure to Cantonese since birth, but most (86.7 %) were exposed to Cantonese since preschool (from 2-5 years of age), and others since school-entry (6–7 years). The mean cumulative number of years of Cantonese exposure was 5.73 (*SD*=1.84, range = 2.0 to 8.58).

No participants from any of the three groups were reported to have hearing impairments or ASD. All children from the MonDLD and MonTD groups passed a pure tone audiometry hearing screening test. They also obtained standard scores of above 70 in Raven's Progressive Matrices (Raven et al., 1996), screening out the likelihood of intellectual disability, although the MonTD group (*M*=110.4, *SD*=13.1, range = 85 to > 135) scored significantly higher than the MonDLD group (*M*=101.6, *SD*=12.2, range = 82 to 127), $t(36) = 2.14$, $p = 0.04$. The L2-TD group did not undergo pure tone audiometry hearing screening and tests for non-verbal intelligence, as the likelihood of having hearing impairments or intellectual disability was deemed to be very low, with no parent suspicion reported through the parental questionnaire for any participant, and all children studying in mainstream schools with no expressed concerns from schools over the children's development in any areas either. All parents gave written consent for their children to participate in the study.

2.2. Materials

2.2.1. Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS)

HKCOLAS (T'sou et al., 2006) is a norm-referenced language

[3] Of these 57 children, the NWR data of 32 children (16 MonTD and 16 MonDLD) were also included in our previous study (Fu et al., 2024a) which analyzed the effects of lexicality of the language-specific stimuli on performance and group differentiation in monolingual children.

[4] We were only able to recruit older children (above the age range covered in most NWR studies), as there was a reluctance from parents to enroll younger children in research studies during the pandemic. Given that TD/DLD significant group differences in NWR have been reported for children and adolescents across different ages, at least up to age 15;4 (Riches et al., 2011; Schwob et al., 2021), we believed that examining children aged 8 to 11 would still allow our research objectives to be appropriately addressed.

[5] See footnote 1 on predominantly monolingual children.

[6] There were no significant group differences in age. As for gender, there were more females in the L2-TD group (12 out of 19) than the monolingual TD and DLD groups (5 out of 19 in each), and unsurprisingly, a chi-square test yielded a significant group difference in gender ratio. However, since studies have reported no gender effects on children's NWR performance (see e.g. Chiat & Roy, 2007; Washington & Craig, 2004), we do not think this gender ratio difference in the L2-TD group would account for findings on the performance of this group relative to the other two groups.

assessment tool published by Child Assessment Services, Department of Health, Government of the Hong Kong SAR. It is designed to examine Cantonese oral language abilities of five- to twelve-year-old children in Hong Kong. HKCOLAS includes six subtests (Test of Hong Kong Cantonese Grammar, Textual Comprehension Test, Word Definition Test, Lexical-Semantic Relations Test, Narrative Test, and Expressive Nominal Vocabulary Test), where children who score 1.25 SDs below age means in two or more subtests were given a diagnosis of language disorder. HKCOLAS has good levels of clinical accuracy (sensitivity: 0.95; specificity: 0.98), test–retest reliability (coefficient alpha: 0.80–0.97 across all subtests), and is widely used by SLTs in Hong Kong to diagnose language disorder or DLD in children.

2.2.2. Pure tone audiometry hearing screening test

The Interacoustics AD226 diagnostic audiometer was used to perform a pure tone audiometry hearing screening test. Children are asked to raise their hands when they hear a beep, where pure tones were presented at 25 dB hearing levels (HL) at frequencies of 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. Children pass the hearing screening if they are able to respond to pure tones at all test frequencies at 25 dB HL in both left and right ears.

2.2.3. Raven's Progressive Matrices

A Hong Kong Chinese adapted version of Raven's Progressive Matrices (Chan, 1984; Raven et al., 1996) was used as a measure of non-verbal intelligence quotient, to screen out the possibility of intellectual disability when assessing children suspected of DLD. In Raven's Progressive Matrices, children answer 60 multiple choice questions requiring them to identify a missing piece from six to eight options that completes a pattern. Children with standard scores of 70 or above are considered to be within the normal range.

2.2.4. Parental questionnaire for establishing TD status

The parental questionnaire used was an adapted version of the LITMUS-PaBiQ (Tuller, 2015). Parents were asked: 1) whether the child was born at full term; 2) whether there were significant delays in the child's early language milestones; 3) whether the child had received speech and language assessments; 4) whether the child had received a diagnosis of any developmental disorders or language impairments; 5) whether the child had hearing impairments or frequent ear infections; 6) whether the parent had concerns over the child's development in any areas; 7) whether there was a history of speech and language impairments in any family members; 8) years living in Hong Kong; 9) age at which regular exposure to Cantonese had begun; and 10) parent's subjective ratings on the child's ability to speak and understand Urdu and Cantonese (rated on a scale of 1 to 7, with 1 being "poor" and 7 being "excellent"). The parental questionnaire was administered in the form of a phone interview by an Urdu-speaking SLT, such that parents were able to use a language they were familiar with (Urdu in this case) when responding to the questionnaire.

2.2.5. Nonword repetition stimuli

Three sets of language-specific nonwords, reported in Fu et al. (2024a), and one set of quasi-universal, Cantonese-adapted cross-linguistic nonwords were used (see Supplementary Materials for full list).

2.2.5.1. Language-Specific nonwords

2.2.5.1.1. High-Lexicality nonwords. High-Lexicality nonwords were created solely with morphemic syllables in Cantonese that do not form a meaningful combination, e.g. *fe1 ji1 maa1*[7] (*fe1* meaning brown, *ji1* meaning clothing, *maa1* meaning mom). The morphemic status of each syllable was validated through native speaker judgements by the first and second authors, and cross-checked with the Cantonese syllabary (Bauer & Benedict, 1997). There were 24 items in total (3 items x 4 lengths x 2 syllable complexity)

2.2.5.1.2. Low-Lexicality nonwords. Low-Lexicality nonwords contained only syllables that were non-morphemic in Cantonese across all six contrastive lexical tones, e.g. *ngu1 fi1 hu1* (*ngu*, *fi*, and *hu* each have no meaning regardless of lexical tone and are meaningless when combined). Essentially, low lexicality syllables are accidental phonological gaps in Cantonese. These stringent syllable selection criteria resulted in Low-Lexicality nonwords having a smaller vowel range than High-Lexicality nonwords. There were 24 items in total (3 items x 4 lengths x 2 syllable complexity)

2.2.5.1.3. High-Lexicality-Vowel-Matched nonwords. To enable cleaner comparisons, High-Lexicality-Vowel-Matched nonwords were created, such that they matched with High-Lexicality nonwords in terms of lexicality, but with Low-Lexicality nonwords in terms of vowel range[8]; e.g. *lo1 fo1* (*lo1*, sentence final particle, *fo1* meaning science). There were 24 items in total (3 items x 4 lengths x 2 syllable complexity)

In all three nonword sets described above, syllables that sounded like real English words (e.g., *wet* or *fit*) and syllable combinations that resembled Cantonese multi-syllabic words were avoided. Each of the sets were also manipulated on length, where items ranged from two to five syllables; and rime structure (rime refers to the sequence of all phonemes following the onset of a syllable), where half of the items were constructed with CV syllables (i.e. with a simpler rime structure of V), and the other half CVC syllables (i.e. with a more complex rime structure of VC). All syllables within all nonwords were set to be articulated in Cantonese tone one (i.e. high, level tone) with even length and stress.

2.2.5.2. Cross-linguistic (CL-NWR) nonwords.

The Cantonese adapted nonwords from the quasi-universal, CL-NWR test (Chiat, 2015) contained only cross-linguistically frequent consonants and vowels, had neutral prosodic features, and simple CV structures (there were no CVC nonwords). Like the other three sets of nonwords, CL-NWR nonwords ranged from two to five syllables in length. When compared with the three sets of language-specific nonwords in terms of lexicality, CL-NWR nonwords had a medium level of lexicality when adapted to Cantonese, in that 57 % (12/21) of their constituent syllables were morphemic in Cantonese, e.g. *sibu* (*si1* meaning poem or silk, and *bu* being meaningless across lexical tones). To facilitate cross-linguistic comparisons, the prosodic pattern applied in the original design of the CL-NWR nonwords – even length, pitch and stress on all syllables except the final syllable, which was lengthened and articulated with a lower pitch to mark the end of an utterance – was retained in this Cantonese version of the nonword set. Such prosodic pattern resembled Cantonese tone one (i.e.

[7] Examples of nonword stimuli are transcribed in the Jyutping phonetic transcription system. In Jyutping, the six Cantonese contrastive lexical tones are denoted by numbers 1–6 following each syllable – the number 1 following all syllables in our examples indicate Cantonese tone one (high-level tone).

[8] Lexicality was a key variable in the study, comparing performance on High-Lexicality Nonwords, in which all syllables are morphemes, with Low-Lexicality Nonwords, in which no syllable is a morpheme in any of the six contrastive lexical tones within the Language-Specific Nonwords by design. The criteria for Low-Lexicality nonwords yielded a restricted set of syllables and we noted that these contained a smaller vowel range than the High-Lexicality nonwords. To et al. (2013) reported that L1 Cantonese speech production accuracy of the same consonant can vary across vowel contexts in their population study of L1 Cantonese-speaking children's acquisition of Hong Kong Cantonese consonants, vowels, and tones, and the age at which children reached the 90% acquisition criterion could differ significantly when this required correct production in all three vowel contexts (the more stringent approach they chose) versus when the target consonant was produced correctly in two out of three vowel contexts; we therefore added the High-Lexicality-Vowel-Matched condition to match the limited range of vowels in the Low-Lexicality condition (just 2 different vowels in CV constituent syllables and 3 different vowels in CVC constituent syllables). This avoided the confound of high/low lexicality with differences in vowel range, enabling 'cleaner' comparisons between lexicality conditions.

high, level tone) on all non-final syllables, and Cantonese tone six (low-mid, level tone) on final syllables. Since such a tone pattern (i.e., having the same, high, level tone across consecutive syllables) is unusual in Cantonese, we argue that even though real monosyllabic words (i.e. morphemic syllables) were included within items in this nonword set, it is unlikely that children are extracting morphosyntactic cues from the constituent syllables to support their repetition. There were 16 items in total (4 items x 4 lengths x 1 syllable complexity)

Across all four sets of nonwords, all consonants and vowels included were expected to be acquired by age 4;0 in speech production by monolingual Cantonese-speaking children according to developmental norms (To et al., 2013). All nonwords were recorded by a female native Cantonese-speaking student SLT.

2.3. Procedures

All experimental tasks were administered by trained native Cantonese-speaking experimenters, and diagnostic testing was conducted by Cantonese-speaking SLTs or student SLTs under the supervision of an experienced SLT. Children from the MonDLD and MonTD groups attended the testing session at a SLT clinic, where they completed a hearing screening, followed by the NWR task, then a standardised language assessment and finally a test of non-verbal intelligence quotient; the session lasted for approximately two hours. Children from the L2-TD group were tested through a home-visit session, as they did not have to undergo diagnostic testing at a clinic due to the lack of norm-referenced language assessment tools. The session started with a warm-up task, where children verbally named items, locations, occupations and items of clothing shown in pictures, before moving on to the NWR task. The session lasted for about 45 min.

The NWR task was separated into two experimental blocks, which were the quasi-universal block (consisting of Cantonese adapted CL-NWR items) and the language-specific block (consisting of High-Lexicality, Low-Lexicality and High-Lexicality-Vowel-Matched items); CL-NWR items were presented independently to facilitate future cross-linguistic comparisons. The order of presentation of items within each block was randomised, and the order of presentation of the two experimental blocks was counterbalanced across participants.

The NWR task was embedded into a picture story presented through PowerPoint slides, following the design of Polišenská and Kapalková's (2014) computerised NWR task. Nonword stimuli and task instructions were pre-recorded, and participants listened to the recordings through noise cancelling headphones in a quiet room. The task began with two practice trials, where children were instructed to listen to and repeat magic words (i.e. nonwords) exactly as they heard them. After each attempt made by the child, a bead appeared on screen. To ensure that the task requirements were understood, replays of the stimuli were permitted and children were given feedback on accuracy during the practice trials. The two experimental blocks followed, where the NWR task was presented as stories about helping story characters repair a broken necklace or bracelets by repeating nonwords exactly as they heard them. Beads appeared on screen after an attempt had been made, until the necklace or bracelets were repaired, marking the end of an experimental block. Replays were not allowed, except when the presentation of stimuli was interrupted (e.g. by the child talking) which was infrequent. Feedback on accuracy was not given during experimental trials.

2.4. Scoring

Responses were audio recorded and transcribed. Performance on all nonwords was scored on both whole-nonword-level accuracy (i.e. responses must contain all and only the target segments in the correct order to be regarded as correct) and syllable-level accuracy (i.e. each correctly repeated syllable within a response gets a score of one). Two sound variations in the responses were not regarded as incorrect, which are the omission of initial /ŋ/ consonant, and substitutions between final /k/ and final /t/ consonants, because these are well documented free variants that are prevalent even in adult native Cantonese speakers (To et al., 2013). Under both scoring approaches, accuracy was scored on consonants and vowels only but not tones, as instances of tone changes were extremely rare and where they occurred, they resembled hesitations rather than repetition failures based on judgements of experimenters and authors. All raters were instructed to disregard any changes in tone in children's responses.

2.5. Inter-rater reliability

Data from the two groups of monolingual Cantonese-speaking children (MonTD and MonDLD) were transcribed and scored by seven native Cantonese speakers with linguistics training. Five completed the first round of transcriptions and scoring, and three independently transcribed 36.8 % of the data (one transcriber acted as the first transcriber for some children and the second transcriber for other children). For NWR scores at whole-nonword-level of accuracy, the average measure Intra-class Coefficient (ICC) using a two-way mixed model and absolute agreement was 0.94 for High-Lexicality-Vowel-Matched nonwords (95 % CI of 0.87 to 0.91); 0.89 for High-Lexicality nonwords (95 % CI of 0.86 to 0.91); 0.89 for Low-Lexicality nonwords (95 % CI of 0.70 to 0.98), and 0.93 for CL-NWR nonwords (95 % CI of 0.91 to 0.95), indicating good to excellent levels of reliability between raters. For NWR scores at syllable-level of accuracy, the average measure Intra-class Coefficient (ICC) using a two-way mixed model and absolute agreement was 0.89 for High-Lexicality-Vowel-Matched nonwords (95 % CI of 0.88 to 0.90); 0.88 for High-Lexicality nonwords (95 % CI of 0.87 to 0.89); 0.85 for Low-Lexicality nonwords (95 % CI of 0.84 to 0.87), and 0.93 for CL-NWR nonwords (95 % CI of 0.92 to 0.94), indicating good to excellent levels of reliability between raters.

The same approach was used to compute inter-rater reliability in the data from the L2-TD group. Five native Cantonese speakers with linguistic training transcribed and scored the data. One completed the first round of transcriptions and scoring, and four independently transcribed 36.8 % of the data and scored NWR accuracy both at whole-item level and syllable level. When NWR was scored at whole-nonword level, the average measure Intra-class Coefficient (ICC) using a two-way mixed model and absolute agreement was 0.89 for High-Lexicality-Vowel-Matched nonwords (95 % CI of 0.84 to 0.92); 0.87 for High-Lexicality nonwords (95 % CI of 0.82 to 0.91); 0.79 for Low-Lexicality nonwords (95 % CI of 0.71 to 0.85), and 0.92 for CL-NWR nonwords (95 % CI of 0.89 to 0.95), indicating good to excellent levels of reliability between raters. When NWR was scored at syllable level, the average measure Intra-class Coefficient (ICC) using a two-way mixed model and absolute agreement was 0.88 for High-Lexicality-Vowel-Matched nonwords (95 % CI of 0.83 to 0.88); 0.85 for High-Lexicality nonwords (95 % CI of 0.82 to 0.87); 0.83 for Low-Lexicality nonwords (95 % CI of 0.80 to 0.86), and 0.86 for CL-NWR nonwords (95 % CI of 0.82 to 0.88), indicating good to excellent levels of reliability between raters.

2.6. Data analysis

NWR scores were analysed with mixed effects logistic regression models, using the R package lme4 (Bates & Maechler, 2010) in R (version 4.1.3, R Core Development Team, 2021). Mixed effects logistic regression models are used as they are well suited for analyzing binary outcome variables, such as correct/incorrect responses, since they can directly model the probability of the binary outcome (Agresti, 2002, p. 565). In addition, they allow for the inclusion of random effects, which can help provide more accurate estimates of the fixed effects (Spieler & Schumacher, 2020, p.5), and the logistic link function in logistic mixed models allows for direct interpretation of the effects in terms of odds ratios, which is better than interpreting coefficients from linear models fitted to the percent correct (Hosmer et al., 2013, p. 322). All assumptions required by mixed effects logistic regression models were met and no data transformation was conducted. While the models analyze NWR accuracy as a categorical variable (i.e. correct/incorrect responses; see following paragraph for details), descriptive statistics for NWR accuracy will be reported in percentages, for intelligibility of the data.

Two statistical models addressed each of the two RQs, the first focussing on NWR scoring at a whole-nonword level; and the second focussing on NWR scoring at a syllable level. As both RQs asked whether each nonword set was able to capture differences between TD and DLD groups, whilst minimizing differences between L1 and L2 groups, the same independent variables and random effects were added to the two models. Independent variables included Group (MonDLD vs. MonTD vs. L2-TD), Nonword Set (High-Lexicality vs. Low-Lexicality vs. High-Lexicality-Vowel-Matched vs. CL-NWR), and their interaction. Random effects included Participant and Nonword Item. The dependent variable of Model 1 was NWR accuracy at whole-nonword level (scored as a categorical variable, correct vs. incorrect repetition on each trial, i.e., each nonword); and for Model 2, NWR accuracy at syllable level (scored as a categorical variable, correct vs. incorrect repetition on each trial, i.e., each syllable within each nonword). Any significant interactions are interpreted through post-hoc analyses and plotting predicted probabilities of NWR accuracy by Nonword Set and Group using the function, ggpredict() in the ggeffects package (Lüdecke, 2018) in R. This function computes predicted values (i.e., predicted probabilities of scoring successfully on whole nonwords/syllables) for all possible levels of a model's predictors (i.e., for each participant group and each nonword set). In other words, as the models analysed NWR accuracy as a categorical variable (correct vs. incorrect), the predicted probabilities in the figures describe the probability that a group of children (MonTD or MonDLD or L2-TD) would score successfully on the NWR test.

3. Results

3.1. NWR accuracy at whole-nonword level

Model 1 addressed RQ1, by examining the effects of participant group, nonword set, and their interaction on NWR accuracy at whole-nonword level. Table 1 shows the descriptive statistics for NWR accuracy (in percentages) according to participant group and nonword set, for whole-item and syllable level scoring.

The fixed effects of Model 1 are shown in Table 2. When NWR was scored at whole-nonword level, there was a significant main effect of Group, where MonTD children performed significantly better than MonDLD children ($p < 0.001$), but there was no significant difference in NWR accuracy between MonDLD and L2-TD children ($p = 0.15$). There was also a significant main effect of Nonword Set, where children scored significantly lower on Low-Lexicality nonwords compared to High-Lexicality-Vowel-Matched nonwords ($p < 0.001$), and there were no significant differences between performance on High-Lexicality and High-Lexicality-Vowel-Matched nonwords ($p = 0.92$). Children also scored significantly higher on CL-NWR nonwords than High-Lexicality-Vowel-Matched nonwords ($p = 0.01$), despite High-Lexicality-Vowel-Matched nonwords having a higher lexicality level, likely because CL-NWR have only simple CV syllables while Language-Specific nonwords have simple CV syllables and more complex CVC syllables by design.

Table 1
Descriptive statistics showing mean % correct, SD and range, according to group and nonword set, for whole-item and syllable-level scoring.

Scoring	Nonword Set	MonDLD	MonTD	L2-TD
Whole-item	High-Lexicality	61.7 (19.8) 20.8–87.5	82.9 (11.9) 58.3–100	74.8 (10.5) 50.0–95.8
	High-Lexicality-Vowel-Matched	60.5 (17.2) 16.7–83.3	82.7 (8.9) 66.7–95.8	68.2 (15.0) 25.0–83.3
	Low-Lexicality	26.1 (12.2) 4.2–50.0	41.4 (22.4) 16.7–87.5	32.0 (10.3) 12.5–54.2
	CL-NWR	79.6 (12.3) 56.3–100	85.5 (10.2) 62.5–100	84.6 (15.5) 43.8–100
Syllable	High-Lexicality	80.4 (12.5) 58.3–97.6	92.6 (5.5) 81.0–100	88.1 (5.5) 76.2–98.8
	High-Lexicality-Vowel-Matched	81.4 (13.2) 44.0–95.2	94.0 (4.0) 86.9–100	87.0 (8.0) 59.5–94
	Low-Lexicality	64.5 (12.4) 36.9–82.1	76.5 (10.6) 58.3–96.4	71.9 (8.2) 53.6–84.5
	CL-NWR	89.1 (7.4) 73.2–100	94.4 (5.2) 78.6–100	94.3 (5.1) 83.9–100

There was also a significant interaction between Group and Nonword Set, when NWR was scored at whole-nonword level. The interpretation of the interaction was assisted by plotting predicted probabilities of NWR accuracy at whole-nonword level, by Nonword Set, for each participant group separately (see Fig. 1). Fig. 1 shows that within each Nonword Set, MonTD children were predicted to have the highest NWR accuracy, followed by L2-TD children, then the MonDLD group; but the degree of group differentiation varied across the nonword sets. In terms of group differences between MonTD and MonDLD groups, all language-specific nonword sets captured substantial group differences, but CL-NWR nonwords did not, with the MonDLD group predicted to achieve similarly high levels of NWR accuracy as the MonTD group. Focussing on group differences between L2-TD and MonDLD groups, High-Lexicality Nonwords appeared to be the most effective in capturing such differences, while there were large overlaps in predicted probabilities of NWR accuracy in L2-TD and MonDLD groups on all remaining nonword sets, particularly when the 95 % CI were taken into consideration. Regarding group differences between MonTD and L2-TD groups, there was a close to complete overlap on predicted performance of the two groups on CL-NWR nonwords, suggesting that CL-NWR may be the most effective in minimizing disadvantages for L2-TD children, relative to MonTD children. There was some degree of disadvantage for L2-TD children, relative to the MonTD children, on all remaining nonword sets, which was especially prominent on High-Lexicality-Vowel-Matched nonwords.

Post-hoc pairwise comparisons were also conducted to examine the levels of group differentiation within each nonword set separately (see Table 3). Results showed there were significant group differences between the MonDLD and MonTD groups in High-Lexicality, High-Lexicality-Vowel-Matched and Low-Lexicality nonwords ($p \leq 0.008$) with small to medium effect sizes, but CL-NWR nonwords did not yield

Table 2
Fixed effects of Model 1 for analysis of NWR accuracy at whole-nonword level in MonDLD, MonTD and L2-TD groups.

Fixed Effect	β	*SE*	*z*	*p*
(Intercept)	0.64	0.40	1.61	0.11
Group (MonTD)	1.66	0.34	4.93	<0.001***
Group (L2-TD)	0.47	0.32	1.45	0.15
Nonword_Set (CL-NWR)	1.38	0.56	2.48	0.01*
Nonword_Set (Low-Lexicality)	−2.32	0.50	−4.66	<0.001***
Nonword_Set (High-Lexicality)	−0.05	0.49	−0.11	0.92
Group (L2-TD): Nonword_Set (CL-NWR)	−0.06	0.30	−0.19	0.85
Group (MonTD): Nonword_Set (CL-NWR)	−1.05	0.31	−3.37	<0.001***
Group (L2-TD): Nonword_Set (Low-Lexicality)	0.06	0.26	0.24	0.81
Group (MonTD): Nonword_Set (Low-Lexicality)	−0.43	0.27	−1.59	0.11
Group (L2-TD): Nonword_Set (High-Lexicality)	0.53	0.25	2.17	0.03*
Group (MonTD): Nonword_Set (High-Lexicality)	0.17	0.27	0.61	0.54

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; MonTD>MonDLD; L2-TD=MonDLD; CL-NWR>High-Lexicality-Vowel-Matched; Low-Lexicality < High-Lexicality-Vowel-Matched; High-Lexicality = High-Lexicality-Vowel-Matched; Significant interaction between Group and Nonword Set (see Fig. 1 and texts to assist interpretation).

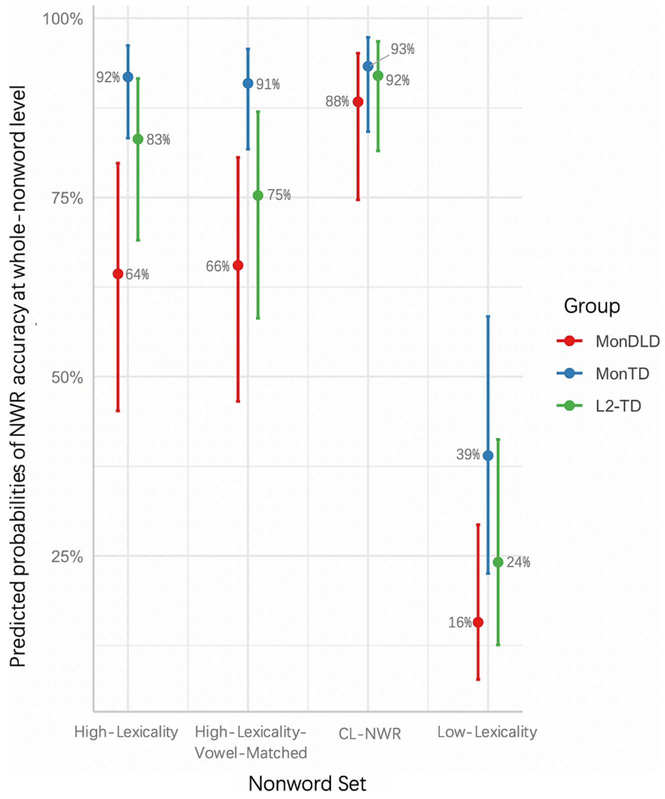

Fig. 1. Predicted probabilities of NWR accuracy at whole-nonword-level in the MonDLD, MonTD and L2-TD groups. Note. Circles represent mean predictive probabilities, and the lines represent 95% confidence intervals.

Table 3
P-values (and effect sizes in odds ratios) for pairwise comparisons between predicted group means of NWR accuracy at whole-nonword level and syllable level in MonDLD, MonTD & L2-TD children, for each nonword set.

Scoring	Nonword Set	MonDLD & MonTD	MonDLD & L2-TD	MonTD & L2-TD
Whole-item	High-Lexicality-Vowel-Matched	<0.001*** (4.27, medium effect)	0.23	0.001*** (0.37, small effect)
	CL-NWR	0.32	0.34	1.00
	High-Lexicality	**<0.001*** (5.80, medium effect)**	**0.02* (2.59, small effect)**	**0.08**
	Low-Lexicality	0.008** (3.42, small effect)	0.36	0.23
Syllable	High-Lexicality-Vowel-Matched	<0.001*** (3.57, medium effect)	0.22	0.001** (0.41, small effect)
	CL-NWR	**0.03* (2.26, small effect)**	**0.03* (2.26, small effect)**	**1.00**
	High-Lexicality	<0.001*** (3.76, medium effect)	0.09	0.03* (0.47, small effect)
	Low-Lexicality	<0.001*** (2.02, small effect)	0.07	0.25

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; Effect size interpretations are based on Chen et al. (2010), who suggested that OR of 1.68, 3.47 and 6.71 are equivalent to Cohen's *d* of 0.2, 0.5 and 0.8 respectively. Bold font indicates Lexicality Levels where the DLD group differs significantly from both TD groups which do not differ significantly from each other.

significant monolingual TD/DLD group differences. Regarding group differences between L2-TD and MonDLD groups, only High-Lexicality nonwords yielded significant group differences ($p = 0.02$), while other nonword sets did not. Comparing MonTD and L2-TD groups, while CL-NWR, High-Lexicality and Low-Lexicality nonwords did not capture significant group differences, demonstrating no disadvantage to the L2-TD children, High-Lexicality-Vowel-Matched nonwords yielded significant group differences, suggesting a disadvantage for L2-TD children. Taken together, High-Lexicality nonwords appeared to be the best at capturing TD/DLD group differences between both MonTD and MonDLD, as well as L2-TD and MonDLD groups, without disadvantaging L2-TD children compared to MonTD children, when NWR was scored at whole-item level.

3.2. NWR accuracy at syllable level

Model 2 addressed RQ2, by examining the effects of participant group, nonword set, and their interaction on NWR accuracy at syllable level. Descriptive statistics for syllable level scoring of NWR accuracy are shown in Table 1 above.

The fixed effects of Model 2 are shown in Table 4. On syllable-level scoring, NWR accuracy was significantly better in the MonTD group, compared to the MonDLD group ($p < 0.001$), and in the L2-TD group compared to the MonDLD group ($p = 0.003$). The main effect of Nonword Set was also significant, with better performance on syllables within CL-NWR items relative to syllables within High-Lexicality items ($p = 0.008$), and lower NWR accuracy on syllables within Low-Lexicality items compared to syllables within High-Lexicality items ($p < 0.001$); there was no difference between performance on syllables within High-Lexicality and High-Lexicality-Vowel-Matched items ($p = 0.92$), which shared the same level of lexicality.

There was also a significant interaction between Group and Nonword Set when NWR was scored at syllable level. The interpretation of the interaction was assisted by plotting predicted probabilities of NWR accuracy at syllable level, by Nonword Set, for each participant group separately (see Fig. 2). Fig. 2 shows that among the four sets of nonwords, CL-NWR nonwords were the only nonword set that was able to effectively minimize group differences between MonTD and L2-TD groups, whilst also capturing higher NWR accuracy in the two TD groups compared to the MonDLD group. On the remaining sets of nonwords, the MonDLD group was consistently predicted to score lower

Table 4
Fixed effects of Model 2 for analysis of NWR accuracy at syllable-level accuracy in MonDLD, MonTD and L2-TD groups.

Fixed Effect	β	*SE*	*z*	*p*
(Intercept)	1.99	0.25	7.99	<0.001***
Group (MonTD)	1.38	0.23	5.91	<0.001***
Group (L2-TD)	0.66	0.22	2.93	0.003**
Nonword_Set (High-Lexicality-Vowel-Matched)	−0.03	0.29	−0.11	0.92
Nonword_Set (CL-NWR)	0.89	0.34	2.65	0.008**
Nonword_Set (Low-Lexicality)	−1.17	0.29	−4.10	<0.001***
Group (MonTD): Nonword_Set (High-Lexicality-Vowel-Matched)	0.01	0.18	0.08	0.94
Group (L2-TD): Nonword_Set (High-Lexicality-Vowel-Matched)	−0.24	0.15	−1.61	0.11
Group (MonTD): Nonword_Set (CL-NWR)	−0.56	0.22	−2.60	0.009**
Group (L2-TD): Nonword_Set (CL-NWR)	0.08	0.21	0.39	0.70
Group (MonTD): Nonword_Set (Low-Lexicality)	−0.66	0.15	−4.32	<0.001***
Group (L2-TD): Nonword_Set (Low-Lexicality)	−0.26	0.14	−1.87	0.06

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; MonTD>MonDLD; L2-TD>MonDLD; CL-NWR>High-Lexicality; Low-Lexicality < High-Lexicality; High-Lexicality-Vowel-Matched = High-Lexicality; Significant interaction between Group and Nonword Set (see Fig. 2 and texts to assist interpretations).

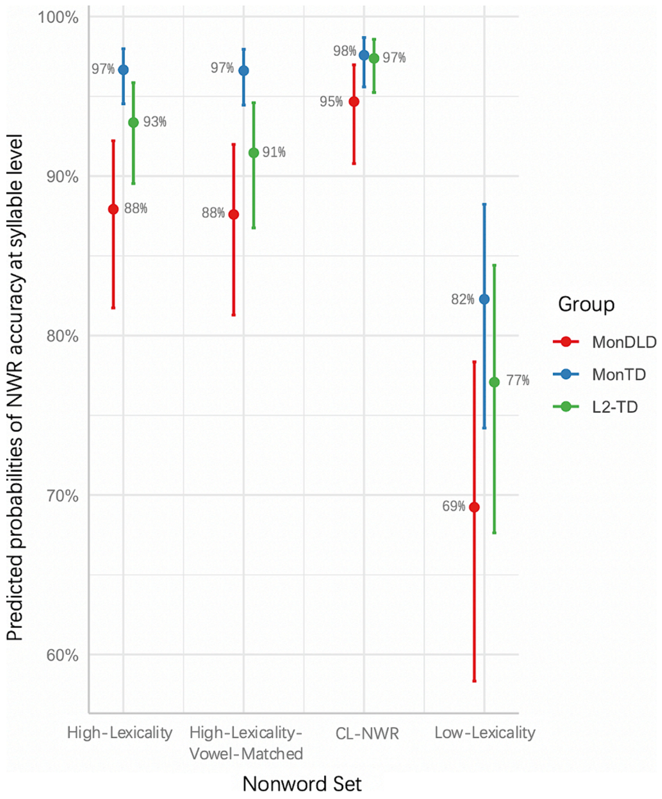

Fig. 2. Predicted probabilities of NWR accuracy at syllable-level in the MonDLD, MonTD and L2-TD groups. Note. Circles represent mean predictive probabilities, and the lines represent 95% confidence intervals.

than the MonTD group, and the L2-TD group was predicted to score between the two monolingual groups – on High-Lexicality and Low-Lexicality nonwords, there was substantial overlap in the predicted performance of the L2-TD group and both MonTD and MonDLD groups, but on High-Lexicality-Vowel-Matched nonwords, the predicted performance of the L2-TD group was noticeably below that of the MonTD group, and overlapped greatly with the MonDLD group.

Post-hoc pairwise comparisons were also conducted to examine the levels of group differentiation within each nonword set (see Table 3). Table 3 shows that consistent with previous findings, all language-specific nonwords (i.e. High-Lexicality, Low-Lexicality and High-Lexicality-Vowel-Matched nonwords) were able to generate statistically significant differences between MonDLD and MonTD groups ($p < 0.001$), with MonDLD scoring lower than MonTD. In addition, when scored at syllable level, Cantonese CL-NWR nonwords generated significant group differences between MonDLD and MonTD groups ($p = 0.03$) in the expected direction. Regarding the group difference between the MonDLD and L2-TD groups, CL-NWR nonwords were the only stimulus set that successfully captured significant differences between the two groups ($p = 0.03$), while the other nonword sets did not. Furthermore, focussing on the difference in predicted NWR accuracy between the MonTD and L2-TD groups, no significant group differences were captured on CL-NWR and Low-Lexicality nonwords – in particular, on CL-NWR nonwords, the p-value of 1.00 indicated that the L2-TD children performed as accurately as MonTD children, suggesting that they were not disadvantaged despite their L2 and bilingual status. Overall, when NWR was scored at syllable level, Cantonese CL-NWR nonwords appeared to be the best at capturing TD/DLD group differences between both MonTD and MonDLD, as well as L2-TD and MonDLD groups, without disadvantaging L2-TD children compared to MonTD children.

4. Discussion

This study explored the potential of Cantonese language-specific and quasi-universal NWR stimuli in capturing significant TD/DLD group differences, even for L2-Cantonese TD children with reduced language experience, using two scoring approaches. Specifically, it investigated whether Cantonese language-specific nonwords (High-Lexicality, Low-Lexicality, High-lexicality-Vowel-Matched nonwords) and cross-linguistic nonwords can capture significant group differences between L2-TD and MonDLD children, as well as between MonTD and MonDLD children, while minimizing group differences between MonTD and L2-TD children, when NWR is scored on (1) whole-item accuracy (2) syllable accuracy.

4.1. NWR performance in MonDLD, MonTD and L2-TD groups

This study is the first to document that Cantonese NWR stimuli have potential to capture significant TD/DLD group differences, even for L2-Cantonese TD children with reduced language experience. As established earlier, the ideal scenario for using NWR to assess bilingual L2 children is if nonwords maximise differences between TD and DLD groups, regardless of monolingual or bilingual status, while simultaneously minimizing the gap between monolingual and bilingual TD groups, despite bilingual L2 learners having weaker lexical and sub-lexical representations in their L2 to support NWR than monolingual children due to reduced language exposure and proficiency associated with bilingual language acquisition. Our results showed that at both whole-nonword and syllable levels of scoring, most of our sets of Cantonese nonwords yielded significant TD/DLD group differences among monolingual children, and certain sets of NWR stimuli were also able to capture significant differences between L2-TD and MonDLD groups, whilst minimizing differences between MonTD and L2-TD children. Therefore, our data are the first to provide evidence supporting the development of Cantonese NWR tests into assessment tools for DLD in bilingual L2 learners of Cantonese. Findings also bear on the type of nonwords that may be optimal for this purpose.

4.2. Optimal NWR tasks for assessing bilingual children in their L2

As certain nonword sets were found to be better suited than others for assessing bilingual L2-Cantonese-speaking children, in that they did not disadvantage L2-TD children, we first focus our discussion on these sets of nonwords.

Cantonese-adapted, CL-NWR nonwords were one of the stimulus sets that was able to capture significant group differences between MonTD and MonDLD and between L2-TD and MonDLD groups, while avoiding differences between L2-TD and MonDLD children, when scored at syllable level. This demonstrates for the first time that CL-NWR nonwords, when adapted to Cantonese, can effectively capture TD/DLD group differences among monolingual-Cantonese-speaking children, as well as between L2-TD and MonDLD children, adding to findings on the Dutch CL-NWR test (using whole-item scoring; Boerma et al., 2015) with evidence from a typologically distinct language. Having said this, it is also relevant to note that, similar to findings on the Swedish CL-NWR task (Öberg & Bohnacker, 2022), the present data also indicated substantial overlap in performance between the MonDLD and MonTD groups, and between the MonDLD and L2-TD groups (see Fig. 2). Despite significant TD/DLD differences yielded at the group level, effect sizes remained small, and the overlap in performance between groups indicated that the clinical accuracy of the Cantonese CL-NWR task may be lower at an individual level. This may be related to the older age of participants within this study, more than how well this set of nonwords could work for Cantonese speakers, as evidenced by a near-ceiling effect on Cantonese CL-NWR test at syllable-level scoring for all groups of children (see Table 1), indicating that even children with a DLD diagnosis did not find the task challenging. These findings are also in line with previous data suggesting that at above six years of age, the magnitude of TD/DLD group differences and clinical accuracy of the CL-NWR task reduces, even though TD/DLD group differences remain statistically significant (Boerma & Blom, 2021). Based on the current significant findings at group level, our data suggest that quasi-universal, CL-NWR nonwords, have potential to be further developed into informative assessment tools for L2 Cantonese-speaking children, especially for children below eight years of age.

While significant L2-TD/MonDLD group differences were only observed with syllable-level scoring in our study (see later section on scoring approaches for a more thorough discussion), it nevertheless provides further evidence supporting the conclusion from our previous study (Fu et al., 2024a) that Cantonese is not a true cross-linguistic exception in NWR, and for the potential utility of CL-NWR nonwords in assessing monolingual and bilingual children for DLD in a Cantonese context. Moreover, the finding that MonTD and L2-TD groups achieved close to the exact same levels of NWR accuracy in the present study demonstrates that CL-NWR nonwords truly do not disadvantage L2-TD children, even though they have less experience with the testing language. This is presumably because CL-NWR nonwords, which only include cross-linguistically frequent consonants and vowels arranged into simple syllable structures with limited prosodic structure, allow for bilingual TD children to use their linguistic knowledge from any and all languages they are acquiring to support redintegration during NWR. In fact, all consonants and vowels in the Cantonese-adapted version of CL-NWR test are also present in the Urdu phonemic inventory (Ambreen & To, 2024), meaning that L2-TD children could also draw on their lexical and sub-lexical knowledge of Urdu (i.e. their L1) when repeating Cantonese CL-NWR nonwords.

Cantonese language-specific, High-Lexicality nonwords were the second stimulus set found not to disadvantage L2-TD children, in this case, when NWR was scored at whole-nonword level. High-Lexicality nonwords were also able to capture significant group differences between MonTD and MonDLD groups, as previously demonstrated in Fu et al. (2024), and between L2-TD and MonDLD groups, albeit with small effect sizes. Indeed, High-Lexicality nonwords captured larger group differences between MonTD and MonDLD groups, and between L2-TD and MonDLD groups, compared to CL-NWR nonwords, suggesting that they are optimal for maximising differences between TD children and those with DLD regardless of the monolingual or bilingual status of the TD children. This is presumably because nonwords with higher levels of lexicality and sub-lexicality allowed monolingual and bilingual L2-TD children, who have better access to lexical and sub-lexical representations than those with DLD, to draw on their long-term linguistic knowledge in the redintegration process during NWR. L2-TD children were also not disadvantaged by High-Lexicality nonwords relative to MonTD children, suggesting that this set of High-Lexicality nonwords allowed L2-TD children to benefit from their lexical and sub-lexical representations in one or both languages (Urdu and Cantonese in this case) to support NWR. Together, the findings suggest that a combination of High-Lexicality nonwords and Cantonese CL-NWR nonwords might be most effective in capturing group differences between TD children, monolingual or bilingual, and monolingual children with DLD.

Interestingly, High-Lexicality-Vowel-Matched nonwords, which were of the same level of lexicality as High-Lexicality nonwords by design, did not generate the same pattern of group differences – at both whole-item and syllable levels of scoring, there was substantial overlap between performance by L2-TD and MonDLD groups, yielding non-significant differences between these two groups, while L2-TD children scored significantly lower than MonTD children. These findings suggested that some nonwords with high lexicality levels, like High-Lexicality-Vowel-Matched nonwords, may still disadvantage L2-TD learners, presumably because there are factors other than lexicality that affect NWR performance, especially in L2-TD children. Notably, studies on NWR in monolingual children have demonstrated that sub-lexical representations also support NWR, where children repeat nonwords more accurately when nonwords had higher levels of phonotactic probability (McKean et al., 2013; Szewczyk et al., 2018) and neighbourhood density (Fu et al., 2024b). If L2 Cantonese-speaking TD children also draw on their sub-lexical representations to support NWR, the difference in their performance on High-Lexicality and High-Lexicality-Vowel-Matched nonwords could perhaps be explained by differences in sub-lexical factors, such as phonotactic probability and neighbourhood density, between the two sets of nonwords (despite them sharing equally high levels of lexicality in terms of morphemicity); future studies could examine how these sub-lexical factors affect NWR in L2-TD learners of Cantonese.

In contrast, Low-Lexicality nonwords were not effective for capturing significant TD/DLD group differences for L2-Cantonese TD children with reduced language experience, given that these nonwords did not yield significant group differences between MonDLD and L2-TD groups at either whole-item or syllable levels of scoring. Although the L2-TD group did not differ from the MonTD group either, the lack of significant group difference between L2-TD and MonDLD groups suggested that at least some children from the L2-TD group were disproportionally challenged by Low-Lexicality nonwords compared to MonTD children. Even though Low-Lexicality nonwords should theoretically be equally challenging to monolingual and bilingual TD children in terms of their non-morphemic status, L2-TD children may be further taxed by language-specific elements in the present stimulus set. For example, Low-Lexicality nonwords included the Cantonese initial velar consonant /ŋ-/, initial rounded labial-velar approximant /w/, and final unreleased stop consonants /-p/, /-t/ and /-k/, all of which occur in Cantonese but not in Urdu (Ambreen & To, 2024). Therefore, at least some L2 learners could still be disproportionally disadvantaged due to reduced experience of Cantonese as a L2 and lack of support from L1. Additionally, our plotted predicted probabilities of NWR accuracy also indicated wide 95 % confidence intervals on Low-Lexicality nonwords, demonstrating great within-group variability and substantial overlap in NWR performance across groups evident in Figs. 1 and 2.

4.3. Scoring of NWR accuracy at Whole-Nonword level vs. Syllable level

The present data suggested that when using Cantonese NWR stimuli to assess Cantonese-speaking MonTD, MonDLD and L2-TD groups, scoring at both whole-nonword and syllable levels was able to maximise TD/DLD group differences while minimizing monolingual/bilingual TD group differences depending on the set of nonword stimuli. As NWR accuracy at whole-nonword level is already commonly adopted in NWR studies and has been demonstrated to differentiate between TD and DLD groups in both monolingual and bilingual children (Schwob et al., 2021), we focus our discussion on the less-used, syllable-level scoring approach. The different patterns of findings on the two scoring approaches may be related to the level of detail captured by each. NWR accuracy at syllable level could be seen as a more lenient level of scoring, as children are still able to score when they correctly repeat only certain components of a nonword, instead of being penalised as soon as one mistake has been made within a nonword, which would be the case when NWR is scored at whole-nonword level. Our finding that Cantonese CL-NWR nonwords only differentiated between MonTD and MonDLD groups and between MonDLD and L2-TD groups when syllable-level scoring was adopted demonstrated the phenomenon that while MonDLD children were repeating whole nonwords at a similar level of accuracy to both TD groups, both TD groups accurately repeated more components within each nonword compared to MonDLD children. Such nuanced differences across the groups could only be captured by a more detailed level of scoring, such as syllable-level accuracy. In addition to its benefits as a more fine-grained measure of NWR, scoring NWR performance at syllable level is also relatively quick and straightforward compared to even more fine-grained measures documented in the literature, such as scoring NWR in percentage of phonemes correct (PPC), suggesting that future studies might usefully explore this method of NWR scoring in other language versions of NWR tests.

4.4. *Limitations and future directions*

Whilst being the first study to report that Cantonese NWR stimuli are capable of generating group differences between MonDLD and L2-TD children, and simultaneously minimizing group differences between monolingual and bilingual L2-TD children in a Cantonese context, this study represents only the first steps in research on developing NWR as an assessment tool helping to identify DLD in both monolingual and bilingual Cantonese-speaking children.

First, this study did not include a bilingual DLD group, thus it is yet to be determined how Cantonese-speaking children with both reduced language experience (L2) and impaired language learning capacity (DLD) perform in NWR compared to other groups of children. Future studies will be in a better position to examine NWR performance in an L2-DLD group, when guidelines and methods for identifying DLD in L2 Cantonese children are better established.

Second, regarding participant sample, it would be beneficial to match gender ratios across groups and increase the sample size, particularly that of the bilingual L2-TD group, given the substantial heterogeneity in children acquiring multiple languages. We also note that, unlike the monolingual groups, non-verbal intelligence scores were not obtained for the L2-TD group in the present study, thus whether differences in non-verbal intelligence contributed to any group differences (or lack thereof) requires future verification. With that said, different studies have documented NWR performance to be largely independent of non-verbal intelligence scores (Boerma & Blom, 2021; Szewczyk et al., 2018; Weismer et al., 2000), thus it is unlikely that significant differences would arise in the pattern of findings when non-verbal intelligence scores are taken into consideration. In addition, future work is required to investigate whether NWR performance in Cantonese-speaking children and patterns of group differentiation are influenced by SES. Furthermore, given the ceiling effects observed in the present CL-NWR findings at syllable-level scoring, the older age of the participants is also a limitation of this study, and future work is required to examine how patterns of group differentiation may differ when younger children are studied.

Third, the current findings may be specific to L1-Urdu-L2-Cantonese-speaking children residing in Hong Kong. Whether these Cantonese NWR stimuli have potential to be developed into assessment tools for assessing bilingual Cantonese-speaking children acquiring languages other than Urdu awaits further investigation. We would expect the findings on CL-NWR nonwords to be more generalizable to children acquiring other L1s, as CL-NWR nonwords are designed to minimize the potential influence of language-specific knowledge. As well as evaluating this expectation, future studies could examine whether the present findings on language-specific High-Lexicality nonwords generalize to L2-TD Cantonese-speaking children with other L1s. It would also be worthwhile to investigate whether our findings on bilingual L2 Cantonese children generalize to bilingual L1 Cantonese children with and without DLD who are developing their first language under heavy influence from another language (e.g. children who acquire Cantonese as their first, heritage and minority language in countries having another language as the majority community language), providing further evidence on the diagnostic potential of our Cantonese NWR stimuli.

Finally, the present study only addressed the ability of Cantonese NWR stimuli to capture differences between MonTD and MonDLD and between L2-TD and MonDLD at a group level, and small effect sizes were yielded. For Cantonese NWR to be developed into an assessment tool with diagnostic value, further research is needed to determine how accurately Cantonese NWR classifies individual children into TD and DLD groups, regardless of monolingual or bilingual status, by investigating sensitivity and specificity of the NWR stimuli in a Cantonese context.

4.5. *Conclusions*

This study investigated the potential of Cantonese NWR stimuli not to disadvantage bilingual L2 Cantonese TD children with reduced target language experience, by examining their ability to capture significant group differences between DLD and L2-TD groups, while minimizing group differences between MonTD and L2-TD groups. When NWR accuracy was scored at whole-nonword level, High-Lexicality nonwords were best at capturing group differences between DLD and TD groups (both monolingual and L2), while not disadvantaging the L2-TD group relative to MonTD children. When NWR accuracy was scored at syllable level, the Cantonese version of CL-NWR was the only set of nonwords that did not disadvantage L2-TD children relative to monolingual TD children, while still being able to generate significant group differences between the monolingual DLD and L2-TD groups. These findings suggest that a combination of both language-specific, High-Lexicality nonwords, and quasi-universal, Cantonese CL-NWR nonwords can yield significant group differences between monolingual DLD and L2-TD groups, despite both groups having reduced language knowledge in Cantonese to support NWR, with the former being affected by impaired language learning capabilities and the latter being affected by reduced input conditions when acquiring more than one language. These findings also contribute evidence, at a group level, that the quasi-universal, CL-NWR test is able to capture significant TD/DLD group differences, even for L2-Cantonese TDs with reduced language experience, from a typologically distinct and understudied language. At an individual level, Cantonese CL-NWR test may have less than ideal clinical accuracy, though this may reflect the ceiling effects in this older age group studied more than how well this set of nonwords could work for Cantonese speakers, and may improve in children below eight years of age, a possibility that requires future evaluation.

Future studies should also explore whether the present findings generalize to L2-Cantonese-speaking children acquiring languages other than Urdu, bilingual children acquiring Cantonese as L1 under heavy influence from another language, and examine sensitivity and specificity of Cantonese NWR in classifying individual children into TD and DLD groups for both monolingual and bilingual populations. Moreover, future studies examining participant-related factors that tap into individual differences in cognitive and linguistic abilities that support NWR performance and how these cognitive and linguistic foundational abilities predict children's NWR performance might shed light on why some children with DLD found NWR significantly more challenging than their TD age peers, while other children with DLD overlapped in performance with TD age peers, as reported in Öberg & Bohnacker (2022). Overall, the findings of the present study bring Cantonese NWR research in line with international research on NWR, supporting the potential of NWR as a clinical marker of DLD to be included in language assessment crosslinguistically.

5. Ethics statement

This study was carried out in accordance with the recommendations of the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University (reference number: HSEARS20161230004). Written informed consent was also obtained from the parents of each participant.

Funding

This research was supported by two research grants (G-UAK5 and 4-ZZR3; Principal Investigator: Chan; Co-Investigators: Chen, Polišenská, Chiat), awarded by the Hong Kong Polytechnic University; and a research grant titled "Revisiting Nonword Repetition as a diagnostic tool in Cantonese-speaking children with Developmental Language Disorder" (GRF PolyU 15609420; Principal Investigator: Chan, Co-Investigators: Polišenská, Chiat and Kan), awarded by The Hong Kong

Research Grants Council General Research Fund.

CRediT authorship contribution statement

Nga Ching Fu: Writing – review & editing, Writing – original draft, Project administration, Investigation, Data curation. **Angel Chan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Si Chen:** Writing – review & editing, Visualization, Validation, Supervision, Formal analysis. **Kamila Polišenská:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Shula Chiat:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Ashley Chan, Hazel Chau, Christy Chu, Saboor Hamdani, Circle Kai, Eva Kwan, Mars Kwan, Carmen Lai, Annie Lam, Maggie Lee, Cheri Li, Imran Muhammad, Sylvia Wan, Angel Wo, Jovial Wong, Katrina Wong, Joanne Yiu for help as (student) speech therapists and research assistants, and all the children and schools who took part in this study. We also thank The Hong Kong Polytechnic University and The Hong Kong Research Grants Council General Research Fund for their funding support (G-UAK5; 4-ZZR3; GRF PolyU 15609420).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2024.105450.

References

Agresti, A. (2002). *Categorical Data Analysis. Wiley Series in Probability and Statistics* (2nd ed). Hoboken, New Jersey: John Wiley & Sons Inc.

Ambreen, S., & To, K. S. C. (2024). Review of the phonological system of contemporary Urdu spoken in Pakistan. *International Journal of Speech-Language Pathology*. https://doi.org/10.1080/17549507.2024.2324905

Armon-Lotem, S. (2012). Introduction: Bilingual children with SLI – the nature of the problem. *Bilingualism: Language and Cognition, 15*(1), 1–4. https://doi.org/10.1017/S1366728911000599

Armon-Lotem, S. (2018). SLI in bilingual development: How do we approach assessment? In A. Bar-On, D. D. Diskin Ravid, & E. Dattner (Eds.), *Handbook of communication disorders* (pp. 617–641). De Gruyter Mouton. https://doi.org/10.1515/9781614514909-031.

Armon-Lotem, S., & Chiat, S. (2012). In *How do sequential bilingual children perform on non-word repetition tasks?* (pp. 53–62). Cascadilla Press.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes (R Package Version 0.999375-33). http://CRAN.R-project.org/package=lme4.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology: De Gruyter Mouton.*. https://doi.org/10.1515/9783110823707

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus Study. Identifying language impairments in children. *PLOS ONE, 11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Boerma, T., & Blom, E. (2021). Crosslinguistic nonword repetition and narrative performance over time: A longitudinal study on 5- to 8-year-old children with diverse language skills. In S. Armon-Lotem, & K. K. Grohmann (Eds.), *Trends in Language Acquisition Research* (Vol. 29, pp. 302–328). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.29.10boe.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research, 58*(6), 1747–1760. https://doi.org/10.1044/2015_JSLHR-L-15-0058

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *Advances in Speech Language Pathology, 9*(4), 312–322. https://doi.org/10.1080/14417040701624474

Chan, J. (1984). Raven's progressive matrices test in Hong Kong. *New Horizons: The Journal of Education, Hong Kong Teachers' Association, 25*, 43–49.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation, 39*(4), 860–864. https://doi.org/10.1080/03610911003650383

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children—Disentangling bilingualism from language impairment* (pp. 125–150). Multilingual Matters. https://www.degruyter.com/document/doi/10.21832/9781783093137-008/html.

Chiat, S., & Roy, P. (2007). The preschool repetition test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research, 50*, 429–443. https://doi.org/10.1044/1092-4388(2007/030)

Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Real-word and nonword repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 56*(1), 323–336. https://doi.org/10.1044/1092-4388(2012/11-0304)

Fu, N. C., Chen, S., Polišenská, K., Chan, A., Kan, R., & Chiat, S. (2024a). Nonword repetition in children with Developmental Language Disorder: Revisiting the case of Cantonese. *Journal of Speech, Language, and Hearing Research, 1–13*. https://doi.org/10.1044/2024_JSLHR-22-00397

Fu, N. C., Chan, A., Chen, S., Polišenská, K., & Chiat, S. (2024b). *Sublexical predictors of nonword repetition in Cantonese-speaking children. [Manuscript in preparation].* Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University.

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*(3), 336–360. https://doi.org/10.1016/0749-596X(90)90004-J

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), 196–206. https://doi.org/10.1044/0161-1461.3002.196

Guiberson, M., & Rodríguez, B. L. (2013). Classification accuracy of nonword repetition when used with preschool-age Spanish-speaking shildren. *Language, Speech, and Hearing Services in Schools, 44*(2), 121–132. https://doi.org/10.1044/0161-1461(2012/12-0009)

Hamdani, S. Z., Chan, A., Kan, R., Chiat, S., Gagarina, N., Haman, E., Łuniewska, M., Polišenská, K., & Armon-Lotem, S. (2024). Identifying developmental language disorder (DLD) in multilingual children: A case study tutorial. *International Journal of Speech-Language Pathology*, 1–15. https://doi.org/10.1080/17549507.2024.2326095

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.* New Jersey: John Wiley & Sons Inc.

Kohnert, K., Windsor, J., & Yim, D. (2006). Do language-based processing tasks separate children with language impairment from typical bilinguals? *Learning Disabilities Research and Practice, 21*(1), 19–29. https://doi.org/10.1111/j.1540-5826.2006.00204.x

le Clercq, C. M. P., van der Schroeff, M. P., Rispens, J. E., Ruytjens, L., Goedegebure, A., van Ingen, G., & Franken, M.-C. (2017). Shortened nonword repetition task (NWR-S): A simple, quick, and less expensive outcome to identify children with combined specific language and reading impairment. *Journal of Speech, Language, and Hearing Research, 60*(8), 2241–2248. https://doi.org/10.1044/2017_JSLHR-L-16-0060

Lee, S. A. S., & Gorman, B. K. (2013). Nonword repetition performance and related factors in children representing four linguistic groups. *International Journal of Bilingualism, 17*(4), 479–495. https://doi.org/10.1177/1367006912438303

Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software, 3*(26), 772. https://doi.org/10.21105/joss.00772

McGregor, K. K. (2009). Semantics in child language disorders. In *Handbook of child language disorders* (pp. 365–387). Psychology Press.

McKean, C., Letts, C., & Howard, D. (2013). Developmental change is key to understanding primary language impairment: The case of phonotactic probability and nonword repetition. *Journal of Speech, Language, and Hearing Research, 56*(5), 1579–1594. https://doi.org/10.1044/1092-4388(2013/12-0066)

Messer, M. H., Leseman, P. P. M., Boom, J., & Mayo, A. Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology, 105*(4), 306–323. https://doi.org/10.1016/j.jecp.2009.12.006

Montgomery, J. W. (2002). Understanding the language difficulties of children with Specific Language Impairments: Does verbal working memory matter? *American Journal of Speech-Language Pathology, 11*(1), 77. https://doi.org/10.1044/1058-0360(2002/009)

Öberg, L., & Bohnacker, U. (2022). Non-word repetition and vocabulary in Arabic-Swedish-speaking 4–7-year-olds with and without Developmental Language Disorder. *Languages, 7*(3), 204. https://doi.org/10.3390/languages7030204

Ortiz, J. A. (2021). Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy. *American Journal of Speech-Language Pathology, 30*(5), 2275–2295. https://doi.org/10.1044/2021_AJSLP-20-00237

Pham, G., & Ebert, K. D. (2020). Diagnostic accuracy of sentence repetition and nonword repetition for Developmental Language Disorder in Vietnamese. *Journal of Speech,* *Language, and Hearing Research, 63*(5), 1521–1536. https://doi.org/10.1044/2020_JSLHR-19-00366

Polišenská, K., & Kapalková, S. (2014). Improving child compliance on a computer administered nonword repetition task. *Journal of Speech, Language, and Hearing Research, 57*(3), 1060–1068. https://doi.org/10.1044/1092-4388(2013/13-0014)

Poon, A. Y. K. (2010). Language use, and language policy and planning in Hong Kong. *Current Issues in Language Planning, 11*(1), 1–66. https://doi.org/10.1080/14664201003682327

R Core Development Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Raven, J. C., Court, J. H., & Raven, J. (1996). *Raven's Standard Progressive Matrices*. Oxford Psychologists Press. https://doi.org/10.1007/978-1-4615-0153-4_11

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2011). Non-word repetition in adolescents with Specific Language Impairment and Autism plus Language Impairments: A qualitative analysis. *Journal of Communication Disorders, 44*(1), 23–36. https://doi.org/10.1016/j.jcomdis.2010.06.003

Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using nonword repetition to identify Developmental Language Disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 64*(9), 3578–3593. https://doi.org/10.1044/2021_JSLHR-20-00552

Sharp, K. M., & Gathercole, V. C. M. (2013). Can a novel word repetition task be a language-neutral assessment tool? Evidence from Welsh-English bilingual children. *Child Language Teaching and Therapy, 29*(1), 77–89. https://doi.org/10.1177/0265659012465208

Spieler, D., & Schumacher, E. (2020). *New Methods in Cognitive Psychology*. New York, NY: Routledge.

Stark, R. E., & Blackwell, P. B. (1997). Oral volitional movements in children with language impairments. *Child Neuropsychology, 3*(2), 81–97. https://doi.org/10.1080/09297049708401370

Stokes, S. F., Wong, A.-M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword Repetition and Sentence Repetition as Clinical Markers of Specific Language Impairment: The case of Cantonese. *Journal of Speech and Hearing Research, 49*(2), 219–236. https://doi.org/10.1044/1092-4388(2006/019)

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition, 179*, 23–36. https://doi.org/10.1016/j.cognition.2018.06.002

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 46*(1), 1–16. https://doi.org/10.1016/j.jcomdis.2012.08.002

To, C. K. S., Cheung, P. S. P., & McLeod, S. (2013). A population study of children's acquisition of Hong Kong Cantonese consonants, vowels, and tones. *Journal of Speech, Language, and Hearing Research, 56*(1), 103–122. https://doi.org/10.1044/1092-4388(2012/11-0080)

T'sou, B., Lee, T., Tung, P., Man, Y., Chan, A., To, C., & Chan, Y. (2006). *Hong Kong Cantonese Oral Language Assessment Scale*. City University of Hong Kong.

Tuller, L. (2015). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing Multilingual Children—Disentangling Bilingualism from Language Impairment* (pp. 301–330). Multilingual Matters. https://doi.org/10.21832/9781783093137-013.

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13*(4), 329–340. https://doi.org/10.1044/1058-0360(2004/033)

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research, 38*(6), 1349–1355. https://doi.org/10.1044/jshr.3806.1349

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43*(4), 865–878. https://doi.org/10.1044/jslhr.4304.865

Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology, 19*(4), 298–310. https://doi.org/10.1044/1058-0360(2010/09-0064)