Nonword repetition performance of Arabic-speaking children with and without Developmental Language Disorder: a study on diagnostic accuracy

Article

Accepted Version

To link to this article DOI: http://dx.doi.org/10.1044/2021_JSLHR-20-00556

Publisher: ASHA

Nonword repetition performance of Arabic-speaking children with and without Developmental Language Disorder:

A study on diagnostic Accuracy

Juhayna Taha[1], Vesna Stojanovik[1] and Emma Pagnamenta[1]

Keywords: Developmental Language Disorder, Specific Language Impairment, nonword repetition, Arabic, cross-linguistic.

[1] School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

Address correspondence to Juhayna Taha, University of Reading, School of Psychology and Clinical Language Sciences, Early Gate Reading RG6 6AL, United Kingdom.

E-mail: j.taha@pgr.reading.ac.uk

Funding statement: This work was funded by [REMOVED FOR REVIEW] awarded to the first author.

Conflict of Interest Statement: The authors report no conflict of interest.

Abstract

Purpose: This study evaluates the effectiveness of a nonword repetition (NWR) task in discriminating between Palestinian Arabic-speaking children with Developmental Language Disorder (DLD) and age-matched typically-developing (TD) children.

Methods: Participants were 30 children with DLD aged between 4;00 and 6;10 and 60 TD children aged between 4;00 and 6;8 matched on chronological age. The Arabic version of a Quasi-Universal Nonword Repetition task was administered. The task comprises 30 nonwords that vary in length, presence of consonant clusters (CC) and wordlikeness ratings. Responses were scored using an item-level scoring method. To assess the diagnostic accuracy of the task. ROC curve analysis was conducted to determine the best cut-off point with the highest sensitivity and specificity values and likelihood ratios were calculated.

Results: Children with DLD scored significantly lower on the NWR task than their age-matched TD peers. Only the DLD group was influenced by the phonological complexity of the nonwords, with nonwords with two CC being more difficult than nonwords with no or only one CC. For both groups, three-syllable nonwords were repeated less accurately than two and one-syllable nonwords. Also, high word-like nonwords were repeated more accurately than nonwords with low wordlikeness ratings. The best cutoff score had sensitivity and specificity of 93% and highly informative likelihood ratios.

Conclusion: NWR was an area of difficulty for Palestinian Arabic-speaking children with DLD. NWR showed excellent discriminatory power in differentiating Arabic-speaking children diagnosed with DLD from their age-matched TD peers. NWR appears to hold promise for clinical use as it is a useful indicator of DLD in Arabic. These results need to be further validated using population-based studies.

Introduction

Developmental Language Disorder (DLD) affects approximately 7% of children at school entry (Norbury et al., 2016) and it refers to difficulties in understanding and/or using language without a known biomedical etiology. These difficulties interfere with everyday life, educational achievement, and are likely to persist into school age and beyond (Bishop et al., 2016, 2017). Given the negative impact of DLD on the quality of life of affected children, early identification of the disorder is imperative.

Clinical markers are tasks that can reliably capture the difficulties experienced by children with DLD and exclude those with typical language development. Therefore, these tasks play an important role in accurate identification and appropriate treatment of DLD. Cross-linguistic evidence shows that Nonword Repetition (NWR) may be a reliable clinical marker of DLD in monolingual and bilingual children speaking a variety of languages (for a review, see Chiat, 2015; Leonard, 2014). Our study aims to investigate NWR abilities of Palestinian Arabic-speaking children with DLD aged 4 to 6 years relative to chronological age-matched typically-developing (TD) peers. Importantly, the study will evaluate the diagnostic accuracy of NWR as a potential clinical marker of DLD in Arabic. Exploring the diagnostic accuracy will inform clinicians to what extent NWR can accurately distinguish Palestinian Arabic-speaking children with and without DLD. We begin with an overview of the cross-linguistic evidence for NWR deficits in children with DLD, followed by a review of the usefulness of NWR tasks as possible diagnostic markers of DLD, and factors that may influence performance on NWR tasks.

NWR deficits in children with DLD: cross-linguistic evidence

NWR tasks assess the ability to encode, temporarily store, retrieve and imitate an unfamiliar string of phonemes that conform to the phonotactics of the child's native language, yet lack any meaning. NWR resembles a crucial skill that underlies early word learning: children's ability to spontaneously repeat the new, unfamiliar words they hear. NWR has been reported to correlate with TD children's concurrent vocabulary size (e.g., Gathercole, 2006; Melby-Lervåg et al., 2012) and to predict vocabulary acquisition (e.g., Gathercole et al., 1997).

Studies have consistently reported that English-speaking children with DLD are significantly less accurate in repeating nonwords compared to their TD peers and that these group differences persist across development (for a review, see Graf Estes et al., 2007). The finding that NWR is impaired in children with DLD has been replicated in many languages, including Italian (Bortolini et al., 2006); Spanish (Girbau & Schwartz, 2007), French (Thordardottir et al., 2011), Dutch (Rispens & Parigger, 2010), Swedish (Kalnak et al., 2014), Slovak (Kapalková et al., 2013) and Turkish (Topbaş et al., 2014) among others.

In contrast, Cantonese-speaking children with DLD (age range: 4;2 to 5;7 years) have been reported to perform as well as age-matched TD children on a NWR task, suggesting that NWR is not a clinical marker of DLD in this language (Stokes et al., 2006). As the NWR task in Stokes et al.'s (2006) study was based on the phonotactic rules of Cantonese, these findings were attributed to the phonologically less complex nature of Cantonese compared with other languages. According to Stokes et al. (2006), Cantonese is a tonal language with a small phonemic inventory, basic syllabic structure (CV only), and only a limited set of syllabic combinations are allowed. Additionally, syllables in multisyllabic words are equally stressed (i.e., quite salient). Therefore, it could be that the nonwords used in Stokes et al. (2006) were not as complex as the nonwords used in other languages with more complex syllabic structures and stress variations (e.g., English). Notably, a subsequent study found that 5-year-old, Cantonese-speaking children with DLD scored below their age-matched TD controls on NWR (Wong et al., 2010). Although the between-group difference was only marginally significant (p = 0.06), Wong et al. (2010) argued that Cantonese-speaking children's weak performance on the nonword (and word) repetition tasks relative to age norms suggests that these children have an impairment in this domain. The contradictory results of the two Cantonese studies were attributed to differences in the NWR tasks and scoring methods (for discussion, see Wong et al., 2010). Recently, Pham and Ebert (2020) found that Vietnamese-speaking children with DLD performed poorly on NWR relative to same-age TD peers. In line with the results of Wong et al (2010)'s study and contrary to those of Stokes et al. (2006), Pham and Ebert (2020) found that NWR could discriminate between Vietnamese-speaking children with and without DLD which suggests NWR tasks may have potential in detecting DLD in Asian tonal languages.

Several studies have examined NWR abilities in Arabic speaking children with and without DLD: NWR has been reported as impaired in monolingual school-age children with DLD acquiring Qatari Arabic ($N = 11$, mean age = 7;8; Shaalan, 2010), Hijazi Arabic ($N = 52$, mean age = 8;4 years; Balilah, 2017), and in kindergarten ($N = 25$, mean age = 5;5) and first grade ($N = 25$, mean age = 6;11) children with DLD acquiring Palestinian Arabic (Saiegh-Haddad & Ghawi-Dakwar, 2017). NWR was also found to be problematic for preschool-age Qatari Arabic speaking children at risk of DLD ($N = 15$, age range: 2;3 to 3;11 years; Khater, 2016) and bilingual Arabic-French/English children with DLD ($N = 16$, mean age = 5;8 to 7;8 years; Abi-Aad & Atallah, 2012). The consistent group differences between Arabic-speaking children with DLD and same-age TD children indicate the potential of NWR in discriminating between clinical and non-clinical groups.

Factors influencing NWR performance

It is well documented that nonword length, i.e., the number of syllables, affects how accurately children repeat nonwords (e.g., Coady & Evans, 2008). TD children, as well as children with DLD, typically show accurate repetition of short nonwords (i.e. one and two syllables). As the nonwords increase in length (three or more syllables), the repetition accuracy decreases for both groups, particularly for children with DLD group (Archibald & Gathercole, 2006; Chiat & Roy, 2007; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990; Jones et al., 2010; Weismer et al., 2000). According to Chiat (2015), this length effect has been replicated in all languages studied to date.

Phonological complexity is another factor that influences NWR accuracy. Phonologically complex nonwords with consonant clusters are repeated less accurately than phonologically simple nonwords that only contain singleton consonants. Although articulatory complexity affects children with and without DLD (Edwards & Lahey, 1998; Gathercole & Baddeley, 1990), children with DLD are more adversely affected by the presence of consonant clusters relative to TD peers (Briscoe et al., 2001; Leclercq et al., 2013; Munson et al., 2005).

Although NWR is a processing-dependent measure, long-term language knowledge also plays a role. NWR accuracy appears to be influenced by two closely related factors: Wordlikeness (the extent to which a nonword resembles a real word based on native speakers' judgment) and phonotactic probability (an objective measure of the frequency of the occurrence of a specific sound or sound combination in a given language). Nonwords that sound like real words in a given language receive high ratings from adults as being word-like. Nonwords with high word-like ratings are repeated by children more accurately than nonwords that are rated as less word-like (Archibald & Gathercole, 2006; Briscoe et al., 2001; Coady et al., 2010; Gathercole, 2006; Gathercole & Baddeley, 1990; Munson et al., 2005). High word-like nonwords overlap with real lexical items in long-term memory, thus will be more easily repeated than nonwords with low word-likeness ratings (Bowey, 2001; Metsala, 1999; Snowling et al., 1991; Szewczyk et al., 2018). Furthermore, nonwords containing high phonotactic probability sequences are repeated more accurately than nonwords containing low phonotactic probability sequences (Munson et al., 2005). Some studies have found that wordlikeness and phonotactic probability have a larger effect on NWR accuracy of children with DLD relative to TD peers (Jones et al., 2010; Leclercq et al., 2013; Munson et al., 2005). For instance, Munson et al. (2005) reported that the difference in NWR accuracy between children with DLD and TD children was larger on items with low phonotactic ability than on those with high phonotactic probability. However, others have found no differences between children with and without DLD (Coady et al., 2010).

NWR as a clinical marker of DLD

The statistically reliable difference between children with and without DLD on NWR tasks is important. However, it does not inform us about its clinical usefulness for the identification of DLD. This requires determining its diagnostic accuracy. Diagnostic accuracy is indexed by measures of sensitivity, i.e., the proportion of children with a DLD diagnosis correctly identified by the task (true positive rate), and specificity, i.e., the proportion of children without a disorder correctly identified by the task (true negative rate). A threshold score should be set as a cutoff point for the analysis of sensitivity and specificity. The classification accuracy of a cutoff point with specificity and sensitivity values above 80% is considered acceptable, with values above 90% being excellent (Plante & Vance, 1994). Dollaghan and Campbell (1998) recommend also calculating positive Likelihood Ratio (LR+), i.e., the probability to be identified as impaired if impaired, and Negative Likelihood Ratio (LR-), i.e., the probability to be identified as unimpaired if unimpaired. Following the guidelines of Sackett et al. (1991), Dollaghan (2007) indicated that values of $LR+ \geq 10.0$ and $LR- \leq 0.1$ can be interpreted with high confidence to rule in or rule out the disorder, respectively, whereas values of $LR+ \geq 3.0$ and $LR+ \leq 0.3$ are suggestive but insufficient to rule in or rule out the disorder, respectively.

The findings of studies that have examined the use of NWR in distinguishing English-speaking children with DLD from TD children are inconclusive (for a full review, see Pawłowska, 2014). Using a cutoff score equal or less than 70% accuracy on the Nonword Repetition Test (NRT), Dollaghan and Campbell (1998) documented LR+ value of 25.15 for children aged 5 to 12 years. A total score of 70% or less on the NRT was 25 times more likely to come from a child with language impairment than from a TD child, suggesting that the NRT had a high degree of accuracy in differentiating children with and without language impairment. However, using the same cutoff point with 7 to 8-year-old children, Weismer et al. (2000) found the LR+ to be 2.78 indicating that the diagnostic accuracy of the NRT was "intermediate" and not sufficient to identify language impairment in this age group. Subsequent studies have reported high levels of sensitivity and specificity of NWR in identifying language impairments in preschool-age children (Deevy et al., 2010) and at age 7 (Redmond et al. 2011). Some studies have found lower levels of sensitivity acceptable levels of specificity of NWR in identifying DLD (Conti-Ramsden, 2003; Conti-Ramsden et al., 2001) while other studies have documented low values for sensitivity and specificity (Archibald & Joanisse, 2009; Poll et al., 2010). The discrepancy of results across studies from English-speaking populations may be due to the variability in reference standards used to identify children with DLD, the structure of NWR tasks used and their scoring methods (for a review, see Graf Estes et al., 2007; Pawłowska, 2014). While some studies have followed a one-gate design by recruiting unselected population samples (Poll et al., 2010; Weismer et al., 2000), others have followed a two-gate design by recruiting pre-selected TD and DLD groups (e.g., Conti-Ramsden & Hesketh, 2003; Conti-Ramsden et al., 2001; Deevy et al., 2010; Gray, 2003; Redmond et al., 2011). Pawlowska (2014) argued that one-gate studies include children with DLD across the ability spectrum, some of which could have borderline scores, whereas two-gate studies include children with a prior diagnosis of DLD who are likely to have severe language difficulties as they were enrolled for intervention. Hence, TD and DLD group differences in two-gate studies are likely to be larger than those of one-gate studies leading to variations in diagnostic accuracy levels. The diagnostic accuracy of NWR has also been examined in languages other than English (see Table 1 for a summary). Most studies have documented good sensitivity and specificity values of above 80%, showing the clinical value of NWR in distinguishing children with and without DLD across languages.

INSERT TABLE 1 HERE

Wallan (2018) examined the clinical utility of the adapted Verbal Short Term Memory test (VSTM) which included digit recall, word list recall and nonword list recall tasks. The nonword list recall was administered to a "language concern" group which included children whose parents/teachers had concerns about their language development (N = 14, age range 2;10 to 5;11 years) and a group of TD children matched on age and nonverbal IQ. The "language concern" group scored slightly lower than the TD group on the nonword list recall task. Wallan (2018) found that this task failed to distinguish between the two groups and attributed the poor diagnostic accuracy of the task to the limited range of scores in the TD children. However, the poor diagnostic accuracy of nonword list recall in Wallan (2018)'s study can also be explained in relation to the reference standard according to which children were placed in "language concern group". The sole reliance on parental/teachers' reports as an indicator of language status could mean that some of the children in the "language concern" group did not have language impairment of clinical significance.

In Arabic, previous studies have used group comparisons and revealed that, on average, Arabic-speaking children with DLD scored below their age-matched TD peers on NWR tasks (Abi-Aad & Atallah, 2012; Balilah, 2017; Khater, 2016; Saiegh-Haddad & Ghawi-Dakwar, 2017; Shaalan, 2010). However, group differences are not sufficient to conclude that poor NWR is a clinical marker of DLD in Arabic-speaking children, due to the high degree of variability in individual DLD profiles. Therefore, the extent which NWR can be an accurate indicator of the presence or absence of DLD in Arabic remains unclear. Exploring the diagnostic accuracy is thus necessary as it considers the individual differences among children with DLD. Examination of diagnostic accuracy can also determine the accuracy of NWR in differentiating between Arabic-speaking children with DLD from TD peers.

In Palestine, the identification of DLD is an ongoing challenge, as no standardized language assessments are available. As a result, Palestinian children with DLD are particularly vulnerable to being misdiagnosed or just missed altogether. Diagnostic tools are needed to facilitate effective and efficient identification of DLD in Arabic. In response to this issue, the present study attempts to provide SLTs with evidence of the potential of NWR as a screening measure. This, in turn, can help enhance the accuracy of assessment procedures when diagnosing DLD in Palestinian children.

Aims

Existing studies have provided important insights about the potential of NWR as a clinical marker of DLD in Arabic. However, information about the clinical usefulness of this measure is yet to be determined. In this study, the Arabic version of a Quasi-Universal Nonword Repetition test was employed to address the following questions:

1. How do children with DLD compare to age-matched TD children in terms of their NWR performance accuracy?
2. How accurate is NWR performance in distinguishing Palestinian Arabic-speaking children with DLD from their age-matched TD peers?

Methods

Participants

This study received ethical approval by the [REMOVED FOR REVIEW] Research Ethics Committee. There were 90 participants in two groups: a group of 60 TD children and 30 children with DLD. All participants were monolingual native speakers of Palestinian Arabic. According to parents' and teachers' reports, all participants had normal hearing, and no symptoms or history of neurological deficits, oral-motor impairments, or social-emotional/behavioral difficulties. See Table 2 for demographic information.

INSERT TABLE 2 HERE

The TD children (27 females and 33 males) aged between 4;00 and 6;8 years; months (M = 63.85 months, SD = 10.16 months) were recruited from three kindergartens in the same geographical area as the DLD group. Additional inclusionary criteria for this group were 1) age-appropriate language skills as reported by their caregivers 2) no history of speech-language therapy. The children with DLD (8 females and 22 males) aged between 4;00 and 6;10 years; months (M = 61.50 months, SD = 11.27 months) were recruited from five private speech therapy clinics in [REMOVED FOR REVIEW]. Each child in the TD group was within two months of age of a child in the DLD group. The two groups were matched on chronological age ($t(53.04) = -.96$, $p = .34$, $d = .22$).

All 30 children in the DLD group had been diagnosed with DLD by qualified speech and language therapists (SLTs) independent of this study and were receiving language intervention at the time of testing.The diagnosis of DLD in Palestine is made based on qualitative assessment supported by the clinical judgement of the SLTs. Therefore, it was crucial to ensure that the children with DLD met the criteria for DLD as set out by Bishop et al. (2016, 2017). A brief interview with each of the children's SLT was done to confirm that (1) their language disorder was not limited to expressive phonology, but also affected other language components such as semantics morpho-syntax and pragmatics among others, (2) their hearing was normal according to audiology reports, (3) and their language disorder was not associated with any biomedical conditions (e.g., neurological and genetic syndromes).

A weakness in expressive morpho-syntax is a hallmark of children with DLD (Leonard, 2014). Particularly, Arabic-speaking children with, or at risk of DLD, are known to have difficulties with Sentence Repetition (Shaalan, 2010; Wallan, 2018), the production of verb inflections (Abdallah & Crago, 2008; Fahim, 2017; Shaalan, 2010) and noun plurals (Abdallah et al., 2013; Shaalan, 2010). Accordingly, three non-standardized language tasks were administered to verify the language status of the TD children and to ascertain that the children with DLD had language skills that were considerably below those expected for their chronological age. These included the (a) ***Arabic Sentence Repetition Test (A-SRT)***: The task assesses the production of language-specific structures that are impaired in Arabic-speaking children with DLD and language-independent structures which are documented to be impaired in children with DLD across languages; (b) ***Arabic Verb Elicitation Test (AVET)*****:** a picture-naming task which examines the production of verb tense and agreement inflections; (c) ***Arabic Noun Pluralization Test (ANPT)*****:** a picture-naming task that examines the production of noun plural types. Additionally, we calculated (d) ***Mean Morpheme per Utterance (MPU*****).** *MPU* is an index of grammatical development that accounts for the highly synthetic nature and rich morphology of Semitic languages (Dromi & Berman, 1982). *MPU* is equivalent to the Mean Length of Utterance (*MLU*; Brown, 1973) in English. A language sample of 100 utterances was obtained using the wordless storybook "Frog, where are you" (Mayer, 1969). Using this sample, we followed the guidelines of Shaalan and Khater (2006) for MPU calculations in Arabic. The *MPU* score reflects the total number of morphemes divided by the total number of utterances produced in the narrative task. Clinically, low *MLU* scores are viewed as supporting evidence for the diagnosis of language impairment in children (Rice et al., 2010). In addition to the language tasks, the Colored Progressive Matrices (*CPM*, Raven, 2007) was administered to assess the children's nonverbal abilities.

Given that all the measures are not standardized, the results of the TD group (mean and standard deviation) were used to calculate the z scores for all participants (see Table 3). Each child in the DLD group scored at or below −1.5 SD of the TD mean on at least two of the linguistic measures (*A-SRT*, *AVET, ANPT, MPU*) - see Supplemental Material 1 for the individual scores of all participants. Groups were compared using raw scores. Children with DLD scored significantly below the TD children on the *A-SRT* ($t(47.46) = -15.64$, $p < .001$, $d = 3.63$), *AVET* ($t(31.67) = -9.98$, $p < .001$, $d = 2.52$), the *ANPT* ($t(84.58) = -12.56$, $p < .001$, d

= 2.58), and *MPU* ($t(72.49)$ =-11.28 , $p < .001$, $d = 2.42$). The raw scores on the CPM did not differ significantly between the groups ($t(51.59) = - 1.26$, $p = .214$, $d = 0.29$).

INSERT Table 3 ABOUT HERE

Nonword repetition task

The design of the NWR task used in this study was motivated by the Crosslinguistic Nonword Repetition Framework (CL-NWR; Chiat, 2015) which was established within the COST Action IS0804 "Language Impairment Testing in Multilingual Settings" (LITMUS; Armon-Lotem et al., 2015). The goal of the CL-NWR Framework was to design NWR tasks containing nonwords of minimal language-specific features such that these tasks can discriminate between children with and without DLD regardless of their language background (Chiat, 2015). The framework is comprised of three types of tests that vary in the phonological characteristics of nonwords, one of which is the Crosslinguistic (Quasi-Universal) NWR test (CL-NWRT; Chiat, 2015). The test examines phonological short-term memory and was constructed to be maximally compatible with languages with diverse phonological systems. Specifically, the test contains 16 nonwords varying in length from two to five syllables. The syllables are of CV structure, a simple syllable type that is relatively universal. The syllables of nonwords were composed using a set of consonants /p, b, t, d, k, g, s, z, l, m, n/ and vowels /a, u, i/ that are the most common sounds across languages (Chiat, 2015).

Within the CL-NWR , dos Santos & Ferré (2018) developed the French LITMUS Nonword Repetition Test (LITMUS-NWRT). The test aimed to assess phonology with a particular focus on the effects of phonological complexity. Three phonological aspects (based on French phonology but also applicable to a large number of different languages; dos Santos & Ferré, 2018) were systematically manipulated including syllable structure, segmental complexity and sequential complexity. In line with the CL-NWR Framework (Chiat, 2015), the LITMUS-NWR task contained a set of language-specific nonwords and a set of language-independent

(Quasi-Universal) nonwords. The latter set was created using phonemes and phonotactic rules compatible with a large number of languages (Maddieson et al., 2011). Furthermore, this set was adapted into Lebanese Arabic by Abi-Aad and Atallah (2012) resulting in the Arabic version of the Quasi-Universal LITMUS NWRT (QU-LITMUS-NWRT). The set was adapted to identify Lebanese bilingual children whose first language (L1) was Arabic and second language (L2) was French/English.

With regards to syllabic structure complexity, the items of the Arabic QU-LITMUS-NWR had 13 syllabic structures made of three-syllable types. The first type was CV syllable structure which was the same structure used in the CL-NWRT (Chiat, 2015). The QU-LITMUS-NWR also included CCV and CVC syllables which were not present in the CL-NWRT (Chiat, 2015). While syllables with CV structure are common across all languages, syllables with consonant clusters (CC) or Codas are not. The inclusion of these structures was justified by their known effects on NWR performance in languages that permit them, in this case: French, Arabic and English (e.g., Coady & Evans, 2008; dos Santos & Ferré, 2018; Shaalan, 2010).

Segmental complexity of the nonwords was varied for the consonants. This resulted in a smaller set of consonants compared to the CL-NWRT (Chiat, 2015). The nonwords were created using only four consonants /k,f,b,l/ and three vowels /a,u,i/. The stops /p/ (in the Arabic version /b/) and /k/ were contrasted for their place of articulation with /k/, a dorsal stop, being more complex than /b/ which is a labial stop (dos Santos & Ferré, 2018). These two stops were contrasted with the fricative /f/ of which the manner of articulation is considered to be more complex. Moreover, the liquid /l/ was chosen to enable the formation of nonwords with branching onsets that are permitted across many world's languages (dos Santos & Ferré, 2018). Importantly, these consonants are acquired early in the phonological systems of most languages (Abi-Aad & Atallah, 2012; dos Santos & Ferré, 2018). In Arabic, /k/ and /f/ are acquired by 2;10 years, /b/ is acquired by 3;4 years and /l/ by 3;10 years (Amayreh & Dyson, 1998).

Additionally, Sequential complexity (sequentiality) was taken into account. According to dos Santos and Ferré (2018), sequentiality could increase item complexity at two levels: consonant sequences and syllable sequences (for further details, see dos Santos & Ferré, 2018).

The Arabic QU-LITMUS-NWR contained 30 nonwords varying in length from one to three syllables. Given that the main purpose of the QU-LITMUS-NWR was to assess effects of phonological complexity, the influence of working memory was restricted by limiting the length of nonwords to three syllables (Abi-Aad & Atallah, 2012; dos Santos & Ferré, 2018). Hence, the nonwords in the current task are shorter (up to 3 syllables) compared to those in the CL-LITMUS-NWR test (Chiat, 2015) which increased the nonwords' syllable number (up to 5 syllables) rather than syllable complexity to be compatible with languages that lack complex syllables.

According to Abi-Aad and Atallah (2012), the Arabic QU-LITMUS-NWR has quasi-universal prosody to control for familiarity with lexical phonology of the target. That is, the syllables of the nonwords receive equal stress and they are produced with even length and pitch, with the exception of the final syllable lengthening which typically marks the end of an utterance (Chiat, 2015). In this way, language-specific prosodic patterns were avoided.

Lastly, given that wordlikeness affects NWR performance (Archibald & Gathercole, 2006), a familiarity questionnaire (Abi-Aad & Atallah, 2012) was used to obtain familiarity ratings for the nonwords from 30 Palestinian Arabic-speaking adults (10 males, M_{age} = 25.32 years, *SD* = 5.79). After hearing the auditorily presented nonwords, participants were asked to rate each nonword on a 5-point scale, where 1 = “” this word is very unlike an Arabic word” and 5=” this is a very Arabic-like word”. Nonwords with an average score above 2.5 were considered to be of high wordlikeness and those equal or below 2.5 were considered to be of low wordlikeness. There were 7 nonwords in the high wordlikness category (*M*= 3.43, *SD*=

.74) and 23 nonwords in the low wordlikeness category (M = 1.65, SD = .33). The items on the Arabic QU-LITMUS-NWR test (Abi-Aad & Atallah, 2012) are presented in Appendix 1.

Procedure

Written informed consent was obtained by the parents of all participating children before testing. Children were participating in a larger study and completed a battery of tests in two separate sessions each lasting approximately one hour. In the first session, CPM, a narrative task, ANPT and a sentence repetition task were administered. In the second session, CL-NWR, AVET, a grammatical judgement task and a nonword discrimination task were administered. All tests were conducted by the first author who is a qualified SLT and a native speaker of Palestinian Arabic. Each child was tested individually, in a quiet room, in their kindergarten or the speech and language therapy clinic they were attending.

The NWR task was administered in the form of a stringing beads game. Children were given wooden animal beads and were given the following instruction in Arabic: "Now, you will put the wooden animal block next to your ear and listen to the funny word it will say. Listen carefully and repeat the funny word immediately and exactly as you heard it. After you repeat the funny word, you will insert the bead in the thread. Then, you will pick up another animal bead and listen to another funny word" and so on. The nonwords were produced live by the researcher. Live presentation is less consistent compared to the use of audio-recorded nonwords. However, it is a more natural approach, and it is more relevant to clinical practice in that it is similar to tasks employed in speech and language therapy sessions (Chiat & Roy, 2007). The use of an interactive game alongside the live presentation of nonwords has been used in previous studies and shown to be effective in motivating children and maintaining their attention (Chiat & Roy, 2007; Kapalková et al., 2013). To ensure consistency of the delivery of the stimuli across children, the first author practiced the production of the items and conducted the test with all children.

Two practice items were provided before the test was administrated. The practice nonwords were repeated until the children understood what they had to do. The experimental nonwords were presented in a fixed randomized order to all children. Each experimental nonword was only presented once unless there was an interruption to the first presentation (e.g., loud noise, the child being distracted). If the child self-corrected him/herself, the final response was scored regardless of its accuracy. To keep the children motivated, they were praised with "well done" or "bravo" for their responses irrespective of their accuracy. The children's responses were audio-recorded and were transcribed phonetically off-line by the first author for analysis.

Coding and scoring

Following the Crosslinguistic Nonword Repetition Framework (Chiat, 2015), children's responses were scored using item-level scoring. Each repeated nonword was scored as correct if it contained all the consonants and vowels of the target in the correct order. This scoring method did not allow for typical developmental phonological errors. Repetitions that included any additions, omissions or substitutions were scored as incorrect. Correct repetitions received a score of 1 while incorrect repetitions received a score of 0. The maximum raw score was 30. Item-level (binary) scoring is a straightforward scoring method for SLTs to use in clinical settings. Item-level scoring is commonly used for NWR tests such as the Children's Test of Nonword Repetition (Gathercole & Baddeley, 1996) and the Preschool Repetition Test (Seeff-Gabriel et al., 2008). Calculating the percentage phonemes correct (PPC) is also a common scoring method for NWR tests. Roy and Chiat (2004) compared the item-level scores and PPC scores in a sample of English-speaking children. They concluded that the two scoring methods were equally able to differentiate between TD and clinical samples, but item-level scoring was less-time consuming. Kapalková et al. (2013) explored several NWR scoring methods in a sample of Slovak-speaking children. She found that item-level scores did not discriminate between 3, 4 and 5-year-old TD children, allowing for the use of one cutoff point for all age groups Item-level scoring was more accurate than a vowel scoring method in differentiating children with and without DLD (Kapalková et al., 2013). Furthermore, in Spanish-speaking children, item-level scores have yielded better levels of diagnostic accuracy compared to the PPC scores (e.g., Guiberson & Rodríguez, 2013; Gutiérrez-Clellen & Simon-Cereijido, 2010; Windsor et al., 2010). Across languages, item-level scores on NWR tasks have sufficiently discriminated children with language impairments from TD peers (Dispaldro et al., 2013; Kalnak et al., 2014; Kapalková et al., 2013; Kazemi & Saeednia, 2017; Roy & Chiat, 2004; Topbaş et al., 2014).

To calculate inter-rater reliability, a second native Palestinian Arabic-speaking SLT independently scored the audio-recorded responses of 25 children (27% of the sample). The intra-class correlation coefficient (ICC; absolute) was found to be excellent (ICC = .93).

Results

Analysis 1: Group differences

All statistical analyses were performed using R Studio software, version 3.6.3 (R Core Team, 2020). All raw scores were converted to percentages.

To address the first research question, we examined the differences in accuracy scores of the TD and DLD groups. Table 4 summarizes the overall performance of the two groups on the QU-LITMUS-NWRT task as well as their scores across nonwords that vary in terms of length, presence of consonant clusters (CC) and wordlikeness.

INSERT Table 4 ABOUT HERE

The dependent variable was NWR accuracy (where "correct" response = 1 and "incorrect = 0). Given that this is a binary outcome with assumed binominal distribution, data were analyzed using mixed-effects logistic regression models (Baayen et al., 2008) with *lme4* package (Bates et al., 2015). The independent variables were nonword length (3 levels: one, two and three syllables), the presence of CC (3 levels: none, one and two CC) and wordlikeness

(2 levels: high word-like, low word-like) and group (2 levels: TD, DLD). Age was entered as a covariate. All independent variables were contrast-coded and entered as fixed effects. To account for the variability within participants and items, the model included crossed random intercepts for participant and item (Baayen et al., 2008). Fitted models were compared in terms of Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), with reduced AIC and BIC values indicating a better model fit (Tabachnick & Fidell, 2007). This was supplemented by Likelihood ratio tests conducted to determine if the inclusion of a predictor significantly improved the model fit (Baayen et al., 2008; Tabachnick & Fidell, 2007).

First, we examined whether the inclusion of the random effects structure was permitted. This was done by comparing a baseline generalized linear model without the random intercepts (null model) with a baseline mixed-effects model that only included the random intercepts. Relative to the null model (AIC = 2731), the baseline mixed-effects model provided a substantially better fit for the data ($AIC = 1708$ $\chi^2(2) = 1027$, $p < .001$). Therefore, the inclusion of the random intercepts was justified.

Next, we implemented a step-wise-step up procedure for building the mixed-effects model. Age was entered first as a covariate. Next, the predictors: group, nonword length, consonant clusters and wordlikeness variables were entered into the model, respectively, followed by their interactions. A summary of the model fitting procedure is provided in Supplemental Material 2. The fit of the final model ($M8$) was significantly better than the intercept-only baseline model ($AIC = 1596$, $\chi^2(12) = 1157$, $p < .001$). The output of the final model is presented in Table 5. The significance level of the main effects of the fixed factors was obtained using the Anova() function. The estimated marginal means (EMM) were obtained using the *emmeans* package (Lenth, 2020), with all pairwise comparisons corrected using Tukey's HSD adjustment.

INSERT TABLE 5 ABOUT HERE

There was a main effect of age ($X^2(1) = 7.24$, $p < .01$). There was a main effect of group ($X^2(1) = 114.53$, $p < .001$), with the TD group (*EMM* = 4.42, *SE* = .40) scoring higher than the DLD group on the task (*EMM* = .36, *SE* = .42, $p < .001$). The group by age interaction was not significant ($X^2(1) = 1.60$, $p = .207$).

There was a main effect of nonword length ($X^2(2) = 32.72$, $p < .001$), such that three-syllable nonwords (*EMM* = 1. 06, *SE* = .44) were repeated less accurately compared to one-syllable (*EMM* = 3.54, *SE* = .48, $p < .001$) and two-syllable nonwords (*EMM* = 2.58, *SE* = .39, $p < .001$). The difference in the repetition accuracy of one and two syllable nonwords was not significant ($p = .106$). The group by nonword length interaction was not significant ($X^2(2) = .79$, $p = .673$).

There was a significant effect of the number of consonant clusters ($X^2(2) = 11.41$, $p < .01$), such that nonwords with two consonant clusters (*EMM* = 2.26. *SE* = .68) were repeated less accurately compared to nonwords with no consonant clusters (*EMM* = 3.22, *SE* = .34, p <.01) but were comparable to nonwords with one consonant cluster (*EMM* = 2.70, *SE* = .36, $p = .084$). The repetition accuracy of nonwords with no or one consonant cluster did not differ significantly ($p = .376$).

The group by number of consonant clusters interaction was significant ($X^2(2) = 9.98$, $p < .01$). The interaction is illustrated in Figure 1 which plots the proportion of correctly repeated nonwords as a function of number of consonant clusters for the TD and DLD groups. It can be observed that, for the DLD group, the repetition accuracy decreased more significantly with an increased number of consonant clusters. This reduction in accuracy appears to be much less pronounced for the TD group.

INSERT FIGURE 1 ABOUT HERE

Post-hoc comparisons showed that, within the DLD group, nonwords with two consonant clusters (EMM = -.86, *SE* =. 75) were repeated less accurately than nonwords without consonant clusters (*EMM* = 1.48, *SE* = .40, $p < .05$) or with one consonant cluster (*EMM* = .46, *SE* = .42, $p < .05$). There was no difference in repetition accuracy of nonwords with one or two consonant clusters ($p = .879$).

Within the TD group, the repetition accuracy of nonwords with two consonant clusters (*EMM* = 3.38, *SE* = .74)) was not significantly different to nonwords without consonant clusters (*EMM* =4.96, *SE* = .43, $p < .433$) or with one consonant cluster (*EMM* = 4.94, *SE* = .41, p = .422). There was no difference in repetition accuracy of nonwords without consonant clusters and nonwords with one consonant cluster ($p = 1$). The TD group outperformed the DLD group in repeating nonwords with one, two or no consonant clusters (for all comparisons, $p < .001$).

The effect of wordlikeness was significant ($X^2(1) = 5.72$, p < *.05*). Highly word-like nonwords (*EMM* = 3.01, *SE* =.55) were repeated more accurately than nonwords that were less word-like (*EMM* = 1.77*,* *SE* = .32, $p < .05$). Group by wordlikeness interaction was not significant ($X^2(1) = .37$, p = *.542*).

Analysis 2: Diagnostic accuracy of the nonword repetition task

To address the second research question, we assessed the diagnostic accuracy of the QU-LITMUS-NWRT. Receiver Operating Characteristic (*ROC*) curve was generated using the *pROC* package (Robin et al., 2011). *ROC* curves plot the true positive rate (sensitivity) as a function of false-positive rate (1 – specificity) for every possible cutoff score (Gonçalves et al., 2014). Consequently, the optimal cutoff score with the highest sensitivity and specificity values is determined. Also, the area under the *ROC* curve (*AUC*) was computed. *AUC* is an index of the test classification accuracy and it reflects the probability that a randomly selected child with DLD will have a lower score than a randomly-selected TD child. According to Carter et al. (2016), *AUC* values range from .5 to 1.0. An *AUC* of 1.0 indicates a perfect test, .90– .99 is an excellent test, .8 – .89 a good test, .7 – .79 a fair test, and lower than .7 is a non-useful test. Sensitivity, Specificity, and Likelihood Ratios were calculated for the final cutoff score.

Figure 2 presents the ROC curve for the QU-LITMUS-NWRT using item-level scoring. Based the *ROC* analysis, the optimum cutoff score was 81.67% (equivalent to a raw score 24 out of 30). The diagnostic accuracy of the cutoff score was excellent: AUC = .99 [95% CI = .94 – 1], Sensitivity = .93 [95% CI = .83 – .10], Specificity = .93 [95% CI = .87 – .98], LR+ = 13.93 [95% CI = 5.41 – 36.26], LR- = .07 [95% CI .02 – .27].

INSERT FIGURE 2 ABOUT HERE

Discussion

This is the first study to examine the diagnostic accuracy of NWR for the identification of DLD in Arabic. This study found that 4 to 6-year-old Palestinian Arabic-speaking children with DLD performed below the level of age-matched TD controls on the QU-LITMUS-NWRT. Nonword length and wordlikeness ratings appeared to influence NWR accuracy of TD and DLD groups whereas the presence of CC influenced the NWR accuracy of the DLD group only. The QU-LITMUS-NWRT was found to have excellent diagnostic accuracy in distinguishing children with DLD from TD peers, indicating that it is a promising measure that clinicians could include within their assessment battery to establish DLD diagnosis in Arabic-speaking children.

Evidence that Arabic-speaking children with DLD have poor nonword repetition abilities compared to their TD peers

The accuracy scores of the DLD group were substantially lower than those of the TD group on the QU-LITMUS-NWRT (52% versus 93%). This result aligns with existing literature documenting that children with DLD have considerable difficulty in repeating nonwords compared to age-matched TD peers across languages (Ahufinger et al., 2021; Armon-Lotem & Meir, 2016; de Bree et al., 2007; Girbau, 2016; Graf Estes et al., 2007; Kalnak et al., 2014; Kapalková et al., 2013; Topbaş et al., 2014). Our findings are also consistent with previous studies which showed poor performance of Arabic-speaking children with or at risk of DLD on language-specific NWR tasks (Balilah, 2017; Khater, 2016; Saiegh-Haddad & Ghawi-Dakwar, 2017; Shaalan, 2010). It should be noted that these studies used NWR tests that were language-specific i.e., followed Arabic phonotactics, while in this study we used a quasi-language independent NWR test. The fact that there were significant group differences on the QU-LITMUS-NWRT, -language-independent test- suggests that the test is as sensitive as language-specific Arabic NWR tests to the language difficulties of Arabic-speaking children with DLD.

There was a main effect of age on NWR accuracy in the TD and DLD groups suggesting that scores on the QU-LITMUS-NWRT improved with age. The effect of age replicates studies which have reported that older children outperformed younger children on NWR tasks (e.g., Chiat & Roy, 2007; Guiberson & Rodríguez, 2013; Kapalková et al., 2013; Roy & Chiat, 2004; Weismer et al., 2000).

Several item characteristics appeared to influence task performance. For both groups, repetition accuracy decreased as the nonwords increased in length. Accuracy fell significantly for three-syllable nonwords compared to one and two-syllable nonwords. The non-significant group by nonword length interaction suggests that the effect of length on NWR was equivalent across for both groups. This result contradicts studies showing that, as nonwords increase in length, repetition accuracy decreases for TD and, to a greater degree, DLD groups (Archibald & Gathercole, 2006; Chiat & Roy, 2007; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990; Jones et al., 2010; Weismer et al., 2000). Particularly, research shows differences between TD and DLD groups are larger when repeating nonwords of three or more syllables (Archibald & Gathercole, 2006; Chiat & Roy, 2007; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990; Jones et al., 2010; Weismer et al., 2000). The additional disadvantage noted in DLD groups in repeating long nonwords has been explained in the light of a limitation in their phonological short-term memory (e.g., Archibald & Gathercole, 2006;

Gathercole & Baddeley, 1990). However, as mentioned above, the developers of the QU-LITMUS-NWRT aimed to limit the effect of length on NWR as their focus was to evaluate the effects of phonological complexity (e.g., presence of CC) on NWR. Hence, the fact that the test had relatively short nonwords of one, two and three-syllables could have contributed to the lack of interaction between the two variables. Previous research with Gulf-Arabic speaking children has documented similar findings when using a NWR task containing two and three-syllable nonwords (Shaalan, 2010).

The number of CCs in nonwords seemed to affect the repetition accuracy of the DLD group only. The DLD group repeated nonwords with two CCs less accurately than nonwords with one or no CCs. This is in line with earlier studies showing that nonwords with CCs are more difficult to repeat than nonwords with singleton consonants in children with DLD (Briscoe et al., 2001; Coady & Evans, 2008; Graf Estes & Else-Quest, 2007; Leclercq et al., 2013; Munson et al., 2005). It is suggested that the increased articulatory complexity of nonwords with CC places higher demands on speech motor output processes since their production involves the coordination of many articulatory movements within syllables. This, in turn, increases the likelihood of articulation errors occurring (Archibald et al., 2013). However, given that articulatory control skills were not measured in this study, such a conclusion is not possible.

The TD and DLD groups in our study showed a higher repetition accuracy of high word-like nonwords than low word-like nonwords. This result extends previous research indicating that knowledge stored in long term memory supports NWR (Archibald & Gathercole, 2006; Gathercole & Baddeley, 1990; Jones et al., 2010; Munson et al., 2005). A non-significant interaction between group and wordlikeness ratings revealed that wordlikeness affected both groups similarly, although the scores of the DLD group were lower than those of the TD group on high and low word-like nonwords.

Poor nonword repetition as a possible clinical marker of Arabic DLD

The Arabic version of the QU-LITMUS-NWRT (Abi-Aad & Atallah, 2012) showed an overall excellent diagnostic accuracy in differentiating 4 to 6-year-old, Palestinian Arabic-speaking children with DLD from their age-matched TD peers. ROC analyses using item-level scores revealed that a cutoff score of 81.67% on the task had the best overall classification accuracy (93%). The sensitivity and specificity of the cutoff score were equal to 93% showing a good value in terms of diagnostic accuracy (Plante & Vance, 1994). These results mean that the QU-LITMUS-NWRT correctly identified 28 out of 30 children with DLD as having DLD (sensitivity) and 56 out of 60 TD children as being TD (specificity).

Our findings are in contrast to those of Wallan (2018) who found that a nonword list recall task had inadequate diagnostic accuracy in distinguishing Arabic-speaking children with language concerns (LC) from their TD peers. The nonword list recall task in Wallan (2018)'s study correctly identified 89% of TD children but only 56% of the children with LC. The difference in results can be attributed to several reasons. Firstly, in the task used by Wallan (2018), children were asked to repeat a list of up to four nonwords whereas the QU-LITMUS-NWR used in our study was less demanding as children repeated one nonword at a time.

Secondly, the performance of the TD and LC groups on the nonword recall list was approximately similar with both groups showing floors effects in Wallan's study (2018). Out of a maximum score of 4 points, the mean score for the TD group was 1.63 (*SD* = .47) and 1.16 (*SD* = .35) for the DLD group. This suggests that the nonword recall task used by Wallan (2018) was difficult even for the TD children. In our current study, performance of the TD group was close to the ceiling and significantly higher than the DLD group, showing a large effect size (*d* = 2.62).

Importantly, none of the children in the LC group (*n* = 16) in Wallan's study had a confirmed diagnosis of DLD. Although children in the LC may have weaker language skills compared to their TD peers, the level of their language ability might have not been low enough for a DLD diagnosis. On the other hand, the children in our study had a DLD diagnosis and were receiving language intervention at the time of the study. This means that the DLD group in our study may have had more severe language difficulties compared to the LC group in Wallan's (2018) study. The less demanding nature of the QU-LITMUS-NWR compared to the nonword list recall used in Wallan's (2018) study and the more stringent criteria for the DLD children recruited for our study may have enlarged the differences between the TD and DLD group in our study, positively influencing the diagnostic accuracy of the task.

We further calculated the Likelihood ratios for the QU-LITMUS-NWR. The LR+ was 13.93, and the LR− was equal to .07. Based on Dollaghan, (2007), values of LR+ ≥ 10.0 and LR- ≤ 0.1 can be interpreted with confidence. Thus, based on the QU-LITMUS-NWR alone, one can conclude that a child who scores below the cutoff (81.67%) may have DLD and a child who scores above it may not. Although the 95% confidence intervals for the LRs include values that fall beyond the threshold mentioned above, they remain within the informative range. This points to the diagnostic value of the QU-LITMUS-NWR for the identification of DLD in Arabic.

The finding that NWR has a good level of accuracy in identifying children with DLD and excluding TD children is not trivial. It replicates the existing literature which reported good diagnostic accuracy for NWRTs in identifying children with DLD acquiring typologically different languages (Armon-Lotem & Meir, 2016; Dispaldro et al., 2013; Kalnak et al., 2014; Kapalková et al., 2013; Kazemi & Saeednia, 2017; Thordardottir et al., 2011; Topbaş et al., 2014). The excellent identification accuracy of the QU-LITMUS-NWR and its consistency with the DLD literature provides strong evidence that NWR should be considered as a potential clinical marker of DLD in Arabic-speaking children.

Clinical implications

Our findings form a stepping-stone into advancing the diagnostic procedures for identifying Arabic-speaking children with DLD in the Palestinian context and other Arab countries where speech and language therapy remains a relatively under-developed field. SLTs face difficulty in diagnosing DLD in Arabic due to the poor availability of appropriate language assessments. When examining the language abilities of Arabic-speaking children, the sole reliance on qualitative assessments and/or subjective clinical judgment might not provide sufficient or reliable evidence regarding the presence or absence of DLD. As a result, Palestinian Arabic-speaking children with DLD encounter an increased risk of under-identification and misdiagnosis.

This study offers information that can contribute to a more accurate evaluation of Arabic-speaking children with DLD. Our findings show that poor NWR has good discriminatory power in distinguishing between Arabic-speaking children with and without DLD. Consequently, our results highlight the importance of considering NWR abilities besides the informal language measures when diagnosing DLD in Arabic. Particularly, the study highlights the potential of the Arabic version of the QU-LITMUS-NWR as a useful indicator/index of DLD that is quick to administer.

Previous Arabic studies showed that children with DLD perform poorly on NWR tasks. An important contribution of our study is that we can specify what the threshold performance should be for a child to be considered for further assessment. For the QU-LITMUS-NWR task, a cutoff point of 81.67%, equivalent to a score of 24, could be used to determine whether a child's language abilities need further assessment.

The QU-LITMUS-NWRT was constructed using early acquired sounds and syllabic structures that are common across all Arabic dialects (Watson, 2000) as well as across many languages (Maddison, 2008). This means that the use of the test can be extended beyond identifying DLD in monolingual children acquiring Palestinian Arabic to other Arabic dialects. The design of QU-LITMUS-NWR makes it suitable to be used with bilingual children whose L1 or L2 is Arabic once its diagnostic accuracy in identifying DLD in this population is explored.

Limitations and future directions

Although promising, our findings are preliminary and should be interpreted with caution. Our study followed a two-gate design in which preselected TD and DLD groups were recruited. Two-gate designs are very common in diagnostic studies, however, they could lead to a spectrum bias (Pawłowska, 2014; Redmond et al., 2019). Children with DLD in this study were receiving language intervention and may not be representative of Palestinian Arabic-speaking children with DLD in terms of severity. Population-based one-gate designs are needed to validate our results.

The diagnostic accuracy of the NWR task should be considered with relevance to the reference standards of DLD employed in this study. The first reference standard was the receipt of speech and language therapy intervention Children with DLD were diagnosed prior to the current study. To verify the DLD status of the children, our second reference standard was poor performance (below 1.5 SD) on at least two morpho-syntactic measures. These tasks only assess expressive morphology and their use as a reference standard might be limited with children with DLD whose language difficulties do not involve grammar (e.g., semantics). Notably, reference standards that are used to estimate diagnostic accuracy are not interchangeable (Redmond et al., 2019). Hence, if different reference standards are used, the diagnostic accuracy of the current task may vary.

Live administration of the QU-LITMUS-NWRT was engaging for the children. However, live administration could be associated with inevitable variations in rate, pitch, loudness when the examiner delivered the test to different children. This could have influenced the children's performance in the test. Therefore, future studies should consider the use of audio-recorded stimuli to ensure consistency of delivery of the test.

Although it has been reported that oral motor planning influences NWR performance (e.g., (Archibald et al., 2013), no measures of this ability were taken as part of this study. Future studies of NWR in Arabic should take this measure into account as it could provide us with insights about the underlying cause of NWR difficulties in Arabic-speaking children with DLD. It also needs to be pointed out that there was an imbalance between the number of nonwords in the categories of word-likeness and CCs. Although we reported the significant and insignificant interactions (group and wordlikeness, and group and number of consonant clusters), they are likely to have been conflated with non-word length which limits the interpretation of the analysis of these interactions.

Conclusion

This study offers valuable implications for the assessment of DLD in Palestinian Arabic-speaking children. Children with DLD were found to perform poorly on the Arabic version of the quasi-universal LITMUS Nonword Repetition Test (QU-LITMUS-NWRT; Abi-Aad & Atallah, 2012). In the current study, the QU-LITMUS-NWRT was found to have high diagnostic accuracy, suggesting that it should be considered as a clinical marker of DLD in Arabic-speaking children aged 4 to 6 years. The test could be used by SLTs – alongside other language measures- to improve the accuracy of identifying DLD in Arabic. However, the adaptation of the task for clinical use requires further validation of its diagnostic accuracy. The use of one-gate designs incorporating reference standards that cover different language domains will be needed to include a more representative, heterogeneous group of children with DLD.

Acknowledgments

We would like to thank the children and their parents who participated in the study. We also thank the teachers and speech and language therapists who facilitated the recruitment of the children. This work was funded by [REMOVED FOR REVIEW].

References

Abdallah, F., Aljenaie, K., & Mahfouthi, A. (2013). Plural noun inflection in Kuwaiti Arabic-speaking children with and without Specific Language Impairment. *Journal of Child Language*, *40*(1), 139–168.

Abdallah, F., & Crago, M. (2008). Verb morphology deficits in Arabic-speaking children with specific language impairment. *Applied Psycholinguistics*, *29*(2), 315–340.

Abi-Aad, K., & Atallah, C. (2012). *Phonologie , Plurilinguisme et Trouble Spécifique du Langage Oral au Liban : Etude Pilote sur la Pertinence d ' un Test de Répétition de Non-Mots. (Unpublished Thesis)*. Universite Saint-Joseph.

Ahufinger, N., Berglund-Barraza, A., Cruz-Santos, A., Ferinu, L., Andreu, L., Sanz-Torrent, M., & Evans, J. L. (2021). Consistency of a Nonword Repetition Task to Discriminate Children with and without Developmental Language Disorder in Catalan–Spanish and European Portuguese Speaking Children. *Children*, *8*(2), 85. https://doi.org/10.3390/children8020085

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. *Journal of Speech, Language, and Hearing Research*, *41*(3), 642–653.

Archibald, L. M. D., & Gathercole, S. E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language and Communication Disorders*, *41*(6), 675–693.

Archibald, L. M. D., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, *52*(4), 899–914.

Archibald, L. M. D., Joanisse, M. F., & Munson, B. (2013). Motor control and nonword repetition in specific working memory impairment and SLI. *Topics in Language*

Disorders, *33*(3), 255–267.

Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: evidence from Russian and Hebrew. *International Journal of Language and Communication Disorders*, *51*(6), 715–731. https://doi.org/10.1111/1460-6984.12242

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Balilah, A. M. (2017). *Linguistic and Cognitive Measures in Arabic- Speaking English Language Learners (ELLs) and monolingual children with and without Developmental Language Disorder (DLD).(Unpublished Doctoral Thesis)*. The University of Western Ontario.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword Repetition as a Behavioural Marker for Inherited Language Impairment: Evidence From a Twin Study. *Journal of Child Psychology and Psychiatry*, *37*(4), 391–403.

Bishop, Dorothy V.M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Adams, C., Archibald, L., Baird, G., Bauer, A., Bellair, J., Boyle, C., Brownlie, E., Carter, G., Clark, B., Clegg, J., Cohen, N., Conti-Ramsden, G., Dockrell, J., Dunn, J., Ebbels, S., … House, A. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *58*(10), 1068–1080.

Bishop, Dorothy V.M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Adams, C., Archibald, L., Baird, G., Bauer, A., Bellair, J., Boyle, C., Brownlie, E., Carter, G., Clark, B., Clegg, J., Cohen, N., Conti-Ramsden, G., Dockrell, J., Dunn, J., Ebbels, S., … Whitehouse, A. (2016). CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLoS ONE*, *11*(7), 1–27.

Bortolini, U., Arfé, B., Caselli, M. C., Degasperi, L., Deevy, P., & Leonard, L. B. (2006). Clinical markers for specific language impairment in Italian: The contribution of clitics and non-word repetition. *International Journal of Language and Communication Disorders*, *41*(6), 695–712.

Bowey, J. A. (2001). Nonword repetition and young children's receptive vocabulary: A longitudinal study. *Applied Psycholinguistics*, *22*(3), 441–469.

Briscoe, J., Bishop, D. V. M., & Norbury, C. F. (2001). Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(3), 329–340.

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery (United States)*, *159*(6), 1638–1645.

Chiat, S. (2015). Nonword repetition. In S Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–150). Multilingual Matters.

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*(2), 429–443.

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language and Communication Disorders*, *43*(1), 1–40.

Coady, J., Evans, J. L., & Kluender, K. R. (2010). The role of phonotactic frequency in sentence repetition by children with specific language impairment. *International Journal of Language and Communication Disorders*, *45*(4), 494–509.

Conti-Ramsden, G. (2003). Processing and Linguistic Markers in Young Children With Specific Language Impairment (SLI). *Journal of Speech, Language and Hearing Research*, *46*, 1029–1037.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(6), 741–748.

Conti-Ramsden, G., & Hesketh, A. (2003). Risk markers for SLI: A study of young language-learning children. *International Journal of Language and Communication Disorders*, *38*(3), 251–263.

de Bree, E., Rispens, J., & Gerrits, E. (2007). Non-word repetition in Dutch children with (a risk of) dyslexia and SLI. *Clinical Linguistics and Phonetics*, *21*(11–12), 935–944.

Deevy, P., Weil, L. W., Leonard, L. B., & Goffman, L. (2010). Extending use of the NRT to preschool-age children with and without specific language impairment. *Language, Speech, and Hearing Services in Schools*, *41*(3), 277–288.

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*(February), 323–336.

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

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136–1146.

dos Santos, C., & Ferré, S. (2018). A Nonword Repetition Task to Assess Bilingual Children's Phonology. *Taylor & Francis*, *25*(1), 58–71. https://doi.org/10.1080/10489223.2016.1243692

Dromi, E., & Berman, R. A. (1982). A morphemic measure of early language development: Data from modern Hebrew. *Journal of Child Language*, *9*(2), 403–424.

Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. *Applied Psycholinguistics*, *19*(2), 279–309.

Fahim, D. (2017). Verb Morphology in Egyptian Arabic Developmental Language Impairment. *Arab Journal of Applied Linguistics*, *2*(1), 49–73.

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, *33*(6), 966–979.

Gathercole, Susan E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543.

Gathercole, Susan 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.

Girbau, D. (2016). The Non-word Repetition Task as a clinical marker of Specific Language Impairment in Spanish-speaking children. *First Language*, *36*(1), 30–49.

Girbau, D., & Schwartz, R. G. (2007). Non-word repetition in Spanish-speaking children with Specific Language Impairment (SLI). *International Journal of Language and Communication Disorders*, *42*(1), 59–75.

Gonçalves, L., Subtil, A., Oliveira, M. R., De, P., & Bermudez, Z. (2014). ROC CURVE ESTIMATION: AN OVERVIEW. In *REVSTAT-Statistical Journal* (Vol. 12, Issue 1).

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the Nonword Repetition Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195.

Gray, S. (2003). Diagnostic accuracy and test-retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders*, *36*(2), 129–151.

Guiberson, M., & Rodríguez, B. L. (2013). Classification accuracy of nonword repetition when used with preschool-age spanish-speaking children. *Language, Speech, and Hearing Services in Schools*, *44*(2), 121–132.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2010). Using Nonword Repetition Tasks for the Identification of Language Impairment in Spanish-English-Speaking Children: Does the Language of Assessment Matter? *Learning Disabilities Research & Practice*, *25*(1), 48–58. https://doi.org/10.1111/j.1540-5826.2009.00300.x

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and Frequency in Specific Language Impairment: Accuracy and Error Data from Two Nonword Repetition Tests. *Journal of Speech, Language, and Hearing Research*.

Kalnak, N., Peyrard-Janvid, M., Forssberg, H., & Birgitta, S. (2014). Nonword Repetition –

A Clinical Marker for Specific Language Impairment in Swedish Associated with Parents' Language-Related Problems. *PLoS ONE*, *9*(2), e89544.

Kapalková, S., Polišenská, K., & Vicenová, Z. (2013). Non-word repetition performance in Slovak-speaking children with and without SLI: Novel scoring methods. *International Journal of Language and Communication Disorders*, *48*(1), 78–89.

Kazemi, Y., & Saeednia, S. (2017). The clinical examination of non-word repetition tasks in identifying Persian-speaking children with primary language impairment. *International Journal of Pediatric Otorhinolaryngology*, *93*, 7–12.

Khater, M. (2016). The relationship between nonword repetition, root and pattern effects, and vocabulary in Gulf Arabic speaking children.(Unpublished thesis). In *Methods*. City, University of London.

Leclercq, A. L., Maillart, C., & Majerus, S. (2013). Nonword repetition problems in children with specific language impairment: A deficit in accessing long-term linguistic representations? *Topics in Language Disorders*, *33*(3), 238–254.

Lenth, R. (2020). *Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.6*. https://cran.r-project.org/package=emmeans

Leonard, L. B. (2014). *Children with specific language impairment* (2nd ed.). MIT Press.

Mayer, M. (1969). *Frog, where are you?* Dial books for Young Readers.

Melby-Lervåg, M., Lervåg, A., Lyster, S. A. H., Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-Repetition Ability Does Not Appear to Be a Causal Influence on Children's Vocabulary Development. *Psychological Science*, *23*(10), 1092–1098.

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. In *Journal of Educational Psychology* (Vol. 91, Issue 1, pp. 3–19).

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The Influence of Vocabulary Size, Phonotactic Probability, and Wordlikeness on Nonword Repetitions of Children With and Without Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *57*(11), 1247–1257.

Pawłowska, M. (2014). Evaluation of Three Proposed Markers for Language Impairment in English: A Meta- Analysis of Diagnostic Accuracy Studies. *Journal of Speech, Language, and Hearing Research* •, *57*, 2261–2273.

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.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15–24.

Poll, G. H., Betz, S. K., & Miller, C. A. (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research*, *53*(2), 414–429.

Redmond, S. M., Ash, A. C., Christopulos, T. T., & Pfaff, T. (2019). Diagnostic Accuracy of Sentence Recall and Past Tense Measures for Identifying Children's Language Impairments. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2438–2454.

Redmond, S. M., Thompson, H. L., & Goldstein, S. (2011). Psycholinguistic Profiling Differentiates Specific Language Impairment From Typical Development and From

Attention-Deficit/Hyperactivity Disorder. *Journal of Speech Language & Hearing Research*, *54*, 99–117.

Rice, M. L., Smolik, F., Rytting, N., & Blossom, M. (2010). Mean Length of Utterance Levels in 6-month Intervals for Children 3 to 9 Years with and without Language Impairments. *Hearing Research*, *53*(April), 333–349.

Rispens, J., & Parigger, E. (2010). Non-word repetition in Dutch-speaking children with specific language impairment with and without reading problems. *British Journal of Developmental Psychology*, *28*(1), 177–188.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). *pROC: an open-source package for R and S+ to analyze and compare ROC curves*. *BMC Bioinf*, 1471–2105.

Roy, P., & Chiat, S. (2004). A Prosodically Controlled Word and Nonword Repetition Task for 2- to 4-Year-Olds: Evidence from Typically Developing Children. *Journal of Speech, Language, and Hearing Research*, *47*(1), 223–234.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine*. Little, Brown.

Saiegh-Haddad, E., & Ghawi-Dakwar, O. (2017). Impact of diglossia on word and non-word repetition among language impaired and typically developing Arabic native speaking children. *Frontiers in Psychology*, *8*(NOV), 1–17.

Shaalan, S. (2010). *Investigating Grammatical Complexity in Gulf Arabic Speaking Children with Specific Language Impairment (SLI).(Unpublished Thesis)*. University College London.

Shaalan, Saleh, & Khater, M. (2006). *A comparison of two measures of assessing spontaneous language samples in Arabic speaking children*.

Snowling, M., Chiat, S., & Hulme, C. (1991). Words, nonwords, and phonological processes: Some comments on Gathercole, Willis, Emslie, and Baddeley. *Applied Psycholinguistics*, *12*(3), 369–373.

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, Language, and Hearing Research*, *49*(2), 219–236.

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.

Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics* (5th ed.). Pearson.

Team R Core. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.r-project.org/

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau, N., & Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research*, *54*(2), 580–597.

Topbaş, S., Kaçar-Kütükçü, D., & Kopkalli-Yavuz, H. (2014). Performance of children on the Turkish Nonword Repetition Test: Effect of word similarity, word length, and scoring. *Clinical Linguistics and Phonetics*, *28*(7–8), 602–616.

Wallan, A. (2018). *Evaluation of Arabic tests of sentence repetition and verbal short term memory for Saudi preschoolers. (Unpublished Thesis)*. City, University of London.

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.

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)

Wong, A. M. Y., Kidd, J. C., Ho, C. S. H., & Au, T. K. F. (2010). Characterizing the overlap between SLI and dyslexia in Chinese: The role of phonology and beyond. *Scientific Studies of Reading*, *14*(1), 30–57.

Figure captions

Figure 1. Nonword repetition accuracy across nonwords with different numbers of CCs for the Typically developing (TD) children and children with Developmental Language Disorder (DLD)

Figure 2. Receiver Operating Characteristics (ROC) curve for the Item-level scoring method

Figure 1

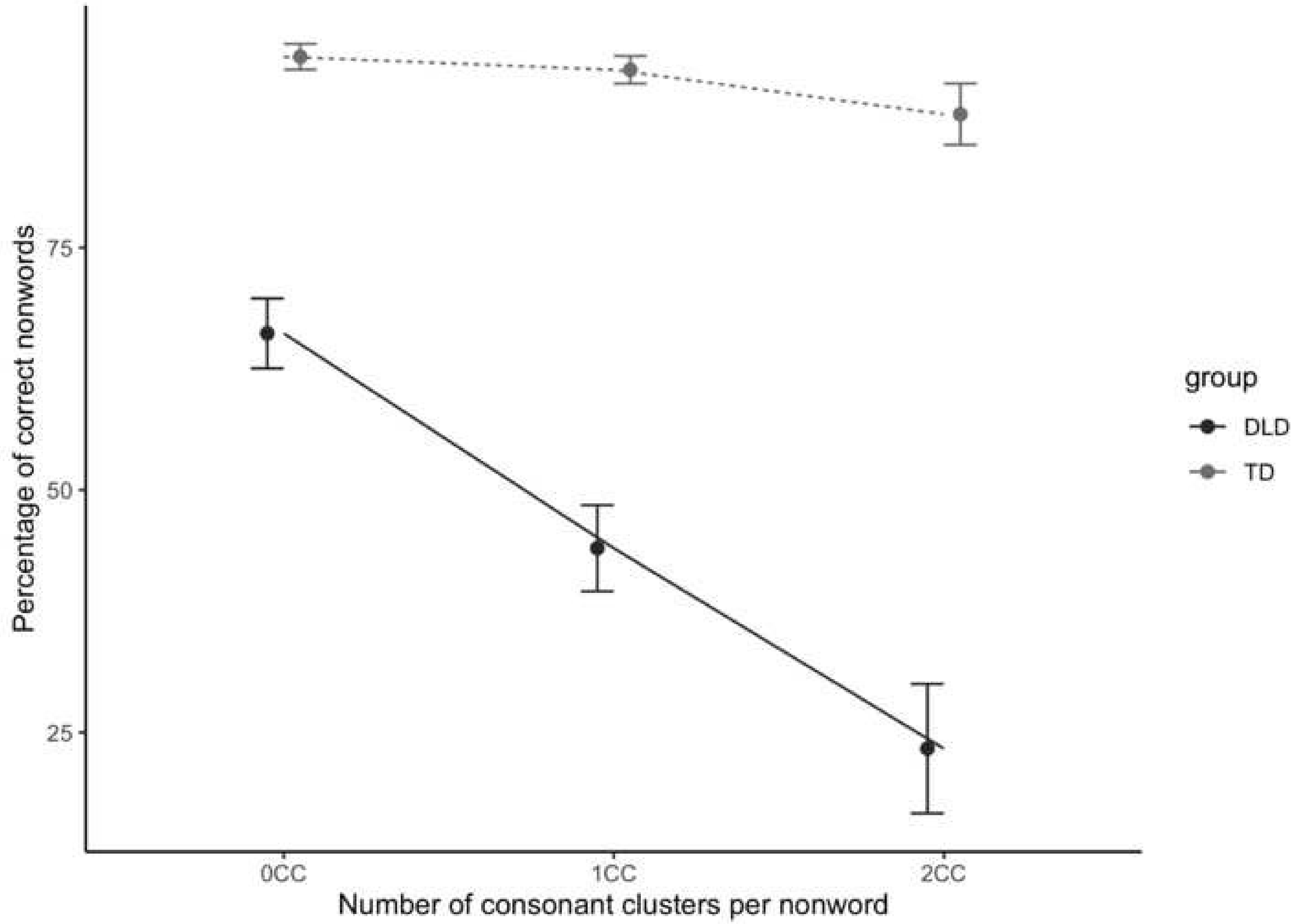
Percentage of correct nonwords
75
50
25
0CC
1CC
2CC
Number of consonant clusters per nonword
group
DLD
TD

Figure 2

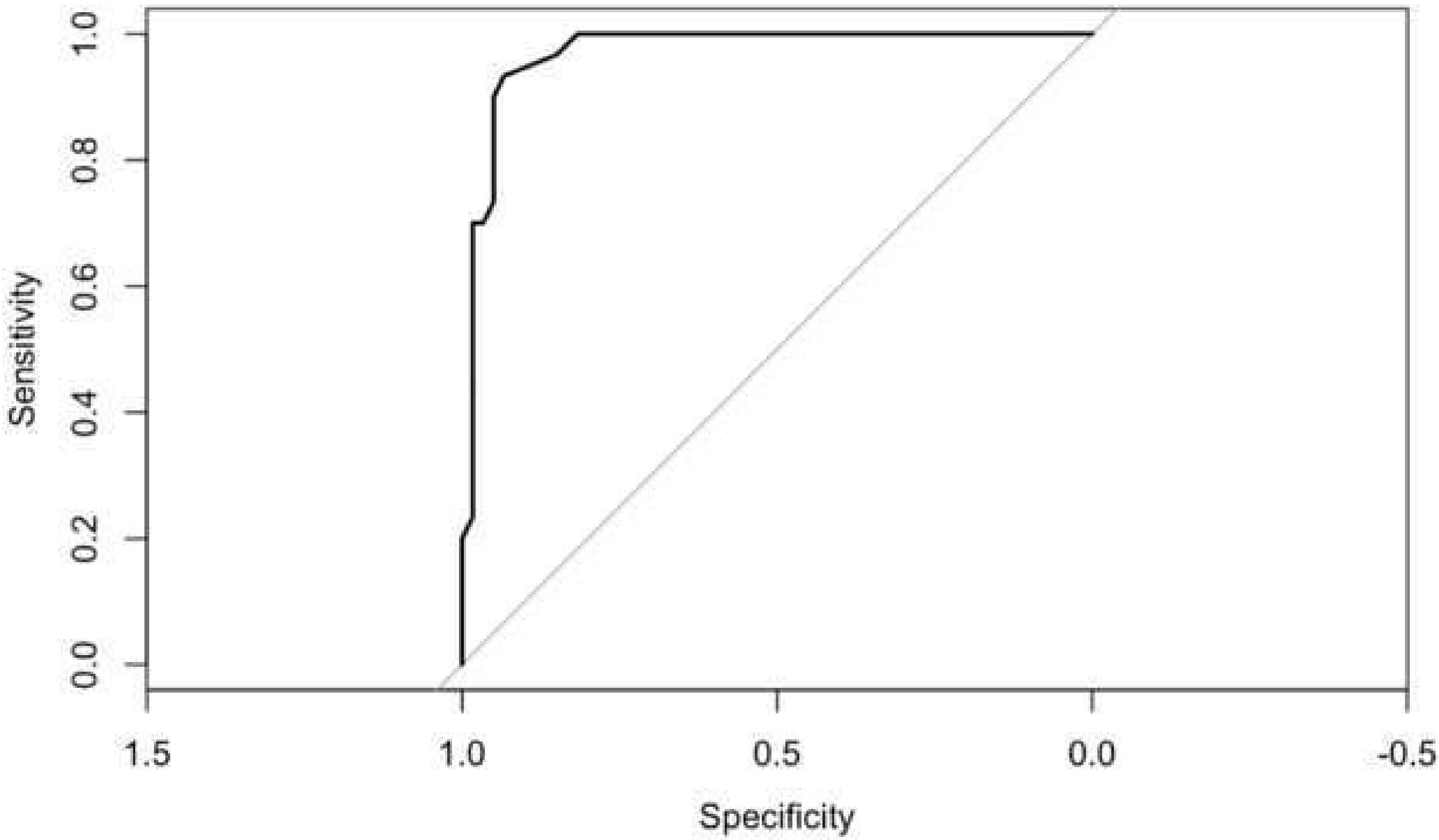

Table 1. *Summary of cross-linguistic findings on the diagnostic accuracy of nonword repetition in identifying DLD in monolingual children*

Reference	Language	*TD*		*DLD*		Sensitivity %	Specificity %	LR+	LR-
		N	Age in years	N	Age in years				
Ahufinger et al. (2021)	**Portuguese**	75	7;0 – 11;11	75	7;0 – 11;11	47	99	35.92	.54
Armon-Lotem & Meir (2016)	**Hebrew**	38	6 (.17)	14	6;1 (.33)	93[a]	66	2.71	.11
Armon-Lotem & Meir (2016)	**Russian**	20	6;1 (.17)	14	5;10 (.25)	86	90	8.57	.16
Bortolini et al. (2006)	**Italian**	11	3;7 - 5;5	11	3;7 - 5;6	82	82	4.56	.22
Dispaldro et al. (2013)	**Italian**	17	3;11-5;8	17	4;1 - 5;7	100	1	ND [b]	ND
Girbau (2016)	**Spanish**	20	8;1- 10;3	20	8;0 - 9;11	100	85	6.67	0
Guiberson & Rodríguez (2013)	**Spanish**	23	4;1 (.82)	21	3;11(.81)	71	74	2.74	.39
Kalnak et al. (2014)	**Swedish**	86	9;4 (1.3)	61	9;3 (1.2)	90	98	38.8	.10
Kapalková et al. (2013)	**Slovak**	16	4;3 - 5;6	16	4;2 - 5;6	94	100	ND	.06
Kazemi & Saeednia (2017)	**Farsi**	31	4;8 (.7)	20	4;5 (.74)	90	96	27.9	.10
Pham & Ebert (2020)	**Vietnamese**	194	5;8 (.4)	10	5;5 (.3)	90	79	4.53	.13
Thordardottir et al. (2011)	**French**	78	4;1 - 5;11	14	4;6 - 5;11	85	88	6.77	.18
Topbaş et al. (2014)	**Turkish**	120	4;4 - 8;0	20	4;2 - 8;3	89	87	6.85	-.02

Note. **TD**: Typically Developing. **DLD:** Developmental Language Disorder. **LR+**: Positive Likelihood Ratio. **LR-** : Negative Likelihood Ratio. **ND**: not defined.
[a] Sensitivity and Specificity and LR values are reported for the best cutoff points.
[b] When the specificity is 100, the LRs are undefined

Table 2. *Participants' characteristics*

	Group	
	TD	***DLD***
Family characteristics	**%(*N*)**	
Mother's education		
High school	20(12)	33.33(10)
University degree/college diploma	75(45)	53.34(16)
Postgraduate degree	5(3)	13.33(4)
Family history of communication disorders	6.67(4)	30(9)**
Language milestones	**Age in months *Mean*(*SD*)**	
Babbling	6.22(1.69)	6.33(1.71)
First word	11.72(2.06)	20.43(6.94)***
Word combinations	19.44(3.53)	35.60(9.37)***
Follow simple commands	18.89(5.07)	26.13(7.33)***

Note. **TD**: Typically Developing. **DLD:** Developmental Language Disorder.

* $p < .05$, ** $p < .01$, *** < .001

Table 3. *A summary of the raw and z scores of the TD and DLD groups on the background measures*

	Group							
	TD				***DLD***			
	Raw scores		***Z scores***		***Raw scores***		***Z scores***	
Measures	M (SD)	Range	M (SD)	Range	M (SD)	Range	M(SD)	Range
A-SRT (out of 100)	82.78(13.95)	30.56 – 100	0(1)	-3.11 – 1.32	24.78(17.76)	1.39 – 56.94	-4.16(1.27)	-5.83 – -1.85
AVET (out of 100)	96.63(5.81)	73.96 –100	0(1)	-3.90 – . 58	60.83(19.21)	14.58 – 89.58	-6.16(3.31)	-14.12 – -1.21
ANPT (out of 100)	74.67(24.68)	20 – 100	0(1)	-2.22 – 1.03	21.99(14.97)	0 – 73.33	-2.14(.61)	-3.03 – -.05
MPU	5.35(.97)	3.15 – 7.48	0(0)	-2.27 – 2.20	3.25(.75)	1.89 – 4.61	-2.17(.78)	-3.57 – -.76
CPM (out of 36)	15.89(3.68)	9 – 23	0(1)	-1.87 – 1.94	14.76(3.99)	9 – 23	-.30(1.09)	-1.87 – 1.94

Note. **TD**: Typically Developing. **DLD**: Developmental Language Disorder. ***A-SR****:* Arabic Sentence Repetition Test. ***AVET***: Arabic Verb Elicitation Task. ***ANPT***: Arabic Noun Plurals Test. **MPU**, Mean Morpheme per Utterance***.*** ***CPM***: Colored Progressive Matrices (Ravens, 2007).

Table 4. *Mean percentages of correct nonwords (with standard deviations) of the TD and DLD groups on the CL-NWR task.*

	Group	
	TD	***DLD***
Overall performance	93.61(10.61)	52.22(19.89)***
Nonword length		
One syllable	98.89(4.19)	79.44(23.44)***
Two syllables	95.24(10.49)	53.57(22.02)***
Three syllables	88.17(19)	34(23.43)***
Presence of consonant clusters		
none	94.68(10.29)	66.15(19.75)***
One CC	93.33(11.05)	44(24.36)***
Two CC	88.75(24.54)	23.33(36.51)***
Wordlikenss		
High wordlikeness	98.96(4.77)	80(24.91)***
Low wordlikeness	92.79(11.73)	47.95(20.19)***

Note. **CL-NWR**: Crosslinguistic Nonword Repetition Test. **TD**: Typically Developing. **DLD**: Developmental Language Disorder. **CC**: Consonant Cluster
* $p < .05$, ** $p < .01$, *** $p < .001$

Table 5. *Parameter estimates of the final logistic mixed-effects model (M8)*

Parameters	β	SE (β)	Z statistic
Fixed Effects			
Intercept	.25	1.24	.20
Age	.05**	.02	2.69
Group: TD (compare with DLD)	3.48***	.42	8.21
Nonword length:2 Syllables (compared with 1 syllable)	-.96*	.46	-2.10
Nonword length: 3 syllables (compared with 1 syllable)	-2.48***	.47	-5.29
CC: 1 CC (compared with no CC)	-1.02**	.35	-2.93
CC:2 CC (compared with no CC)	-2.34***	.71	-3.32
Wordlikeness: low wordlikeness (compared with high wordlikeness)	-1.25*	.52	-2.39
Group X CC interaction			
Group: TD x CC number: 1 CC	1.01**	.32	3.14
Group: TD x CC number: 2 CC	.76	.53	1.44
Random Effects	**Variance**	**SD**	
Participant (Intercept)	2.18	1.48	
Item (Intercept)	.57	.76	
Observations 2730, participants: 90, items: 30			

Note. **TD:** Typically Developing. **DLD**: Developmental Language Disorder. **CC**: Consonant Cluster.
* = $p < .05$, ** = $p < .01$, *** = $p < .001$

Click here to access/download
Appendix
20-00556R2 Appendix 1 19.12.2020.docx