Sentence repetition as a clinical marker of developmental language disorder: evidence from Arabic

ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Sentence Repetition as a Clinical Marker of Developmental Language Disorder: Evidence From Arabic

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

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

ARTICLE INFO

Article History:
Received April 30, 2021
Revision received July 9, 2021
Accepted August 1, 2021

Editor-in-Chief: Stephen M. Camarata
Editor: Megan York Roberts

https://doi.org/10.1044/2021_JSLHR-21-00244

ABSTRACT

Purpose: Research on the typical and impaired grammatical acquisition of Arabic is limited. This study systematically examined the morphosyntactic abilities of Arabic-speaking children with and without developmental language disorder (DLD) using a novel sentence repetition task. The usefulness of the task as an indicator of DLD in Arabic was determined.
Method: A LITMUS (Language Impairment Testing in Multilingual Settings) sentence repetition task was developed in Palestinian Arabic (LITMUS-SR-PA-72) and administered to 30 children with DLD (M = 61.50 months, SD = 11.27) and 60 age-matched typically developing (TD) children (M = 63.85 months, SD = 10.16). The task targeted grammatical structures known to be problematic for Arabic-speaking children with DLD (language specific) and children with DLD across languages (language independent). Responses were scored using binary, error, and structural scoring methods.
Results: Children with DLD scored below TD children on the LITMUS-SR-PA-72, in general, and in the repetition of language-specific and language-independent structures. The frequency of morphosyntactic errors was higher in the DLD group relative to the TD group. Despite the large similarity of the type of morphosyntactic errors between the two groups, some atypical errors were exclusively produced by the DLD group. The three scoring methods showed good diagnostic power in the discrimination between children with DLD and children without DLD.
Conclusions: Sentence repetition was an area of difficulty for Palestinian Arabic–speaking children with DLD. The DLD group demonstrated difficulties with language-specific and language-independent structures, particularly complex sentences with noncanonical word order. Most grammatical errors made by the DLD group resembled those of the TD group and were mostly omissions or substitutions of grammatical affixes or omissions of function words. SR appears to hold promise as a good indicator for the presence or absence of DLD in Arabic. Further validation of these findings using population-based studies is warranted.
Supplemental Material: https://doi.org/10.23641/asha.16968043

Developmental language disorder (DLD) is a condition where the child has significant impairment in understanding and/or using spoken language such that it impairs everyday social functioning and educational progress; this difficulty is not associated with an obvious cause and is likely to persist beyond childhood (Bishop et al., 2017). Research has focused on identifying the psycholinguistic phenotypic markers that are characteristically associated with DLD and can be used as indicators of the disorder (e.g., Rice & Wexler, 1996). These can either be (a) distinct grammatical behaviors that are observed in spontaneous and elicited language, for example, deficits in marking verb tense and agreement in English (e.g., Ash & Redmond, 2014) and

Correspondence to Juhayna Taha: j.taha@pgr.reading.ac.uk. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

omission of articles and object clitics in Spanish and Italian (e.g., Guasti et al., 2016; Jackson-Maldonado & Maldonado, 2017), or (b) poor performance on language-based processing tasks such as nonword (see Schowb et al., 2021) and sentence repetition (SR; see Rujas et al., 2021).

SR tasks have gained traction as reliable screening measures for identifying DLD in monolingual and bilingual children in different languages (e.g., Armon-Lotem & Meir, 2016; Conti-Ramsden et al., 2001; Fleckstein et al., 2018). To date, little is known about the usefulness of SR in identifying DLD in Arabic. This study investigates the morphosyntactic abilities of Palestinian Arabic–speaking children with DLD and their typically developing (TD) peers using a novel SR task. First, we compare the two groups on accuracy and error patterns in the repetition of grammatical structures known to be problematic for children with DLD acquiring Arabic and other languages. Then, we assess the accuracy of SR for discriminating Palestinian Arabic–speaking children with DLD from their TD peers.

SR as a Measure of Morphosyntactic Abilities

The exact mechanisms underlying SR have been debated. The central question has been whether performance on SR tasks reflects linguistic knowledge (Klem et al., 2015; Polišenská et al., 2015) or memory capacity (e.g., Alloway & Gathercole, 2005). Early accounts proposed that if sentence length exceeds the individual's immediate memory, repetition of the stimulus will involve linguistic representations in long-term memory, in addition to short-term memory. Such repetitions are suggested to be filtered through the individual's productive linguistic system (Slobin & Welsh, 1973). This view was supported by later studies, suggesting that when a sentence is long enough to tap into the individual's grammatical system, grammatical reconstruction takes place. Thus, after hearing a sentence, individuals use recently activated lexical items to create a conceptual message of the sentence in short-term memory to regenerate the sentence using morphosyntactic representations they are holding in long-term memory (Lombardi & Potter, 1992; Potter & Lombardi, 1990, 1998). Short sentences, however, are imitated in a parrot-like fashion, exclusively relying on short-memory rather than linguistic competence (Vinther, 2002).

Conversely, Riches (2012) proposed that the roles of short- and long-term memory in SR are not length dependent, but they work effectively together at all sentence lengths. This is supported by evidence showing that when the sentence length is constant, increasing syntactic complexity of sentences results in a greater number of errors in SR (Frizelle & Fletcher, 2014; Kidd et al., 2007; Riches et al., 2010). Also, Riches found that the best predictor of SR was syntactic knowledge, as indexed by a priming task. Similarly, Polišenská et al. (2015) have suggested that SR is more dependent on morphosyntax and lexical phonology and less so on semantics or prosody. Together, these findings support the view that SR taps into underlying syntactic competence. It is generally agreed that children find it difficult to imitate structures that they do not know (Devescovi & Caselli, 2007) and that there is an overlap between SR errors and errors made in spontaneous contexts (Riches, 2012). This makes SR a valuable tool for evaluating grammatical structures that might not otherwise be present in spontaneous speech (Seeff-Gabriel et al., 2010) and in characterizing the typical and impaired acquisition of linguistic structures in a given language.

Diagnostic Accuracy of SR Tasks

The quality of a clinical marker as an indicator of the presence or absence of DLD can be determined based on diagnostic accuracy metrics. Sensitivity refers to the proportion of children with the disorder (i.e., with DLD) correctly identified by the task, and specificity refers to the proportion of children without a disorder (i.e., TD) correctly identified by the task. Plante and Vance (1994) recommend that sensitivity and specificity values of 90% and above indicate good classification accuracy of the test, values of 80% to 89% indicate fair diagnostic accuracy, and values below 80% indicate unacceptably high rates of misidentification. Alternative measures of diagnostic accuracy include positive likelihood ratio (LR+), that is, the probability of being correctly identified as having DLD if the child has DLD, and negative likelihood ratio (LR−), that is, the probability of being correctly identified as unimpaired if the child has typical language (Sackett et al., 1991). Likelihood ratios have an advantage over sensitivity and specificity because they are less likely to change due to variations in the prevalence of the disorder (Dollaghan & Campbell, 1998). Dollaghan (2007) suggested that values of LR+ $\geq$ 10.0 and LR− $\leq$ 0.1 indicate that the test can indicate, with confidence, the presence or absence of the disorder; values of LR+ $\geq$ 3.0 and LR+ $\leq$ 0.3 indicate that the test is suggestive but insufficient to rule in or rule out the disorder; and values of LR+ $<$ 3.0 and LR− $>$ 3.0 indicate the test does not discriminate between the presence or absence of the disorder.

SR has been shown to be a reliable clinical marker of DLD in English-speaking children (for a review, see Pawłowska, 2014). Conti-Ramsden et al. (2001) found that SR, compared to a third-person singular task, past tense marking, and nonword repetition, was the most accurate in identifying English-speaking children with DLD aged 10;5 to 11;1 (years;months), with sensitivity and specificity values of 90% and 85%, respectively. More recently, Redmond et al. (2019) revealed that SR discriminated 7-year-old English-speaking children with and without DLD with sensitivity and specificity values greater than 80%,

indicating the potential of the task as a diagnostic tool for DLD (Redmond et al., 2019). Several studies have examined the diagnostic accuracy of SR in identifying children with DLD who speak languages other than English (for a summary, see Table 1). The sensitivity and specificity of SR tasks in most cross-linguistic studies varied between 80% and 90%, indicating fair to good levels of accuracy in discriminating between children with and without DLD.

DLD in Arabic: Characteristics of Morphosyntactic Deficits

Arabic-speaking children with DLD have difficulties with verb morphology production (e.g., Abdallah & Crago, 2008; Fahim, 2017; Taha et al., 2021). In a recent study, Taha et al. (2021) reported that 4- to 7-year-old Palestinian Arabic–speaking children with DLD were significantly less accurate than their age-matched TD peers in producing the following forms: past tense masculine singular verbs (e.g., *daras,* study-PAST-3MS, "he studies"), past tense feminine singular morpheme *–at* (e.g., *daras**at,*** study-PAST-3FS, "she studied"), past tense plural morpheme *–u* (e.g., *daras**u***, study-PAST-3PL, "they studies"), present tense masculine singular morpheme *byi–* (e.g., ***byi**drus,* study-PRES-3MS, "he is studying"), present tense feminine singular morpheme *bti–* (e.g., ***bti**drus,* study-PRES-3FS, " she is studying"), and present tense plural circumfix morpheme *byi—u* (e.g., ***byi**drus**u**,* study-PRES-3PL, "they are studying"). The tense errors of the DLD group resembled the use of finite (i.e., wrong tense) or nonfinite/tenseless forms (i.e., imperative and imperfective verbs) in place of the correct tense. The pattern of subject–verb agreement errors was the use of the singular verbs in place of the plural verbs and the use of the masculine verbs in place of the feminine verbs.

Compared to age-matched TD children, Arabic-speaking children with DLD exhibit difficulties with inflecting Arabic noun plurals (Abdallah et al., 2013; Fahim, 2005; Shaalan, 2010). This includes the use of suffixes for the regular masculine sound plural (MSP; e.g., the suffix ***-:in*** as in *najjar**i:n*** "carpenters") and feminine sound plurals (e.g., the suffix ***-a:t*** as in *ward**a:t*** "flowers") and the use of irregular broken plural (BP) forms (e.g., *dafadiʕ* "frogs"). Analysis of error patterns revealed that children with DLD tended to either use a singular noun instead of the plural form (e.g., *tawala* "table" for *tawl**a:t*** "tables") or a nonmorphological form such as a periphrastic expression of number (e.g., *tamanja kalb* "eight dog" for *klab "*dogs") or quantifiers (e.g., *kteer arnab* "many rabbit" for *aranib* "rabbits"; Abdallah et al., 2013).

Another characteristic of DLD in Arabic is the omission of bound pronouns (Abdallah, 2002; Faquih, 2014; Shaalan, 2010). Using an elicitation task, Faquih (2014) found that the production of bound pronouns is impaired in Hijazi Arabic–speaking children with DLD aged 3;2–6;9 compared to TD children. Specifically, Faquih reported that only a few children in the DLD group produced third-person masculine possessive pronouns (e.g., *ktab**o*** book-POSS-3MS, "his book") and feminine singular possessive pronouns (e.g., *ktab**ha,*** book-POSS-3FS, "her book") and failed to produce any third-person plural possessive pronouns (e.g., *kutub**hum,*** book-BP-POSS-3PL, "their book"). Errors made by the DLD group were characterized by pronoun omission or substitution of a bound pronoun with the free possessive pronoun /ħag/ "mine" (Faquih, 2014).

As Examples 1a and 1b illustrate below Arabic has a flexible word order where verb–subject–object (VSO) and subject–verb–object (SVO) structures are commonly used (Mohammad, 2000). Through syntactic movement,

Table 1. Summary of diagnostic accuracy of sentence repetition tasks in identifying developmental language disorder (DLD) across languages.

		TD		DLD					
Reference	Language	*n*	Age (years; months)	*n*	Age (years; months)	Sensitivity %	Specificity %	LR+	LR–
Armon-Lotem & Meir (2016)	Hebrew	38	6;0 (0.17)	14	6;1 (0.33)	100[a]	87	7.60	0
Armon-Lotem & Meir (2016)	Russian	20	6;1 (0.17)	14	5;10 (0.25)	86	90	8.57	0.16
Vang Christensen (2019)	Danish	37	7;9 (1.5)	16	7;9 (1.1)	94	97	34.7	0.06
Vang Christensen (2019)	Danish	50	12;5 (0.8)	11	12;3 (1.1)	91	98	45.5	0.09
Stokes et al. (2006)	Cantonese	15	4;1–6;9	14	4;2–5;7	77	97	25.66	0.24
Pham & Ebert (2020)	Vietnamese	194	5;8 (0.4)	10	5;5 (0.3)	90	71	3.13	0.14
Thordardottir et al. (2011)	French	78	4;1–5;11	14	4;6–5;11	92	86	6.46	0.09
Leclercq et al. (2014)	French	34	10.2 (1.4)	34	9.11 (1.2)	97	88	8.08	0.03
Theodorou et al. (2017)	Cypriot Greek	22	4;5–8;7	16	4;11–8;1	75	82	4.11	0.3

Note. TD = typically developing; DLD = developmental language disorder; LR+ = positive likelihood ratio; LR– = negative likelihood ratio.
[a]Sensitivity and specificity and LR values are reported for the best cutoff points.

the object could be moved to a preverbal position, resulting in an object–verb–subject (OVS) or object–subject–verb (OSV) structure. One can add an object clitic to the verb to refer to the fronted object (see 1c). This process is called clitic left dislocation (CLD; Lalami, 1996). The production and comprehension of sentences with CLD are reported to be challenging for Qatari Arabic–speaking children with DLD. Shaalan (2010) found that children with DLD scored significantly lower on sentences with CLD than their age-matched TD peers.

1. (a) *biyakul il-walad buza* [VSO]
   Eat-PRES-3MS the-boy ice-cream
   "the boy is eating ice-cream"
   (b) *il-walad biyakul buza* [SVO]
   The-boy eat-PRES-3MS ice-cream
   "the boy is eating ice-cream"
   (c) *buza biyakul**ha** il-walad* [$OV_{cl}S$]
   Ice-cream eat-PRES-3MS-CL-3FS the-boy
   "ice-cream the boy ate it"

Shaalan (2010) also showed that Qatari Arabic–speaking children with DLD aged 4;10–8;11 scored significantly lower than their age- and language-matched TD peers when repeating subject relative clauses, suggesting that subject relatives may pose a difficulty for Arabic-speaking children with DLD. The task included only one object relative clause, and although the DLD group repeated this item less accurately (35%) than the TD group (77%), more evidence is needed to determine whether this form is problematic for Arabic-speaking children with DLD. Examples of subject and object relatives in Palestinian Arabic (PA) are provided in Examples 2a and 2b, respectively.

2. (a) *hay il-binit illi fafat il-arnab* [subject relative]
   This the-girl that see-PAST-3FS the-rabbit
   "this is the girl that saw the rabbit"
   (b) *hay il-bisse illi il-sulħafa 3adatha* [object relative]
   This the-cat that the-turtle bite-PAST-3FS-CL-3FS
   "this is the cat that the turtle bit"

Recently, Wallan (2018) developed two SR tasks: a novel SR targeting grammatical structures in Arabic and an anomalous SR (ASR) test, including sets of semantically anomalous and syntactically anomalous sentences. The tasks were administered to a group of Najdi Arabic–speaking TD children between ages 2;6 and 5;11 and a group of children with reported language concerns (LC). The LC group performed poorly on the SR and ASR tasks relative to age and nonverbal IQ–matched TD children. Wallan also found that the SR task correctly identified 81% of children with LC and 93% of TD children. Although the SR had a good level of accuracy in discriminating children with and without LC, the results should be considered in light of the study caveats. None of the children in the LC group were clinically assessed or had a confirmed diagnosis of DLD. It is unclear whether the language difficulties of the LC group were associated with other comorbidities or differentiating conditions (e.g., hearing loss), which could have contributed to the poor performance on the SR tasks. Thus, the diagnostic accuracy of SR in identifying DLD in Arabic remains unknown.

This Study

Although existing findings on the morphosyntactic difficulties in Arabic DLD are informative, they remain preliminary. Most of the studies included small numbers of children with DLD (e.g., $N = 14$ in Faquih, 2014, and Taha et al., 2021; $N = 12$ in Abdallah et al., 2013; and $N = 10$ in Abdallah & Crago, 2008). In some studies, the number of items used to examine the target grammatical structures was very limited (e.g., object relatives and passives in Shaalan's, 2010, study were only assessed using one item each). While two studies have shown that Arabic-speaking children with language impairment (as a group) perform poorly on SR tasks (Shaalan, 2010; Wallan, 2018), the diagnostic accuracy of the task in discriminating between children with and without DLD at the individual level is yet to be established.

There is a scarcity of norm-referenced tests that are available in Arabic (see *Évaluation du langage oral chez l'enfant libanais* for Lebanese Arabic, [Zebib et al., 2019] and Arabic Language: Evaluation of Function for Gulf-Arabic, [Rakhlin et al., 2021]). In Arabic-speaking contexts, speech and language therapists (SLTs) rely on informal assessment tasks (i.e., parental interview, language sample analysis) to establish DLD diagnosis. Thus, diagnostic decisions are not always consistent and vary according to the subjective judgment and clinical experience of the SLTs. Consequently, Arabic-speaking children with DLD continue to be at risk of being under-misdiagnosed. Tasks with good discriminatory power are needed to help facilitate the effective and efficient identification of DLD in Arabic. Accordingly, this study aims to examine the potential of SR as a clinical marker of DLD in Arabic-speaking children. We specifically address the following questions:

1. How do Arabic-speaking children with DLD compare to TD children in terms of their performance accuracy on SR?
2. How do Arabic-speaking children with DLD compare to TD children in terms of the quantity and quality of their grammatical errors in SR?
3. What is the diagnostic accuracy of the SR for the identification of DLD in Arabic?

We predict that the scores of the DLD group on the SR task will be significantly lower than those of TD children. We also predict that the SR task will show good accuracy in differentiating between children with and without DLD. Based on Riches's (2012) findings that errors in SR corresponded to errors made in other production tasks, we expect the morphosyntactic errors made by the DLD group to mirror those reported in the Arabic literature for children with DLD in elicited or spontaneous language samples.

Method

Participants

This study received approval from University of Reading Ethics Committee. A total of 90 monolingual Palestinian Arabic–speaking children aged 4;0–6;10 were recruited from Ramallah, Palestine. According to a parental report, none of the children had a history of hearing loss or cognitive, motor, behavioral, or neurological impairments. There were 30 children (22 boys and eight girls) with DLD, aged between 4;0 and 6;10 (M = 61.50 months, SD = 11.27), recruited through five private speech and language therapy clinics. These children received a diagnosis of DLD by qualified SLTs and were enrolled in language intervention sessions at the time of the study. Given that the DLD diagnosis was based on informal assessments, it was imperative to confirm that these children met the criteria for DLD (Bishop et al., 2016, 2017). Screening of each child's clinical reports was done to confirm that they had (a) language difficulties affecting one or more language aspects (children with expressive phonological difficulties were included only if they also had difficulties in other language domains, e.g., morphosyntax, semantics), (b) passed hearing tests, and (c) had language disorder that was not associated with any differentiating conditions (e.g., neurological or genetic disorders). There were 60 TD children (33 boys and 27 girls) aged 4;0–6;8 (M = 63.85 months, SD = 10.16). They were recruited through three kindergartens and two schools. The additional inclusion criteria for this group were (a) no parental concerns about the child's current language skill and (b) no history of language delay or intervention. Each TD child was within 2 months of age of a child with DLD. The two groups were matched on chronological age, $t(53.04) = -0.96$, $p = .34$, $d = 0.22$, and did not differ in their nonverbal abilities as measured by the Colored Progressive Matrices (CPM; Raven, 2007), $t(51.59) = -1.26$, $p = .214$, $d = 0.29$.

A battery of standardized language tasks was administered to all children to confirm their diagnostic status. The tasks examined language areas known to be problematic for Arabic-speaking children with or at risk of DLD. The tasks included (a) the Arabic Verb Elicitation Test (AVET), a picture-naming task that examines the production of verb tense and agreement inflections; (b) the Arabic Noun Pluralization Test (ANPT), an elicitation task that examines the production noun plural types; (c) the Arabic version of the Quasi-Universal LITMUS (Language Impairment Testing in Multilingual Settings) Nonword Repetition Test (Abi-Aad & Atallah, 2020; dos Santos et al., n.d.), a task that examines the repetition of nonwords with minimal language-specific features. Additionally, we calculated the (d) mean morpheme per utterance (MPU). MPU is a measure of a child's grammatical ability level in semitic languages (Dromi & Berman, 1982). A narrative sample was obtained for each child using *Frog, Where Are You?* (Mayer, 1969), and the first 100 utterances were transcribed. MPU scores were calculated according to guidelines adopted by Shaalan and Khater (2006) for Arabic. The MPU is derived by diving the total number of morphemes by 100, that is, the number of utterances produced in the narrative task. The results of the TD group (mean and standard deviation) were used to obtain z scores for all participants. All children with DLD scored at or below −1.5 SDs below the mean on at least three of the language measures. All TD children scored above the −1.5 SD cutoff point on at least three language measures. The raw and standardized scores of the TD and DLD groups on the language measures are presented in Table 2. The average raw scores of the DLD group were significantly below those of the TD group on the AVET, $t(31.67) = -9.98$, $p < .001$, $d = 2.52$; the ANPT, $t(84.58) = -12.56$, $p < .001$, $d = 2.58$; QU-LITMUS-NWRT, $t(37.23) = -10.73$, $p < .001$, $d = 2.62$; and MPU, $t(72.49) = -11.28$, $p < .001$, $d = 2.42$.

SR Task

The SR task was designed following the principles of the COST Action IS0804 LITMUS (Armon-Lotem et al., 2015). According to Marinis and Armon-Lotem (2015), SR tasks should include grammatical constructions that are vulnerable for children with DLD in the target language (i.e., language specific) and syntactically complex structures, which are problematic for children with DLD across languages (i.e., language-independent structures). Based on the available research on DLD in Arabic, the language-specific structures were tense and verb agreement morphology (Abdallah & Crago, 2008; Fahim, 2017; Taha et al., 2021), noun plural morphology (Abdallah et al., 2013), and bound possessive pronouns (Faquih, 2014). The language-independent structures were syntactically complex sentences and included passives, sentences with CLD, object *wh*-questions, subject and object relative clause, sentences with subordination, and conditionals.

Table 2. A summary of the raw and *z* scores of the typically developing (TD) and developmental language disorder (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-LITMUS-NWR (out of 100)	93.79 (10.47)	40–100	0.02 (0.99)	–5.05 to –60	52.16 (19.91)	3.33–86.67	–3.90 (1.87)	–8.50 to –.65
AVET (out of 100)	96.63 (5.81)	73.96–100	0 (1)	–3.90 to 0.58	60.83 (19.21)	14.58–89.58	–6.16 (3.31)	–14.12 to –1.21
ANPT (out of 100)	74.67 (24.68)	20–100	0 (1)	–2.22 to 1.03	21.99 (14.97)	0–73.33	–2.14 (0.61)	–3.03 to –0.05
MPU	5.35 (0.97)	3.15–7.48	0 (0)	–2.27 to 2.20	3.25 (0.75)	1.89–4.61	–2.17 (0.78)	–3.57 to –0.76
CPM (out of 36)	15.89 (3.68)	9–23	0 (1)	–1.87 to 1.94	14.76 (3.99)	9–23	–0.30 (1.09)	–1.87 to 1.94

Note. A-QU-LITMUS-NWR = Arabic version of the Quasi-Universal LITMUS Nonword Repetition Test (dos Santos et al., n.d.); AVET = Arabic Verb Elicitation Test; ANPT = Arabic Noun Plurals Test; MPU = mean morpheme per utterance; CPM = Colored Progressive Matrices (Raven, 2007).

Additionally, the task included biclausal sentences with coordination and complementizes, which were syntactically simple control structures matching the syntactically complex sentences (i.e., language independent) in length (Marinis & Armon-Lotem, 2015).

According to LITMUS-SR guidelines, sentences should be grouped into levels according to their length and syntactic complexity. Essentially, language-specific structures were assessed using syntactically simple (e.g., SVO structure) and short sentences (average of eight syllables). The language-specific targets emerge early in development and are evident in the language of 4-year-old Arabic-speaking TD children (e.g., Abdallah et al., 2013; Abdu & Abdu, 1986; Al-Akeel, 1998; Aljenaie, 2001; Omar, 1973; Ravid & Farah, 1999). Hence, all language-specific structures were included in Level 1. No data were available on the acquisition of the language-independent structures in Arabic. Therefore, the assignment of these structures to levels of difficulty followed the design of other LITMUS-SR tasks. This was done to ensure that our task was comparable to other SR tasks in other languages.

The initial version of the task was piloted with an additional group of 13 monolingual Palestinian Arabic–speaking TD children aged 4;1–6;5 (M = 62.4 months, SD = 7.44). These children were not included in the main TD group of this study. Pilot findings revealed that the repetition accuracy of the target structures ranged from 54% to 100%. The average repetition accuracy differed significantly across levels, $F(1, 43) = 41.38$, $p < .001$. The average accuracy of repeating Level 1 structures (M = 94.16, SD = 5.18) was significantly lower than that of Level 2 (M = 76.67, SD = 11.22) and Level 3 (M = 68.83, SD = 18.43); for all comparisons, $p < .05$. These results confirmed that the levels of the task were increasing in difficulty. Although conditional sentences were difficult for the TD children, we decided to retain these items as their repetition accuracy (M = 54.81, SD = 13.96) was above chance level, suggesting that these structures are not yet acquired but are emerging.

The vocabulary (verbs, nouns, and adjectives) used in the task was limited to early-acquired words that were selected from children's story books. Age of acquisition data in Lebanese Arabic were available for only 52 of the words used in the task (Łuniewska et al., 2019; see Supplemental Material S1). As an additional measure, the list of words was judged as being appropriate for preschool-age children by five kindergarten teachers: All words included in the test received an overall agreement score of 80% or above as being familiar to preschool-age children.

The final version of the Palestinian Arabic LITMUS sentence repetition task (LITMUS-SR-PA-72) consisted of 72 sentences. The task examined a total of 13 structures (20 substructures). The structures were classified into three levels of increasing difficulty. Each level contained 24 sentences. All language-specific structures were included in Level 1: past tense, present tense, noun plurals, and bound possessive pronouns. Levels 2 and 3 included language-independent structures. Level 2 contained movement-derived structures such as passives, object *wh*-questions and sentences with CLD, and control structures (biclausal sentences with coordination and complementizes). Level 3 included structures with embedding, and these were conditionals, subordinate sentences, and subject relatives. We also included object relatives that involve both movement and embedding. The order of sentences within each level was pseudorandomized so that there were no two consecutive sentences of the same structure (for the full list of items, see the Appendix). The sentences varied in length from three to seven words and seven to 15 syllables. There was a significant difference in length across the levels, $F(1, 70) = 60.06$, $p < .001$. Sentences in Level 2 (M = 10.83 syllables, SD = 2.32) and Level 3 (M = 11.75 syllables, SD = 1.45) did not differ significantly in length ($p = .18$) but were

significantly longer than sentences in Level 1 (M = 7.92 syllables, SD = .88, $p < .001$).

Procedure

Each child was tested individually in a quiet room in the kindergarten, school, or speech and language therapy clinic they attended. The children were participating in a larger research project and were assessed using a battery of tests across two 1-hr sessions. In the first session, CPM, a narrative task, ANPT, and LITMUS-SR-PA-72 were administered; in the second session, QU-LITMUS-NWRT, AVET, and nonword discrimination tasks were administered. Testing was performed by the first author, who is a qualified SLT and native speaker of Palestinian Arabic. The administration of the LITMUS-SR-PA-72 followed the procedures suggested by Marinis and Armon-Lotem (2015). Live voice was used given the young age of the participants. Presenting sentences with live voice makes the task more engaging for the children (e.g., Devescovi & Caselli, 2007; Frizelle et al., 2017; Gavarró, 2017) and allows the examiner to build a better rapport with them. The live presentation of the task is more clinically relevant: SR tasks within standardized language tests are presented live by clinicians (e.g., Newcomer & Hammill, 2008; Seeff-Gabriel et al., 2008; Wiig et al., 2013). To achieve a consistent presentation of the task for all participants, the examiner practiced reading the sentences at an average speed. Sentences were presented according to their level of difficulty, with sentences in Level 1 being presented first, then sentences in Levels 2 and 3, respectively. The LITMUS-SR-PA-72 task was introduced using a tower-building game. Children were given a bucket full of colored blocks. They were instructed to listen carefully to each sentence and to repeat it verbatim. Two practice sentences preceded the task, and the child was given feedback on their repetitions to ensure their understanding of the task. The examiner read each sentence individually and only once. The sentence was read again if the child did not hear it due to ambient noise or if being distracted. After each repetition of the experimental sentences, the child was verbally praised (e.g., good job) and was allowed to add a block to the tower. The task took approximately 20 min. Responses were audio-recorded for later transcription and scoring.

Scoring

The responses of the children on the LITMUS-SR-PA-72 task were transcribed orthographically, coded and scored off-line using different scoring systems as follows:

- Binary scoring: The child received a score of 1 if their repetition is identical to the target sentence and a score of 0 if their repetition contained any omission, substitution, or addition of words and/or affixes of the target sentence. The maximum total binary score was 72.
- Error scoring: A 0–3 scoring scheme was employed based on the number of errors observed in the child's repetition. Identical repetitions of the target sentence were assigned a score of 3, repetitions containing one error were assigned a score of 2, repetitions containing two to three errors were assigned a score of 1, and repetitions containing four or more errors were assigned a score of 0. This yielded a maximum score of 216.
- Structure scoring: This method was based on whether or not the child maintained the grammatical structure targeted by the sentence. Repetitions containing the target grammatical affix or morphosyntactic structure received a score of 1, whereas repetitions in which the target grammatical structure was omitted, substituted, or changed were considered incorrect and received a score of 0. Compared to the binary and error scoring methods, structural scoring was more lenient as the child's repetition was not penalized for errors that did not affect the structure targeted by the sentence (i.e., lexical substitutions). In all scoring methods, phonological errors that were consistent with the child's speech were not considered errors. Dialectal variations in the repetition of words were also disregarded (e.g., *ke:ka* for *kaʕke*). If the child self-corrected and provided more than one response, their final response was scored irrespective of its accuracy. Errors were disregarded if they did not affect the sentence grammatical structure and included the use of a shortened form of the word (e.g., *ʕa* for *ʕala* "on") and the omission of the relative *illi* from relative clauses. Both errors did not affect the grammatical structure or meaning of the sentences.

Error Analysis

The error analysis was applied to sentences that were ungrammatical (i.e., received a structural score of 0). We did not have predefined error categories, rather, for each ungrammatical sentence, error description was provided and the resulting structure was determined. Example 3 below illustrates the scoring and error coding methods. When repeating Item 46, the child omitted the *wh*-word *mi:n*, which is essential for the formulation of the target object *wh*-question and also omitted the relative pronoun *illi*. Given that the repetition deviated from the target sentence, it received a binary score of 0. There were two omission errors, hence, the error score was 1. Finally, the child failed to repeat the object *wh*-structure correctly, and so, the structural score was 0. In this case, the morphosyntactic error category would be a combined omission of

wh-word and relative *illi* leading to a change of structure (sentence with CLD).

Item 46	*mi:n il-be:bi illi taʕmato mama?* [object *wh*-question] Who the-baby that feed-PAST-3FS-RES-3MS mom? "who is the baby that mommy fed?"
DLD–3	*il-be:bi taʕmato mama* The-baby feed-PAST-3FS-CL-3MS mom? "..the baby mom fed (it)?"
Binary score	0
Error score	1
Structural score	0
Error type	Omission of *wh*-word *mi:n* and relative pronoun *illi* (1)
Actual production	Sentence with left clitic dislocation

Reliability

A second Arabic-speaking SLT independently scored 22% of the data (seven DLD and 11 TD). The intraclass correlation coefficient (ICC; absolute) indicated a high interrater reliability for the binary (ICC = .98), error (ICC = .91), and structural (ICC = .89) scoring methods. Within each level, items of each grammatical structure were equally divided across odd and even items. The odd–even split-half reliability was determined, and the resulting Spearman–Brown coefficient was .96. Furthermore, the Cronbach's alpha for all test items was valued at .985. Both values indicate that the LITMUS-SR-PA-72 had a satisfactory level of internal consistency reliability.

Statistical Analyses

All statistical analyses were run using R software (Version 4.0.3; R Core Team, 2020). Raw scores were used for first and second analyses (to address the first and second research questions, respectively). Percentage scores were used for the third analysis to address the third research question.

To address the first research question, accuracy scores of the TD and DLD groups on the task were compared. A series of generalized linear mixed models (Baayen et al., 2008) were fitted to the data using the lme4 package (Bates et al., 2015). The dependent measure was the accuracy of the grammatical structure of each of the child's repeated sentences. This was a binomial categorical variable (two levels: 1 = *correct*, 0 = *incorrect*). We entered age and sentence length as covariates. The predictors were group, level, target structure, and their interactions. A stepwise–step-up procedure was followed for building the mixed-effects models. The random effects were determined initially. First, we included by-participant and by-item random intercepts. This was done to account for the nonindependence of the data (repeated measures; Baayen et al., 2008). The addition of random slopes of the within-subject variables was considered as recommended by Barr et al. (2013). However, their inclusion led to model nonconvergence. Hence, the models did not include any random slopes. We compared a baseline generalized linear model without random effects (null model) with a baseline mixed-effects model that only included crossed random effects for items and participants. The latter model had a significantly better fit to the data, AIC = 2,991, $\chi^2(2) = 2{,}750$, $p < .001$. Hence, the inclusion of the random effects structures was warranted. Next, the covariates and the fixed effects and their interactions were entered incrementally to the baseline mixed-effects model. Likelihood ratio tests (using a chi-square statistics) were conducted to evaluate whether the inclusion of a fixed effect significantly improved the model's fit statistics (Meteyard & Davies, 2020). Only the fixed effects that significantly improved the model fit were retained in the model. Significant interactions were followed, with pairwise comparisons using Bonferroni correction. These were obtained by the emmeans package (Lenth et al., 2020).

To address the second research question, TD and DLD groups were compared with regard to the types and frequency of errors they made when they did not succeed in producing the target grammatical structure. For each error type, the differences in error rates between TD and DLD groups were examined using Mann–Whitney test.

To address the third research question, we assessed the diagnostic accuracy of the LITMUS-SR-PA-72 task. Receiver operating characteristic (ROC) curve was generated using the pROC package (Robin et al., 2011). ROC curves plot true-positive rate (sensitivity) as a function of false-positive rate (1 – specificity) for all possible cutoff points (Gonçalves et al., 2014) and the optimal cutoff score with the best sensitivity and specificity trade-off is determined. The area under the ROC curve (AUC) was computed, and it is a measure of test classification accuracy. Carter et al. (2016) indicate that AUC values could range from .5 to 1.0. An AUC of 1.0 reflects a perfect test, values of .90–.99 refer to an excellent test, values of .8–.89 indicate a good test, values of .7–.79 refer to a fair test, and any values lower than this indicate that the test is uninformative.

Results

Analysis 1: Performance Accuracy

Figure 1 illustrates the average percentage scores of children with and without DLD on LITMUS-SR-PA-72. The DLD group scored significantly lower than the TD group using binary, $t(34.51) = -12.17$, $p < .001$, $d = 3.02$;

error, $t(31.71) = -11.03$, $p < .001$, $d = 2.79$; and structural, $t(31.03) = -10.08$, $p < .001$, $d = 2.56$, scoring methods.

Unlike binary and error scoring methods, structural scoring did not penalize the child for repetition errors that did not alter the grammatical structure assessed by the sentence. The structural scores index the child's ability to repeat the target grammatical structures, irrespective of their ability to exactly imitate all the words in the sentence. Given our focus is the children's grammatical ability, structural scores were used in the first and second analyses to investigate differences between the TD and DLD groups in repeating sentences of increasing grammatical complexity. A summary of structural scores for TD and DLD on the LITMUS-SR-PA-72 is displayed in Table 3.

The fit of the final model was significantly better than the intercept-only baseline model, AIC = 2,757, $\chi^2(17) = 268$, $p < .001$. The results of the final model are presented in Table 4.

The inclusion of age, $\chi^2(1) = 6.89$, $p < .01$, and sentence length, $\chi^2(1) = 36.7$, $p < .001$, significantly improved the model fit. As shown in Table 4, age did not significantly predict SR performance ($\beta = 0.03$, $p = .171$) but sentence length did ($\beta = -0.20$, $p < .01$). As the length of the target sentence increased (number of syllables), children were less likely to repeat it correctly. There was a main effect of group, $\chi^2(1) = 104$, $p < .001$, such that the TD group (estimated marginal means [*EMM*] = 5.50, *SE* = .25) repeated sentences more accurately than the DLD group (*EMM* = .70, *SE* = .27, $p < .001$). The Group × Age interaction, $\chi^2(1) = 2.15$, $p = .14$, and the Group × Sentence Length interaction, $\chi^2(2) = 2.86$, $p = .24$, were nonsignificant. There was a main effect of level, $\chi^2(1) = 21.6$, $p < .001$. The Level × Group interaction was not significant, $\chi^2(1) = 0.58$, $p = .44$. The Age × Level interaction was significant, $\chi^2(1) = 10.5$, $p < .01$. As shown in Figure 2, repetition accuracy of structures in all levels increased with age, but this effect was more prominent in Levels 2 and 3 compared to Level 1. When age was controlled, the proportion of correctly repeated structures in Level 1 (*EMM* = 4.62, *SE* = 0.3) was higher than that of structures in Level 2 (*EMM* = 2.56, *SE* = 0.23, $p < .001$) and Level 3 (*EMM* = 2.23, *SE* = 0.26, $p < .001$). There was no significant difference in the proportion of correctly repeated sentences between Levels 2 and 3 ($p = .325$).

Grammatical target had a significant effect on repetition accuracy, $\chi^2(11) = 88.3$, $p < .001$. Within Level 1, the repetition accuracy of present tense verbs (*EMM* = 2.98, *SE* = 0.38) was significantly lower than that of possessive pronouns (*EMM* = 6.28, *SE* = 0.54), noun plurals (*EMM* = 5.32, *SE* = 0.43), and past tense (*EMM* = 4.44, *SE* = 0.39; all comparisons, $p < .001$).

Within Level 2, the probability of correct repetition did not differ significantly across sentences with CLD (*EMM* = 3.10, *SE* = 2.35), complements (*EMM* = 3.32 *SE* = 0.38), coordination (*EMM* = 2.94 =, *SE* = 0.48), and object *wh*-questions (*EMM* = 2.03, *SE* = 0.28; for all comparisons, $p > .05$). The repetition accuracy of passive sentences (*EMM* = 1.40, *SE* = 0.35) did not differ from that of object *wh*-questions ($p =. 907$) or sentences with coordination ($p = .269$) but was significantly lower than for sentences with complements or coordination (both comparisons, $p < .05$).

Figure 1. Accuracy scores of the typically developing (TD) and developmental language disorder (DLD) groups across binary, error, and structural scoring methods on the Palestinian Arabic LITMUS sentence repetition task.

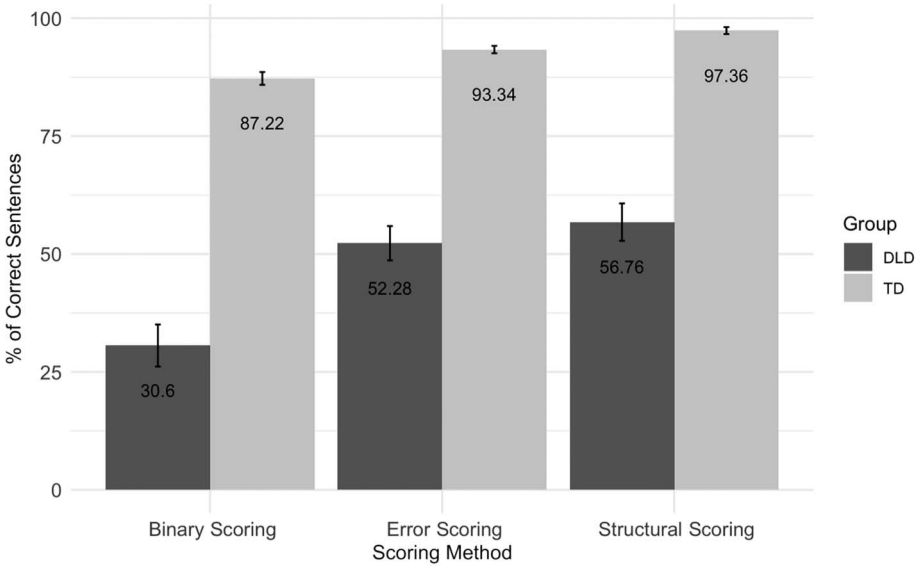


Table 3. Structural scores of the typically developing (TD) and developmental language disorder (DLD) groups across the grammatical targets of the Palestinian Arabic LITMUS sentence repetition task.

Grammatical target	TD *M* (*SD*)	DLD *M* (*SD*)	Significance
Overall performance	97.36 (5.71)	56.76 (21.69)	***
Level 1	99.44 (1.62)	83.33 (13.08)	***
Past tense	99.17 (3.66)	86.67 (17.18)	***
Present tense	98.61 (5.57)	57.78 (33.82)	***
Noun plural	100 (0)	92.22 (12.17)	***
Possessive pronoun	100 (0)	96.67 (8.07)	***
Level 2	96.88 (8.84)	47.22 (25.13)	***
CLD	97.5 (10)	67.5 (25.55)	***
Sentences with complements	98.75 (5.49)	56.67 (34.7)	***
Sentences with coordination	97.08 (8.09)	41.67 (36.16)	***
Wh-object question	95.83 (14.68)	42.92 (36.95)	***
Passive	96.25 (12.02)	31.67 (31.44)	***
Level 3	95.76 (8.10)	39.72 (24.65)	***
Conditional sentences	90.42 (18.46)	23.33 (32.78)	***
Object relatives	95 (11.32)	33.75 (39.41)	***
Subject relatives	98.33 (6.29)	50.83 (44.28)	***
Sentences with subordination	97.5 (9.98)	45.83 (37.76)	***

Note. TD = CLD = clitic left dislocation.
***$p < .001$.

Table 4. Parameter estimates of the final logistic mixed-effects model.

Parameters	β	*SE* (β)	*Z* statistic
Fixed effects			
Intercept	0.35	1.39	0.25
Age	0.03	0.02	1.37
Sentence length	–0.20	0.07	–2.67**
Group: TD (compare with DLD)	4.84	0.36	13.60**
Level 2 (compared with Level 1)	–0.91	0.90	–1.5*
Sentences with complements (compared with CLD)	0.22	0.51	0.42
Conditionals (compared with CLD)	–1.24	0.45	–2.73**
Sentences with coordinates (compared with CLD)	–0.16	0.63	–0.25
Noun plurals (compared with CLD)	2.83	0.44	6.38*
Object relatives (compared with CLD)	–0.95	0.38	–2.49**
Object Wh questions (compared with CLD)	–1.07	0.40	–2.65**
Passives (compared with CLD)	–1.70	0.45	–3.82***
Past tense (compared with CLD)	1.95	0.40	4.92
Possessive pronouns (compared with CLD)	3.80	0.55	6.87***
Subject relatives (compared with CLD)	–0.001	0.39	–0.89
Age × Level	0.02	0.01	1.81***
Random effects	**Variance**	***SD***	
Participant (Intercept)	1.92	1.39	
Item (Intercept)	0.22	0.47	
Observations: 6,480, participants: 90, items: 72			

Note. TD = typically developing. DLD = developmental language disorder; CLD = sentences with clitic left dislocation.
*$p < .05$. **$p < .01$. ***$p <. 001$.

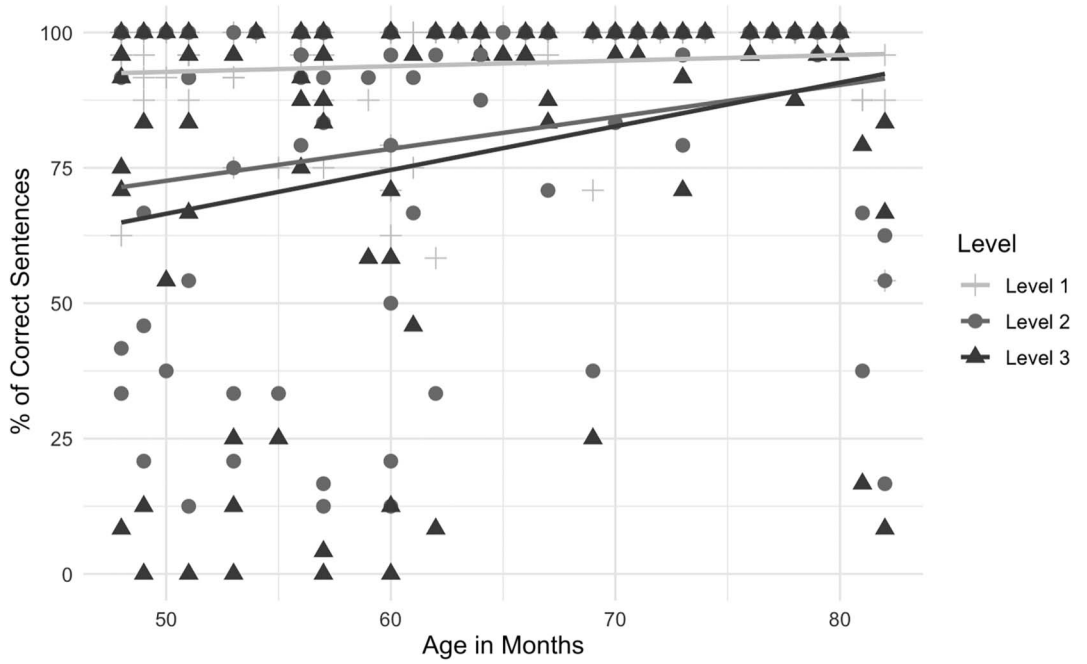
Within Level 3, there were no significant differences between the repetition accuracy of sentences across sentences with conditionals ($EMM = 1.45$, $SE = 0.45$), subject relatives ($EMM = 2.78$, $SE = 0.28$), object relatives ($EMM = 1.83$, $SE = 0.29$), and sentences with subordination ($EMM = 2.78$, $SE = 0.38$; for all comparisons, $p > 0.05$). The Group × Grammatical Target interaction was not significant, $\chi^2(12) = 12.1$, $p = .523$. The proportion of correct sentences repeated by the TD group was significantly higher than that of the DLD group across all of the target structures (for all comparisons, $p < .001$).

We conducted an additional analysis to tease apart the effects of length (indexed by the number of syllables in the sentence) and grammatical complexity (indexed by level of complexity) on repetition accuracy. Levels 2 and 3 only differed in grammatical complexity but did not differ significantly in length. Therefore, we conducted a repeated-measures analysis of variance (ANOVA) with repetition accuracy as the dependent variable, group as a between-subjects variable, and level (i.e., Levels 2 and 3) as a within-subjects variable. The main effects of group, $F(1, 88) = 179$, $p < .001$; level, $F(1, 88) = 6. 45$, $p < .05$; and their interaction, $F(1, 88) = 5.67$, $p < .05$, were all significant. To unpack the interaction, post hoc tests were conducted using Bonferroni corrected p values to account for multiple comparisons (Field, 2009, p. 373). The TD group achieved significantly higher scores than the DLD group in repeating sentences within Level 2 ($p < .001$) and Level 3 ($p < .001$). Within the TD group, there was no significant difference in the average repetition scores of sentences in Levels 2 and 3 ($p = 1$). In contrast, the DLD group scored significantly lower on sentences in Level 2 compared to sentences in Level 1 ($p < .001$).

Analysis 2: Error Patterns

Children with DLD were significantly more likely to produce ungrammatical structures relative to the TD group, $\chi^2(1, N = 90) = 1{,}748$, $p < .001$. As illustrated in Table 5, the most common error in the repetition of past tense verbs included the omission of the entire verb from the sentence or substitution of the plural verb with a singular verb (e.g., *ʃirbib*, drink-PAST-3MS, "he drank" for *ʃirbu*, drink-PAST-3PL, "they drank"). These errors also affected present tense verbs. Additionally, when repeating present tense verbs, the DLD group showed an omission of

Figure 2. A scatter plot showing the change in average structural scores of all children with age across Levels 1, 2, and 3 of the Palestinian Arabic LITMUS sentence repetition task.



the present progressive *b-* and/or gender/person agreement prefix *yi/ti* of the present tense verb, resulting in an imperative or imperfective verb (e.g., *tiqra*, read-IMPER-3FS, "she study" or *iqra*, read-IMP-3MS, "you study" for *btiqra*, rea-PRES-3FS, "she is studying"). In the DLD group, the imperative was used more frequently than the imperfective as a substitute of a present tense verb.

When repeating sentences with noun plurals, the DLD group substituted plural nouns with singular or dual nouns (e.g., *ta:be* "ball" or *ta:bte:n* "two balls" for *taba:t* "balls"). As for possessive pronouns, the DLD group showed omissions of the bound pronouns (e.g., *faʕrat "hair"* for *faʕrathum*, hair-CL-3FS, "their hair"). Overall, the TD group made very few errors in Level 1 structures (all comparisons, $p < .05$). The DLD group was much more likely than the TD group ($p < .001$) to omit the passive prefix *in-*, which resulted in changing the passive sentence to an active one (see Example 3b).

3. (a) **Item 34**: *il- ʃubak infataħ min il-hawa* [passive]
The-window open-PASSIVE-3MS by the-wind
"the window got opened by the wind"
(b) **DLD-3**: *ʃubak fataħ hawa* [SVO]
Window open-PAST-3MS wind
"the window opened wind"

As for sentences with CLD, the TD and DLD groups omitted the clitic pronoun, resulting in a sentence with canonical word order (see Example 4b). The frequency of this error was significantly higher in the DLD group compared to the TD group ($p < .001$).

4. (a) **Item 34**: *il-hadiya fatħatha il-binit* [sentence with CLD]
The-gift open-PAST-3FS-CL-3FS the-girl
"the gift, the girl opened it"
(b) **DLD-10**: *fatħat il-hadiya* [SVO]
Open-PAST-3FS the-gift
"[she] opened it"

When repeating object *wh*-questions, the TD and DLD groups demonstrated omissions of different elements which resulted in repeating object *wh*-questions as subject *wh*-question (5b), a sentence with CLD (5c), or a sentence with canonical word order (5d). An atypical pattern that only appeared in the DLD group was omitting several elements of the questions resulting in a fragmented structure (5e). These errors were significantly more frequent in the DLD group (for all comparisons, $p < .001$).

5. (a) **Item 28:** *ani bisse ħamlatha il-binit?* [object *wh*-question]
Which cat carry-PAST-3FS-CL-3FS the-girl?
"which cat did the girl carry?"
(b) **DLD-2:** *ani: ħamlat il-bisse . . .?* [subject *wh*-question]

Table 5. Frequency and types of grammatical errors of the typically developing (TD) and developmental language disorder (DLD) groups on the Palestinian Arabic LITMUS sentence repetition task.

Target structure	Error pattern (actual production)	Group: TD *n*	Group: DLD *n*	*U*	*z*
Level 1					
Past tense	Omission of verb	0	9	1080	–3.37***
	Omission of plural suffix *-u* (singular for plural verb)	0	6	1080	–3.37***
Present tense	Omission of prefix *bti-/byi-* (imperative)	1	36	1370	–5.77***
	Omission of *b-* clitic (imperfective)	2	8	1081	–2.75**
	Omission of verb	0	6	1020	–2.64***
	Omission of plural *-u* (singular for plural verb)	0	5	1110	–3.69***
Noun plural	Omission of plural suffix (singular for plural noun)	0	4	990	–2.20*
	Substitution of plural suffix (use of dual for plural noun)	0	3	990	–2.20*
Possessive pronouns	Omission of possessive pronoun	0	3	1020	–2.64**
Level 2					
Passive	Omission of passive verb prefix *in-* (past tense verb)	9	73	1587	–6.92***
Sentences with CLD	Omission of clitic pronoun and change to word order (main clause)	3	18	1240	–4.62***
	Omission of clitic pronoun and change to word order (SVO)	3	18	1240	–4.62***
	Omission of clitic pronoun (SVO-wrong meaning)	2	13	1201	–4.06***
Object *wh*	Omission of object clitic (subject *wh*)	10	34	1240	–3.84***
	Omission of *wh* + object clitic (main clause)	7	30	885	–.01***
	Omission of *wh* word (CLD)	2	24	1214	–4.4***
	Omission (fragment structure)	0	22	1170	–4.28***
Coordinate	Omission of coordinator (w) (two main clauses)	4	32	1454	–6.01***
	Omission of coordinator (w) + one clause (one main clause)	1	19	1217	–4.44***
Complement	Omission resulting in one main clause	3	42	1474.5	–6.31***
	Omission (fragment structure)	0	4	1020	–2.64**
Level 3					
Subject relative	Omission of demonstrator *had* and relative *illi* (main clause)	3	86	1398	–5.99***
	Omission (fragment structure)	1	9	1095	–1.19***
Object relative	Omission of demonstrator *had*, relative *illi*, and resumptive clitic (main clause)	6	67	1441.5	–6.14***
	Omission of demonstrator had and relative *illi* (CLD)	1	34	1401.5	–6.04***
	Omission of resumptive clitic (subject relative)	17	26	1073	–1.55**
	Omission (fragment structure)	0	12	1110	–3.7***
Conditional	Omission of conditional *iza* (two main clauses)	13	43	1375	–4.72***
	Omission of conditional *iza* and one clause (main clause)	1	20	1337.5	–5.5***
	Omission (fragment structure)	0	11	1110	–3.67***
Subordinate	Omission of subordinate *ʕashan* (two main clauses)	4	29	1281.5	–4.64***
	Omission of subordinate *ʕashan* + one clause (main clause)	2	12	1181	–4.07***
	Omission (fragment structure)	0	12	1080	–3.37***

Note. CLD = clitic left dislocation.

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

Which carry-PAST-3FS the-cat. . .?
"which carried the cat?"

(c) **DLD-13:** *. . .binit ħamlatha bisse* [sentence with CLD]
girl carry-PAST-3FS-CL-3FS cat
"a girl, a cat carried her"

(d) **DLD-20:** *. . .ħamlat bisse. . .* [SVO]
. . .carry-PAST-3FS cat. . .
"[she] carried a cat"

(e) **DLD-3:** *. . .ħamlat. . .* [Fragment]
. . .carry-PAST-3FS . . .
"[she] carried"

The DLD group showed atypical errors by which they either omitted the coordinator *w* or additional parts of sentences with coordination resulting in two (6b) or one clauses. Both groups showed omissions of several parts of the complement sentences, which resulted in one clause. These errors rarely occurred in the TD group (all comparisons, $p < .001$). A further error that was unique to the DLD group only was the omission of several parts of the sentence resulted in a fragmented structure (6d).

6. (a) **Item 40:** *te:ta ʕimlat ʃa:j w baba akal basko:t* [sentence with coordination]
Grandma make-PAST-3MS tea and dad eat-PAST-3MS biscuits
"grandma made tea and dad ate biscuits"

(b) **DLD-1:** *mama ʕimlat ʃa:j baba akal ...* [two main clauses]

Mom make-PAST-3MS tea and dad eat-PAST-3MS . . .
"mom made tea, dad ate"

(c) **DLD-24:** . . . *akal basko:t* baba [SVO]
. . . eat-PAST-3MS biscuits dad
"dad ate biscuits"

(d) **DLD-26:** *tei:ta .. ʃa:j .. baba ..* [fragment]
grandma .. tea .. dad ..

As for the repetition of subject and object relatives, the TD and DLD groups omitted the demonstrative *had,* relative noun *illi* (and the resumptive clitic pronoun of object relatives), resulting in a clause with canonical word order (7b). Atypical errors of the DLD group included omission of demonstrative *had* and relative noun *illi* of object relatives, resulting in sentences with CLD (7c). All of these errors occurred at a significantly higher frequency in the DLD group relative to the TD group (for all comparisons, $p < .001$).

7. (a) **Item 65:** *ha:d il-ʕasi:r illi ʃirbo il-walad* [object relative]
This the-juice that drink-PAST-3MS-RES-3MS the-boy
"this is the juice that the boy drank"

(b) **DLD-3***: . . .il-walad. . . ʃirb il-ʕasi:r* [SVO]
. . .the-boy. . . drink-PAST-3MS the-juice
"the boy drank the juice"

(c) **DLD-12:** . . .*ʕasi:r ʃirbu walad* [sentence with CLD]
. . . juice drink-PAST-3MS-CL-3MS boy
"juice, a boy drank it"

(d) **DLD-5:** *ha:d ʕasi:r ʃirb il-walad* [subject relative]
This the-juice drink-PAST-3MS the-boy
"this juice drank a the boy"

(e) **DLD-7**: .. *ʕasi:r .. walad ..* [fragment]
.. juice .. boy..

As for conditional sentences, the TD and DLD groups omitted the conditional *iza*, which resulted in two main clauses (8b). In some cases, an additional omission of a clause resulted in only one main clause to be produced (8c). Both error types occurred more often in the DLD group than the TD group ($p < .001$).

8. (a) **Item 62:** *iza il-walad byiʃmal il-wadʒib, raħ yruħ ʕala il- ħadi:qa* [conditional]
If the-boy do-PRES-3MS the-homework, will go-IMPER-3MS to the-park
"if the boy does the homework, he will go to the park"

(b) **DLD-7:** . . .*walad yiʃmal wadʒib,. . . yruħ ʕal ħadi:qa* [two main clauses]
. . .boy do-IMPER-3MS homework, . . .go-IMPER-3MS to the-park
"boy do homework, go to park"

(c) **DLD-27:** . . .*ʕimil wadʒib,. . . ħadi:qa* [SVO]
...do-PAST-3MS homework . . .park
"[he] did homework, park"

(d) **DLD-12**: ... *ʕaħ*adi:qa [fragment]
. . .. park

The most common error type in repeating sentences with subordination was the omission of the subordinator *ʕashan*, which resulted in two main clauses (10b). Sometimes this error was associated with an additional omission of either the main or subordinate clause, which resulted in only one clause (10c). Both errors occurred more often in the DLD group relative to the TD group (both comparisons, $p < .001$). The omission of several elements of sentences in Level 3 resulted in fragmented sentences (see Examples 8e, 9e, and 10e). This was an atypical error specific to the DLD group (see Table 5).

9. (a) **Item 58***: il-walad ʕayyat ʕashan dayyaʕ il-luʃbeh* [subordinate]
The-boy cry-PAST-3MS because lose-PAST-3MS the-toy
"the boy cried because [he] lost the toy"

(b) **DLD-22:** *walad ʕayyat. . . dayyaʕ luʃbeh* [two main clauses]
Boy cry-PAST-3MS. . . lose-PAST-3MS toy
"boy cried. . .[he] lost the toy"

(c) **DLD-14** *il-walad. . . dayyaʕ luʃbeh* [SVO]
The-boy . . .lose-PAST-3MS toy
"the boy lost a toy"

(d) **DLD-27**: *..luʃbeh ..walad..* [fragment]
.. toy..boy ..

Analysis 3: Diagnostic Accuracy

Sensitivity, specificity, and likelihood ratios were calculated for the final cutoff scores across the scoring methods are summarized in Table 6.

For the binary scoring method, a cutoff score of 70.14% or below correctly classified 93% of children with DLD (sensitivity) and 93% of TD children (specificity). A child with DLD was 14 times more likely to obtain a "fail" score (i.e., at or below the 70.14% cutoff) on the LITMUS-SR-PA-72 than a TD child, and only 0.07 times

more likely to obtain a "pass" score (i.e., above the 70.14% cutoff) than a TD child.

With a cutoff score of 79.4%, the error scoring method achieved a good sensitivity level of 93% and a good specificity level of 98%. A child with DLD was 54 times more likely to receive a fail score on the task than a TD child, and only 0.07 times more likely to obtain a pass score than a TD child.

Similarly, the structural scoring method achieved a high level of diagnostic accuracy. A cutoff score at 90.97% correctly classified 97% of children with DLD (sensitivity) and 92% of TD children (specificity). A child with DLD was 11 times more likely to obtain a fail score (i.e., below 90.97% cutoff) on the LITMUS-SR-PA-72 compared to a TD child and was only 0.07 times more likely than a TD child to score above cutoff score. The diagnostic accuracy of the LITMUS-SR-PA-72 achieved using the binary scores did not differ from that achieved using the error (p = .09) or structural scores (p = .986). Similarly, there was no significant difference in the diagnostic accuracy of error scores compared to the structural scores (p = .986).

Discussion

The TD and DLD groups differed significantly in their scores on the LITMUS-SR-PA-72, showing that SR is a locus of difficulty for Arabic-speaking children with DLD. The pattern of grammatical errors in the TD and DLD groups were largely similar, with a higher frequency of grammatical errors in the DLD than in the TD group. The LITMUS-SR-PA-72 discriminated accurately between Arabic-speaking children with DLD and their age-matched TD peers.

Arabic-Speaking Children With DLD Performed Poorly on the LITMUS-SR-PA-72

Our first research question addressed how children with and without DLD differ on performance accuracy on the LITMUS-SR-PA-72. As predicted, we found large and significant differences in the average performance of the TD and DLD groups using binary, error, and grammatical structural scoring methods. These findings are in line with previous studies, suggesting that SR is an area of weakness for children with DLD acquiring a variety of languages (e.g., Conti-Ramsden, 2003; Pham & Ebert, 2020; Thordardottir et al., 2011; Vang Christensen, 2019), including Arabic-speaking children with or at risk of DLD (Shaalan, 2010; Wallan, 2018). The average grammatical structural scores on the LITMUS-SR-PA-72 task improved with age, suggesting that the task captured grammatical developmental changes within the age span of 4–6 years. Although the repetition accuracy of the grammatical structures in Level 1 remained stable against age, it increased significantly with age for structures within Levels 2 and 3 for both groups. The grammatical structures tested within Level 1 are acquired by 3 years of age in Arabic, except for MSPs and BPs, which are acquired gradually into school-age years (Abdallah et al., 2013; Aljenaie, 2001; Faquih, 2014; Omar, 1973; Ravid & Farah, 1999). This explains the limited variation between older and younger children in repetition accuracy of Level 1 structures. No data are available in Arabic on the acquisition of structures within Levels 2 (passive, CLD, object *wh*-questions, coordinates, and complements) and 3 (subject relative, object relative, subordinates and conditionals). Evidence from other languages suggests that the acquisition of these structures extends into school age (e.g., Friedmann & Novogrodsky, 2004; Leonard, 1989; Mastropavlou & Tsimpli, 2011; Stromswold, 1995). They could be emerging and not yet fully acquired by the children in our sample, which may contribute to the observed age effect.

The SR accuracy decreased as sentence length increased. Sentences within Levels 2 and 3 did not differ in length but were significantly longer than sentences within Level 1. For both groups, the average SR accuracy scores for Levels 2 and 3 were significantly lower than those for Level 1, but no difference was observed in the average SR

Table 6. Diagnostic accuracy metrics of the Palestinian Arabic LITMUS sentence repetition task.

Scoring method	Cutoff % (raw)	Sensitivity (correct DLD) [95% CI]	Specificity (correct TD) [95% CI]	LR+ [95% CI]	LR [95% CI]	AUC [95% CI]
Binary scoring	70.14% (52/72)	.93 (28/30) [.83, 1]	.93 (56/60) [.87, .98]	13.93 [5.41, 36.26]	0.07 [5.41, 0.27]	.97 [.94, 1]
Error scoring	79.4% (160/216)	.93 (28/30) [.83, 1]	.98 (59/60) [.95, 1]	54.88 [8, 329]	0.07 [0.02, 0.26]	.98 [.96, 1]
Structural scoring	90.97% (65/72)	.97 (29/30) [.83, 1]	.92 (55/60) [.83, 1]	11.27 [84.83, 26.12]	0.07 [0.02, 0.27]	.99 [.97, 1]

Note. DLD = developmental language disorder; CI = confidence interval; TD = typically developing; LR+ = positive likelihood ratio; LR– = negative likelihood ratio; AUC = area under curve.

accuracy between Levels 2 and 3. The reduction of SR accuracy with increasing length could point out the role of short-term memory in SR, with longer sentences placing greater demands on memory capacity than shorter sentences (e.g., Alloway & Gathercole, 2005). However, the observed decline in SR accuracy across levels occurred despite controlling for sentence length in the analysis. Furthermore, sentences within Levels 2 and 3 were not only longer but also syntactically more complex than sentences in Level 1. Hence, the increased difficulty with the repetition of sentences within Levels 2 and 3 relative to Level 1 cannot be attributed solely to differences in short-term memory load but could also reflect differences underlying syntactic representations in long-term memory (Frizelle et al., 2017).

In an attempt to disentangle the influence of length and grammatical complexity on repetition accuracy, we conducted an additional analysis in which we compared the performance of both groups on Levels 2 and 3, which did not differ significantly in length but differed in grammatical complexity. Before discussing the results, we would like to acknowledge that while the results of this analysis (repeated-measures ANOVA) were largely similar to the original mixed-effects model we conducted, there are slight differences. The difference is likely to be due to the increased complexity of mixed-effects model relative to the repeated-measures model. Specifically, the inclusion of random effects structures may reduce the amount of variance that is attributed to fixed effects and their interactions. That is, if fixed effects or their interactions are small or weak, they could appear as being nonsignificant in mixed models. This could explain the lack of Group × Level interaction in the mixed model but the significance of this interaction in this follow-up analysis. Within the TD group, there was no significant difference in performance between Levels 2 and 3. In contrast, the DLD group showed a significantly lower repetition accuracy of sentences in Level 3 compared to Level 2. This finding suggests that syntactic complexity influenced the repetition accuracy in the DLD group, but not the TD group. That is, the DLD group appeared to be more sensitive and found it more challenging to repeat syntactically complex sentences. The lack of this effect in the TD group could be attributed to the fact that their performance approached ceiling across all levels. This result extends previous evidence, showing that when sentence length was constant, increasing syntactic complexity resulted in a greater number of errors in SR (Frizelle & Fletcher, 2014; Kidd et al., 2007; Riches et al., 2010). This conclusion is in line with accumulating evidence, maintaining that SR is not a pure measure of memory but rather requires interaction between linguistic representations and memory resources (e.g., Marinis & Armon-Lotem, 2015; Moll et al., 2015).

We further examined the accuracy of repetition of the target grammatical structures assessed by the task. The DLD group had lower accuracy scores compared to the TD group in producing past tense and present tense verbs. This finding confirms that the production of verb tense and subject–verb agreement morphology is a weakness for Arabic-speaking children with DLD (Abdallah & Crago, 2008; Fahim, 2017; Taha et al., 2021). The children with DLD in our sample repeated noun plurals and possessive pronouns with high accuracy (> 90%), suggesting that these structures were not problematic for them. This contrasts with findings from previous studies that have used elicitation tasks (Abdallah et al., 2013; Faquih, 2014). Importantly, this finding is inconsistent with the results of the DLD group on the ANPT in which they had an average score of 22%. As mentioned in the Method section, the items used in the LITMUS-SR-PA-72 were limited to early-acquired words. Hence, these items had high frequency and were familiar to the children. These findings could be explained by referring to the critical mass hypothesis, which assumes a relationship between lexical development and morphosyntactic skills in children (Marchman & Bates, 1994; Windfuhr et al., 2002). It proposes that once the children have acquired a critical mass of words (i.e., nouns), acquiring morphological properties (e.g., noun plurals and possessive pronouns) would be facilitated. As the children in our study had acquired all the nouns used in the LITMUS-SR-PA-72 task, they did not have much difficulty with the morphological properties of these nouns (i.e., forming plurals or possessive pronouns) as they would have acquired these nouns.

The DLD group had significantly lower scores than the TD group in repeating all language-independent structures within Level 2 (sentences with CLD, passives, and object *wh*-questions) and Level 3 (subject and object relatives, conditionals, and sentences with subordination). This finding is not surprising as the production and comprehension of syntactic constructs that involve movement (e.g., sentences with CLD, passives, object relatives, and object *wh*-questions) have been identified to be cross-linguistically impaired in children with DLD (e.g., Arosio et al., 2009; Bedore & Leonard, 2001; Deevy & Leonard, 2004; Friedmann & Novogrodsky, 2011; Marinis & Saddy, 2013; Novogrodsky & Friedmann, 2006; Prévost et al., 2014; Shaalan, 2010; Tuller et al., 2011). These noncanonical structures are derived via syntactic movement, which involves building long-distance syntactic dependencies. Surface SVO word order corresponds to a canonical order of arguments (agent–action–theme), whereas noncanonical sentences do not. The DLD group's low scores when repeating movement-derived sentences could be attributed to a difficulty in the assignment of a thematic role to the moved element (Friedmann & Novogrodsky, 2011). The DLD group's poor repetition ability of sentences with clausal embedding (e.g., sentences with complements, subject and object relatives, and sentences with conditionals) has also been identified as an

area of difficulty for children with DLD across languages (e.g., Arosio et al., 2009; Fleckstein et al., 2018; Frizelle & Fletcher, 2014; Gavarró, 2017; Owen & Leonard, 2006).

Our study provides an initial picture of the morphosyntactic difficulties of Arabic-speaking children with DLD. We found significant group differences in the production of verbs, sentences with passives, CLD, object *wh*-questions, subject and object relative clauses, sentences with coordination, complements, subordination, and conditionals. These structures appear to be sensitive to the language differences between children with and without DLD and could potentially support the identification of DLD in Arabic. A further investigation of these structures using other probes (e.g., elicitation tasks and language samples) is warranted to better establish their potential as clinical markers of DLD in Arabic-speaking children.

Grammatical Errors

Our second research question focused on how children with and without DLD differ in terms of their morphosyntactic errors on the LITMUS-SR-PA-72 task. The quantity and, in some instances, the type of grammatical errors differed between the TD and DLD groups. The proportion of errors in the DLD group was significantly higher than that of the TD group.

With regard to the language-specific structures, the DLD group showed either omission or substitution errors when repeating present or past tense verbs. Tense errors consisted of replacing the target tense with a nonfinite form (imperative or imperfective).

Given the fusional properties of Arabic verb morphology, errors in tense were sometimes associated with errors in agreement. Main agreement errors were the use of second-person plural verbs in place of third-person verbs (in cases where the imperative was used) or the use of singular verbs instead of plural verbs. These errors were barely produced by the TD group, suggesting that they are age-inappropriate errors; however, such errors have been observed in Arabic-speaking children with DLD and toddler TD children (Abdallah & Crago, 2008; Ouali, 2018; Qasem & Sircar, 2017; Taha et al., 2021).

As for the language-independent structures, omissions were the dominant error type observed in both groups, with a higher proportion of errors in the DLD compared to the TD group. The omission errors primarily affected grammatical suffixes such as passive prefix *-in* from the passive verb, which resulted in producing an active voice sentence. Object clitic pronouns in sentences with CLD, object *wh*-questions, and object relative clauses were omitted, which resulted in a change of the target grammatical structure. Furthermore, omission errors affected function words such as the coordinator *w* "and," conditional *iza* "if," subordinate *ʕashan* "because," demonstrative *ha:d* "this," relative pronoun *illi* "that," and *wh*-words such as *mi:n* "who" and *ani/u* "which." The omission of the grammatical suffixes or function words frequently co-occurred or was associated with word order changes. These errors occurred in the TD group as well, so could be described as typical. In general, the omission error patterns in the DLD group have been observed in other languages, which extends the evidence that the use of grammatical affixes is an area of weakness in children with DLD (e.g., Bedore & Leonard, 1998; Frizelle & Fletcher, 2014; Grüter, 2005; Hansson & Nettelbladt, 2006; Novogrodsky & Friedmann, 2006; Seeff-Gabriel et al., 2010). One atypical error type that occurred exclusively in the DLD group was the omission of several elements of the target sentences, which resulted in a fragmented structure. This particularly applied to sentences involving syntactic movement: passive sentences, sentences with CLD, object *wh*-questions, and object relatives. The repetition of structures involving movement and/or embedding as fragmented structures could indicate poor morphosyntactic representations of these structures in the long-term memory of children with DLD (Frizelle et al., 2017) or that these structures have not yet been acquired.

LITMUS-SR-PA-72 Could Be a Clinical Marker of DLD in Arabic

Our third research question addressed whether the LITMUS-SR-PA-72 task can reliably distinguish children with and without DLD. ROC analyses were performed to obtain the best cutoff points for the binary, error, and structural scoring methods. The AUC levels associated with the optimal cutoff scores ranged from .97 to .99 for the three scoring methods, suggesting that the LITMUS-SR-PA-72 yielded an excellent diagnostic accuracy. This finding is consistent with previous studies showing a good diagnostic accuracy of SR tasks in identifying DLD in many languages (e.g., Armon-Lotem & Meir, 2016; Leclercq et al., 2014; Pham & Ebert, 2020; Vang Christensen, 2019).

Overall, the binary, error, and structural scoring methods showed sensitivity (proportion of children with DLD correctly identified) and specificity values (proportion of TD children correctly identified) were larger than 90%. Hence, scoring the LITMUS-SR-PA-72 test with any of these scoring methods yielded a good power in differentiating between children with and without DLD (Plante & Vance, 1994). Across the scoring methods, the LRs+ were higher than 10 and the LRs− were less than 0.1. These values suggest that a child with DLD was more than 10 times more likely to obtain a "fail" score (i.e., at or below the specified cutoff) on the LITMUS-SR-PA-72 than a TD child and only less than 0.1 times more likely to obtain a "pass" score (i.e., above the specified cutoff) than a TD child. Together, these findings indicate that a

score above or below the specified cutoff point on the LITMUS-SR-PA-72 could be interpreted with strong confidence as indicative of the presence or absence of DLD (C. A. Dollaghan, 2007).

Despite the good levels of sensitivity, specificity, and strong likelihood ratios, these values should be interpreted in consideration of the associated 95% confidence intervals (CIs). The lower bound of the 95% CIs for the sensitivity of the binary and error scoring methods was 83%, and those for the specificity of the binary and structural scoring methods were 87% and 83%, respectively. These values fall below the 90% threshold, which characterizes tests with good diagnostic accuracy; rather, they are only indicative of adequate diagnostic accuracy (Plante & Vance, 1994). Similarly, the lower bound of 95% CIs for the LR+ of the binary and error methods was ≥ 3, whereas the upper bound of 95% CIs for the LR− of all the scoring methods was ≤ 0.3. These values do not meet the criteria of C. A. Dollaghan (2007) for a clinically informative test, that is, of LR+ ≥ 10.0 and LR− ≤ 0.1; rather, they indicate that the test is suggestive but insufficient to rule in or rule out the disorder. Therefore, we refrain from suggesting that the LITMU-SR-PA is a strong indicator of the presence or absence of DLD. Instead, we propose that the LITMU-SR-PA test is suggestive of DLD and should be used in combination with other assessment tools to achieve accurate DLD diagnosis.

Clinical Implications

The LITMUS-SR-PA-72 task, a theoretically based measure, forms the first step toward a more research-informed approach to DLD diagnosis in the Palestinian Arabic context. Our study suggests that poor SR may characterize DLD in Arabic-speaking children. This leads us to recommend SLTs consider assessing SR as part of the diagnostic procedures of DLD. Particularly, the LITMUS-SR-PA-72 could be used to conduct a systematic evaluation of the morphosyntactic structures known to be problematic for children with DLD. These grammatical structures may be avoided by children with DLD in traditional elicitation tasks (i.e., spontaneous language samples), providing fewer opportunities to assess these structures. Our study emphasizes the potential diagnostic value of the LITMUS-SR-PA-72 task as an index with good diagnostic accuracy in differentiating 4- to 6-year-old Palestinian Arabic–speaking children with DLD from TD children. This good discriminatory power of the task was consistent across the binary, error, and structural scoring methods; hence, either of these scoring systems could be applied according to the purpose of the assessment. The binary scoring method is a simplified scoring system that is quick and easy to administer. It could be most useful when the LITMUS-SR-PA-72 is used to determine whether or not a child's linguistic abilities require further assessment. Error scoring is a more fine-grained method and could be used to determine the severity of a child's language production deficits. Grammatical structure scoring is a precise scoring system that could be used to build a profile of the child's grammatical strengths and weaknesses. Morphosyntactic structures that the child fails to repeat could then be further assessed. The structural scoring method could be used to measure the effectiveness of language intervention and progress in the mastery of the target grammatical structures. Unlike informal language tasks, the LITMUS-SR-PA-72 provides clinicians with language scores that could be compared to different cutoff points according to the scoring method that is being used.

Limitations and Future Directions

The two-gate design alongside the stringent criteria employed for the inclusionary language measures could have resulted in a spectrum bias (C. A. Dollaghan & Horner, 2011; Pawłowska, 2014; Redmond et al., 2019). The children in our study were recruited from preselected samples (e.g., children with a prior DLD diagnosis vs. children with typical language development). A confirmation of the DLD/TD status involved scoring above (for TD children) or at/below cutoff -1.5 *SD*s (for children with DLD) on at least three language measures assessing morphosyntax and phonology. This could have resulted in two groups on the polarized ends of the spectrum of language abilities. The comparison of the DLD group with severe language deficits and the TD group with average language abilities could have led to an overestimation of the diagnostic accuracy levels (Pawłowska, 2014). Furthermore, our DLD group may not be representative of Palestinian children with DLD. Given that DLD diagnosis in Palestine is based on informal language assessments, children whose language difficulties are borderline and/or do not present with comorbid speech sound disorder are more likely to be undiagnosed and, consequently, not entitled to receive language intervention services. Children who receive a clinical diagnosis of DLD usually have more severe language deficits. To address these limitations, Pawłowska (2014) recommends employing one-gate designs in which all children are recruited from a single population (unselected sample) so that heterogeneous and representative samples of children with and without DLD are recruited.

The current study examined the clinical usefulness of the LITMUS-SR-PA-72 task in identifying DLD in 4- to 6-year-old children, hence limiting generalizability of the results to older or younger children. Future work could examine the diagnostic value of the task in identifying DLD in a wider age range and also establishing norms for the acquisition of grammatical structures in Arabic. This information is imperative for the development of age-appropriate

grammatical assessments to inform clinicians and educators of the grammatical structures to watch out for when assessing children across different age groups.

The LITMUS-SR-PA-72 was administered live. This could have resulted in variations in pitch, speed, the loudness of the examiner when reading sentences to different children. For a more consistent task delivery, a computerized version of the task using audio-recorded sentences could be developed. Finally, the vocabulary used in the task was not controlled for frequency or imageability. More research is needed to establish psychometric properties of Arabic vocabulary.

Conclusions

This study found that SR deficits could be a potential clinical marker of DLD in Arabic-speaking children. Compared to age-matched controls, the DLD group scored significantly lower on repetition of both language-specific and language-independent syntactically complex structures. The frequency of morphosyntactic errors was significantly higher in the DLD than in the TD group. Some errors occurred exclusively in the DLD group, suggesting clinicians should consider the type and frequency of error patterns when assessing children's expressive grammar. The LITMUS-SR-PA-72 is moderately accurate in differentiating between Palestinian Arabic–speaking children with and without DLD. The task is only suggestive of the presence or absence of DLD and should be used alongside information from other sources to improve the accuracy of DLD diagnosis. The clinical utility of a refined version of the task should be confirmed in a more representative sample of Palestinian children via larger scale population studies.

Acknowledgments

This work was funded by University of Reading International PhD studentship awarded to the first author. We thank all of the children and parents who took part of this study. We also wish to thank the speech and language therapists and teachers, especially Nairouz Nijm Aldeen, who facilitated the recruitment of children for this study. We also like to thank Theo Marinis for his helpful advice and feedback on the coding of error data.

References

Abdallah, F. (2002). *Specific language impairment in Arabic-speaking children: Deficits in morphosyntax* [Doctoral dissertation, McGill University].

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. https://doi.org/10.1017/S0305000912000499

Abdallah, F., & Crago, M. (2008). Verb morphology deficits in Arabic-speaking children with specific language impairment. *Applied Psycholinguistics, 29*(2), 315–340. https://doi.org/10.1017/S0142716408080156

Abdu, D., & Abdu, S. (1986). *Dirasa fi mufradat tiflayn* [A study on child vocabulary]. Dhaat al-Salaasil Publication.

Abi-Aad, K., & Atallah, C. (2020). Nonword repetition: LITMUS-NWR-Lebanon. In R. Zebib, P. Prévost, L. Tuller, & G. Henry (Eds.), *Multilingualism and specific language disorders in Lebanon* (pp. 79–92). Presses Universitaires de l'Université Saint-Joseph.

Al-Akeel, A. (1998). *The acquisition of Arabic language comprehension by Saudi children* [Unpublished thesis]. Newcastle University.

Aljenaie, K. (2001). The emergence of tense and agreement in Kuwaiti Arabic children. *Reading Working Paper in Linguistics, 4,* 1–24.

Alloway, T. P., & Gathercole, S. E. (2005). Working memory and short-term sentence recall in young children. *European Journal of Cognitive Psychology, 17*(2), 207–220. https://doi.org/10.1080/09541440440000005

Armon-Lotem, S., de Jong, J., & Meir, N. (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual Matters. https://doi.org/10.21832/9781783093137

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 & Communication Disorders, 51*(6), 715–731. https://doi.org/10.1111/1460-6984.12242

Arosio, F., Adani, F., & Guasti, M. T. (2009). Grammatical features in the comprehension of Italian relative clauses by children. In J. Brucart, J. M. Gavarró, & A. Solà (Eds.), *Merging features: Computation, interpretation, and acquisition*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199553266.003.0008

Ash, A. C., & Redmond, S. M. (2014). Using finiteness as a clinical marker to identify language impairment. *SIG 1 Perspectives on Language Learning and Education, 21*(4), 148–158. https://doi.org/10.1044/lle21.4.148

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. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192. https://doi.org/10.1044/jslhr.4105.1185

Bedore, L. M., & Leonard, L. B. (2001). Grammatical morphology deficits in Spanish-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 44*(4), 905–924. https://doi.org/10.1044/1092-4388(2001/072)

Bishop, D. 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), Article e0158753. https://doi.org/10.1371/journal.pone.0158753

Bishop, D. 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. (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. https://doi.org/10.1111/jcpp.12721

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery, 159*(6), 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029

Conti-Ramsden, G. (2003). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research, 46*(5), 1029–1037. https://doi.org/10.1044/1092-4388(2003/082)

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. https://doi.org/10.1111/1469-7610.00770

Deevy, P., & Leonard, L. B. (2004). The comprehension of wh-questions in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(4), 802–815. https://doi.org/10.1044/1092-4388(2004/060)

Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language & Communication Disorders, 42*(2), 187–208. https://doi.org/10.1080/13682820601030686

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

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

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54*(4), 1077–1088. https://doi.org/10.1044/1092-4388(2010/10-0093)

dos Santos, C., Ferré, S., Zebib, R., & Abou Melhem, N. (n.d.). *The Non-Word Repetition Task–Lebanese* [Unpublished material]. Adapted from the LITMUS-NWR-FRENCH task.

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. https://doi.org/10.1017/S0305000900004785

Fahim, D. (2005). *Developmental language impairment in Egyptian Arabic* [Unpublished master's thesis]. Birkbeck, University of London.

Fahim, D. (2017). Verb morphology in Egyptian Arabic developmental language impairment. *Arab Journal of Applied Linguistics, 2*(1), 49–73.

Faquih, N. O. (2014). *Production of Arabic bound pronouns by typically developing children and by children with language disorders* [Doctoral dissertation, Howard University].

Field, A. (2009). *Discovering statistics using IBM SPSS statistics*. Sage.

Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition, 25*(1), 85–101. https://doi.org/10.1080/10489223.2016.1192635

Friedmann, N., & Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language, 31*(3), 661–681. https://doi.org/10.1017/s0305000904006269

Friedmann, N., & Novogrodsky, R. (2011). Which questions are most difficult to understand? *Lingua, 121*(3), 367–382. https://doi.org/10.1016/j.lingua.2010.10.004

Frizelle, P., & Fletcher, P. (2014). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders, 49*(2), 255–264. https://doi.org/10.1111/1460-6984.12070

Frizelle, P., O'Neill, C., & Bishop, D. V. M. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language, 44*(6), 1435–1457. https://doi.org/10.1017/S0305000916000635

Gavarró, A. (2017). A sentence repetition task for Catalan-speaking typically-developing children and children with specific language impairment. *Frontiers in Psychology, 8,* 1–13. https://doi.org/10.3389/fpsyg.2017.01865

Gonçalves, L., Subtil, A., Oliveira, M. R., De, P., & Bermudez, Z. (2014). ROC curve estimation: An overview. *REVSTAT-Statistical Journal, 12*(1), 1–20.

Grüter, T. (2005). Comprehension and production of French object clitics by child second language learners and children with specific language impairment. *Applied Psycholinguistics, 26*(3), 363–391. https://doi.org/10.1017/S0142716405050216

Guasti, M. T., Palma, S., Genovese, E., Stagi, P., Saladini, G., & Arosio, F. (2016). The production of direct object clitics in preschool– and primary school–aged children with specific language impairments. *Clinical Linguistics & Phonetics, 30*(9), 663–678. https://doi.org/10.3109/02699206.2016.1173100

Hansson, K., & Nettelbladt, U. (2006). *Wh*-questions in Swedish children with SLI. *International Journal of Speech-Language Pathology, 8*(4), 376–383. https://doi.org/doi.org/10.1080/14417040600880722

Jackson-Maldonado, D., & Maldonado, R. (2017). Grammaticality differences between Spanish-speaking children with specific language impairment and their typically developing peers. *International Journal of Language & Communication Disorders, 52*(6), 750–765. https://doi.org/10.1111/1460-6984.12312

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes, 22*(6), 860–897. https://doi.org/10.1080/01690960601155284

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*(1), 146–154. https://doi.org/10.1111/desc.12202

Lalami, L. (1996). Clitic left dislocation in Moroccan Arabic. *Amsterdam Studies in the Theory and History of Linguistic Science Series, 4,* 115–130. https://doi.org/10.1075/cilt.141.09lal

Leclercq, A. L., Quémart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A powerful diagnostic tool for French children with specific language impairment. *Research in Developmental Disabilities, 35*(12), 3423–3430. https://doi.org/10.1016/j.ridd.2014.08.026

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). *Estimated marginal means, aka least-squares means.*

R package version 1.4.6. https://cran.r-project.org/package=emmeans

Leonard, L. B. (1989). Language learnability and specific language impairment in children. *Applied Psycholinguistics, 10*(2), 179–202. https://doi.org/10.1017/S0142716400008511

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Memory and Language, 31*(6), 713–733. https://doi.org/10.1016/0749-596X(92)90036-W

Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolík, F., Butcher, M., Chondrogianni, V., Hreich, E. K., Messarra, C., Razak, R. A., Treffers-Daller, J., Yap, N. T., Abboud, L., Talebi, A., Gureghian, M., Tuller, L., & Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *PLOS ONE, 14*(8), Article e0220611. https://doi.org/10.1371/journal.pone.0220611

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21*(2), 339–366. https://doi.org/10.1017/s0305000900009302

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–124). Multilingual Matters.

Marinis, T., & Saddy, D. (2013). Parsing the passive: Comparing children with specific language impairment to sequential bilingual children. *Language Acquisition, 20*(2), 155–179. https://doi.org/10.1080/10489223.2013.766743

Mastropavlou, M., & Tsimpli, I. M. (2011). Complementizers and subordination in typical language acquisition and SLI. *Lingua, 121*(3), 442–462. https://doi.org/10.1016/j.lingua.2010.10.009

Mayer, M. (1969). *Frog, where are you?* SALT Software.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112,* 104092. https://doi.org/10.1016/j.jml.2020.104092

Mohammad, M. A. (2000). *Word order, agreement and pronominalization in Standard and Palestinian Arabic*. John Benjamins. https://doi.org/10.1075/cilt.181

Moll, K., Hulme, C., Nag, S., & Snowling, M. J. (2015). Sentence repetition as a marker of language skills in children with dyslexia. *Applied Psycholinguistics, 36*(2), 203–221. https://doi.org/10.1017/S0142716413000209

Newcomer, P., & Hammill, D. (2008). *Test of Language Development–Primary–Fourth Edition*. Pro-Ed.

Novogrodsky, R., & Friedmann, N. (2006). The production of relative clauses in syntactic SLI: A window to the nature of the impairment. *International Journal of Speech-Language Pathology, 8*(4), 364–375. https://doi.org/10.1080/14417040600919496

Omar, M. (1973). *The acquisition of Egyptian Arabic as a native language*. De Gruyter Mouton. https://doi.org/10.1515/9783110819335

Ouali, H. (2018). The syntax of tense in Arabic. In E. Benmamoun & R. Bassiouney (Eds.), *The Routledge handbook of Arabic linguistics* (pp. 89–103). Routledge.

Owen, A. J., & Leonard, L. B. (2006). The production of finite and nonfinite complement clauses by children with specific language impairment and their typically developing peers. *Journal of Speech, Language, and Hearing Research, 49*(3), 548–571. https://doi.org/10.1044/10902-4388(2006/040)

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*(6), 2261–2273. https://doi.org/10.1044/2014_JSLHR-L-13-0189

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

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. https://doi.org/10.1044/0161-1461.2501.15

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders, 50*(1), 106–118. https://doi.org/10.1111/1460-6984.12126

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language, 38*(3), 265–282. https://doi.org/10.1006/jmla.1997.2546

Prévost, P., Strik, N., & Tuller, L. (2014). *Wh*-questions in child L2 French: Derivational complexity and its interactions with L1 properties, length of exposure, age of exposure, and the input. *Second Language Research, 30*(2), 225–250. https://doi.org/10.1177/0267658313519814

Qasem, F., & Sircar, S. (2017). Imperative as root infinitive analogue in Yemeni Ibbi Arabic: Two case studies. *Arab Journal of Applied Linguistics, 2*(1), 2490–4198.

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

Rakhlin, N. V., Aljughaiman, A., & Grigorenko, E. L. (2021). Assessing language development in Arabic: The Arabic Language: Evaluation of Function (ALEF). *Applied Neuropsychology: Child, 10*(1), 37–52. https://doi.org/10.1080/21622965.2019.1596113

Raven, J. (2007). *Coloured Progressive Matrices*. Pearson.

Ravid, D., & Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language, 19*(56), 187–206. https://doi.org/10.1177/014272379901905603

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. https://doi.org/10.1044/2019_JSLHR-L-18-0388

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*(6), 1239–1257. https://doi.org/10.1044/jshr.3906.1239

Riches, N. G. (2012). Sentence repetition in children with specific language impairment: An investigation of underlying mechanisms. *International Journal of Language & Communication Disorders, 47*(5), 499–510. https://doi.org/10.1111/j.1460-6984.2012.00158.x

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language & Communication Disorders, 45*(1), 47–60. https://doi.org/10.3109/13682820802647676

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 Bioinformatics, 12*(1), 1471–2105. https://doi.org/10.1186/1471-2105-12-77

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children, 8*(7), 598. https://doi.org/10.3390/CHILDREN8070578

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Little, Brown and Co.

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*(7), 2750–2765. https://doi.org/10.1044/2021_JSLHR-20-00552

Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language & Communication Disorders, 45*(6), 691–702. https://doi.org/10.3109/13682820903509432

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *Early Repetition Battery (ERB)*. Pearson Assessment.

Shaalan, S. (2010). *Investigating grammatical complexity in Gulf Arabic speaking children with specific language impairment (SLI)* [Doctoral dissertation, University College London].

Shaalan, S., & Khater, M. (2006). *A comparison of two measures of assessing spontaneous language samples in Arabic speaking children* [Poster presentation]. Child Language Seminar 2006, University of Newcastle Upon-Tyne, United Kingdom.

Slobin, D., & Welsh, C. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. Slobin (Eds.), *Studies of child language development*. Holt, Rinehart, & Winston.

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. https://doi.org/10.1044/1092-4388(2006/019)

Stromswold, K. (1995). The acquisition of subject and object *wh*-questions. *Language Acquisition, 4*(1–2), 5–48. https://doi.org/doi.org/10.1080/10489223.1995.9671658

Taha, J., Stojanovik, V., & Pagnamenta, E. (2021). Expressive verb morphology deficits in Arabic-speaking children with developmental language disorder. *Journal of Speech, Language, and Hearing Research, 64*(2), 561–578. https://doi.org/10.1044/2020_JSLHR-19-00292

Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2017). Sentence repetition as a tool for screening morphosyntactic abilities of bilectal children with SLI. *Frontiers in psychology, 8*, 2104. https://doi.org/10.3389/fpsyg.2017.02104

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. https://doi.org/10.1044/1092-4388(2010/09-0196)

Tuller, L., Delage, H., Monjauze, C., Piller, A. G., & Barthez, M. A. (2011). Clitic pronoun production as a measure of atypical language development in French. *Lingua, 121*(3), 423–441. https://doi.org/10.1016/j.lingua.2010.10.008

Vang Christensen, R. (2019). Sentence repetition: A clinical marker for developmental language disorder in Danish. *Journal of Speech, Language, and Hearing Research, 62*(12), 4450–4463. https://doi.org/10.1044/2019_JSLHR-L-18-0327

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics, 12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Wallan, A. A. (2018). *Evaluation of Arabic tests of sentence repetition and verbal short term memory for Saudi preschoolers* [Doctoral dissertation, University of London].

Wiig, E., Secord, W., & Semel, E. (2013). *Clinical evaluation of language fundamentals: CELF-5*. NCS Pearson.

Windfuhr, K. L., Faragher, B., & Conti-Ramsden, G. (2002). Lexical learning skills in young children with specific language impairment (SLI). *International Journal of Language & Communication Disorders, 37*(4), 415–432. https://doi.org/10.1080/1368282021000007758

Zebib, R., Henry, G., Messarra, C., Hreich, E. K., & Khomsi, A. (2019). ELO-L: A norm-referenced language screening test for 3 to 8-year-old Lebanese children. *Arab Journal of Applied Linguistics, 4*(2), 24–53.

Appendix (p. 1 of 3)

List of Items in the LITMUS-SR-PA-72 Task

Level	Target structure Mean length (SD)	Substructure	Item	Arabic sentences "Approximate English translation"
1	*Past tense* 7.83 syllables (0.75)	PAST-3MS	1	بابا اشترى * سيارة "Daddy bought a car"
			6	رسم الولد شجرة "The boy drew a tree"
		PAST-3FS	7	ماما غسلت الصحن "Mommy washed the dish"
			24	أكلت البسة جبنة "The cat ate cheese"
		PAST-3PL	12	لحقوا البقرات الولد "The cows chased the boy"
			21	الأولاد شربوا عصير "The boys drank juice"
	Present tense 7.66 syllables (0.82)	PRES-3MS	14	سيدو بسوق السيارة "Grandpa is driving the car"
			17	الولد بلقط وردة "The boy is picking a flower"
		PRES-3FS	2	ماما بتقرأقصة "Mommy is reading a story"
			11	بتلون البنت الحيط "The girl is painting the wall"
		PRES-3PL	4	يو الاولاد فوتبولبلع " The boys are playing football"
			13	البنات بنضفوا البيت "The girls are cleaning the house"
	Noun plurals 7.83 syllables (1.17)	Feminine sound plurals	3	تيتا كسرت الكاسات "Grandma broke the glasses"
			22	رمى الولد الطابات " The boy threw the balls"
		Masculine sound plurals	9	تيتيا نادت الطباخين "Grandma called the cooks"
			16	نادى الولد االبياعين "The boy called the salesmen"
		Broken plurals	5	أكل القرد الموز " The monkey ate the bananas"
			10	البنت ضيعت المفاتيح "The girl lost the keys"
	Possessive pronouns 8.33 syllables (.82)	CL-3MS	15	الولد ضيع طابته "The boy lost his ball"
			20	الولد كسر لعبته "The boy broke his game"
		CL-3FS	18	حممت البنت لعبتها "The girl washed her doll"
			23	البنت مشطت شعراتها "The girl brushed her hair"
		CL-3PL	8	البنات نضفوا بيتهم "The girls cleaned their house"
			19	بياولاد نشفوا ايدي يهم "The boys dried their hands"

(*table continues*)

Appendix (p. 2 of 3)

List of Items in the LITMUS-SR-PA-72 Task

Level	Target structure Mean length (SD)	Substructure	Item	Arabic sentences "Approximate English translation"
2	*Passives* 9.5 syllables (.58)		25	الكاسة انكسرت من الولد "The glass got broken by the boy"
			34	الشباك انفتح من الهوا "The window got opened by the wind"
			38	السيارة توسخت من المطر "The car got dirty by the rain"
			41	الولد انضرب على بطنه "The boy got his on his stomach"
	Clitic left dislocation 8.5 syllables (1.29)		26	البلون الولد فقعه "The balloon, the boy popped it"
			36	الكعكة ماما عملته "The cake, mom made it"
			39	الهدية فتحتها البنت "The gift, the girls opened it"
			45	البنطلون لبسو الولد "The trouser, the boy wore it"
	Complement clauses 12 syllables (2.31)	finite	29	فكرت ماما إنو البسة أكلت فار "Mommy thought that the cat ate the mouse"
			44	فكر بابا انو الولد ضيع اللعبة "Dad thought that the boy lost the toy"
		Nonfinite	32	بدها البنت تاكل جزرة "The girl wants to eat a carrot"
			47	سيدو بحب ياكل شوكلاطه "Grandpa likes to eat chocolate"
	Wh Object 10.25 syllables (1.04)	Who	27	مين البنت اللي تيتا باستها Who is the girl that grandma kissed?"
			31	مين الولد اللي ساعدو بابا؟ "Who is the boy that dad helped?"
			42	مين الولد اللي الكلب لحقه؟ Who is the boy that the dog chased?"
			46	مين البيبي اللي طعمته ماما؟ "Who is the baby that mommy fed?"
		Which	28	أني بسة حملتها البنت؟ "Which cat did the girl carry?"
			30	أنو تلفون الولد خربوا؟ "Which phone did the boy break?"
			37	أني هدية البنت اعطتها؟ "Which gift did the girl give?"
			43	أنو صندوق فتحه بابا؟ "Which box did dad open?"
	Coordinates 13.9 syllables (1.63)		33	ماما قرأت قصة و الولد نام "Mommy read a story and the boy slept"
			35	القرد طلع عالشجرة و العصفور طار "The monkey climbed the tree and the bird flew"
			40	تيتا عملت شاي و بابا أكل بسكوت "Grandma made tea and dad ate biscuits"
			48	سيدو غسل السيارة و تيتا كتبت رسالة "Grandpa washed the car and grandma wrote a letter"

(*table continues*)

Appendix (p. 3 of 3)

List of Items in the LITMUS-SR-PA-72 Task

Level	Target structure Mean length (SD)	Substructure	Item	Arabic sentences "Approximate English translation"
3	*Object relatives* 11.5 syllables (.93)	Reversible	52	هاي البنت اللي نيمتها تيتيا "This is the girl that mom put to sleep"
			59	هاد الحمار اللي الكلب دفعه "This is the donkey that the dog chased"
			64	هاد سيدو اللي ساعدو الولد "This is the grandpa that the boy helped"
			69	هاي البسة اللي السلحفاة عضتها "This is cat that the turtle bit"
		Irreversible	53	هاي القصة اللي تيتيا قرأتها "This is the story that grandma read"
			56	هاي البسة اللي لاقتها البنت "This is the cat that the girl found"
			65	هاد العصير اللي شربوا الولد "This is the juice that the boy drank"
			70	هاد الشباك اللي الولد سكره "This is the window that the boy closed"
	Subject relatives 10.7 5 syllables (.46)	Reversible	49	هاد الولد اللي باس البنت "This is the boy that kissed the girl"
			54	هاد الولد اللي حضن ماما "This is the boy that hugged mom"
			60	هاي تيتا اللي صورت البنت "This is the grandpa who took a picture of the girl"
			67	هاي البنت اللي شافت الأرنب "This is the girl that saw the rabbit'
		Irreversible	50	هاي البنت اللي اشترت كتاب "This is the girl that bought a book"
			63	هاد الولد اللي غسل السيارة "This is the boy that washed the car"
			71	هاد الولد اللي نفخ البلون "This is the boy that blew the balloon"
			57	هاي البنت اللي رسمت بيت "This is the girl who drew a house"
	Conditionals 13.75 syllables (1.50)		51	اذا الولد بيعمل الواجب، راح ياخد نجمة "If the boy does the homework, he will take a sticker"
			62	اذا الولد برتب الغرفة، راح يروح عالحديقة "If the boy tidies the room, he will go to the park"
			66	اذا ماما بتشتري بيض، راح تعمل كعكة "If mommy buys eggs, she will make a cake"
			71	اذا البنت بتساعد ماما، راح تشتري فستان "If the girl helps mommy, she will buy a dress"
	Subordinates 12.25 syllables (1.71)		55	البنت وقعت عشان الأرض مبلولة "The girl fell because the floor is wet"
			58	الولد عيط عشان عشان ضيع اللعبة "The boy cried because he lost the toy"
			61	الولد بيدرس عشان عنده امتحان "The boy is studying because he has a test"
			68	البنت بتشرب مي عشنها عطشانة "The girl is drinking water because she is thirsty"

Note. PAST-3MS = third-person masculine singular past; PAST-3FS = third-person feminine singular past; PAST-3PL = third-person plural past; PRES-3MS = third-person masculine singular present; PRES-3FS = third-person feminine singular present; PRES-3PL = third-person plural present; CL-3MS = third-person masculine singular clitic; CL-3FS = third-person feminine singular clitic; CL-3PL = third-person plural clitic.

*Underlined word is the target language-specific structure.